PERBANDINGAN ALGORITMA C4.5 DAN NAÏVE BAYES UNTUK MEMPREDIKSI PRESTASI SISWA/SISWI

SKRIPSI

DISUSUN OLEH

FADLI DWI YULIANTO

NPM. 2009020032

PROGRAM STUDI TEKNOLOGI INFORMASI

FAKULTAS ILMU KOMPUTER DAN TEKNOLOGI INFORMASI

UNIVERSITAS MUHAMMADIYAH SUMATERA UTARA

MEDAN

2024

PERBANDINGAN ALGORITMA C4.5 DAN NAÏVE BAYES UNTUK MEMPREDIKSI PRESTASI SISWA/SISWI

SKRIPSI

Diajukan sebagai salah satu syarat untuk memperoleh gelar Sarjana Komputer (S.Kom) dalam Program Studi Teknologi Informasi pada Fakultas Ilmu Komputer dan Teknologi Informasi, Universitas Muhammadiyah Sumatera Utara

FADLI DWI YULIANTO
NPM. 2009020032

PROGRAM STUDI TEKNOLOGI INFORMASI
FAKULTAS ILMU KOMPUTER DAN TEKNOLOGI INFORMASI
UNIVERSITAS MUHAMMADIYAH SUMATERA UTARA
MEDAN
2024

LEMBAR PENGESAHAN

Judul Skripsi	:	PERBANDINGAN ALGORITMA C4.5 DAN NAÏVE BAYES UNTUK MEMPREDIKSI PRESTASI SISWA/SISWI
Nama Mahasiswa	:	FADLI DWI YULIANTO
NPM	:	2009020032
Program Studi	:	TEKNOLOGI INFORMASI

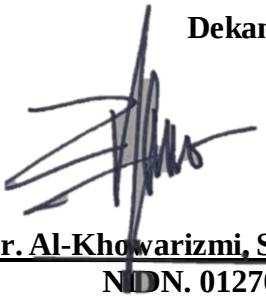
Menyetujui
Komisi Pembimbing

(Dr. Al-Khowarizmi, S.Kom., M.Kom.)
NIDN. 0127099201

Ketua Program Studi

(Fatma Sari Hutagalung, S.Kom., M.Kom)
NIDN. 0117019301

Dekan



(Dr. Al-Khowarizmi, S.Kom., M.Kom.)
NIDN. 0127099201

PERNYATAAN ORISINALITAS

PERBANDINGAN ALGORITMA C4.5 DAN NAÏVE BAYES UNTUK MEMPREDIKSI PRESTASI SISWA/SISWI

SKRIPSI

Saya menyatakan bahwa karya tulis ini adalah hasil karya sendiri, kecuali beberapa kutipan dan ringkasan yang masing-masing disebutkan sumbernya.

Medan, Agustus 2024

Yang membuat pernyataan

Fadli Dwi Yulianto

NPM. 2009020032

PERNYATAAN PERSETUJUAN PUBLIKASI KARYA ILMIAH UNTUK KEPENTINGAN AKADEMIS

Sebagai sivitas akademika Universitas Muhammadiyah Sumatera Utara, saya bertanda tangan dibawah ini:

Nama	: Fadli Dwi Yulianto
NPM	2009020032
Program Studi	: Teknologi Informasi
Karya Ilmiah	: Skripsi

Demi pengembangan ilmu pengetahuan, menyetujui untuk memberikan kepada Universitas Muhammadiyah Sumatera Utara Hak Bedas Royalti Non-Eksekutif *(Non-Exclusive Royalty free Right)* atas penelitian skripsi saya yang berjudul:

PERBANDINGAN ALGORITMA C4.5 DAN NAÏVE BAYES UNTUK MEMPREDIKSI PRESTASI SISWA/SISWI

Beserta perangkat yang ada (jika diperlukan). Dengan Hak Bebas Royalti Non-Eksekutif ini, Universitas Muhammadiyah Sumatera Utara berhak menyimpan, mengalih media, memformat, mengelola dalam bentuk database, merawat dan mempublikasikan Skripsi saya ini tanpa meminta izin dari saya selama tetap mencantumkan nama saya sebagai penulis dan sebagai pemegang dan atau sebagai pemilik hak cipta.

Demikian pernyataan ini dibuat dengan sebenarnya.

Medan, Agustus 2024

Yang membuat pernyataan

Fadli Dwi Yulianto

NPM. 2009020032

RIWAYAT HIDUP

DATA PRIBADI

Nama Lengkap	: Fadli Dwi Yulianto
Tempat dan Tanggal Lahir	: Medan, 08 Juli 2002
Alamat Rumah	: Jl. Al-Falah VI No. 20
Telepon/Faks/HP	082278844918
E-mail	: fadlydwiyuli02@gmail.com
Instansi Tempat Kerja	: -
Alamat Kantor	: -

DATA PENDIDIKAN

SD	: SD Muhammadiyah 02 Medan	TAMAT: 2014
SMP	: SMPN 37 Medan	TAMAT: 2017
SMA	: SMAN 7 Medan	TAMAT: 2020

KATA PENGANTAR

Pendahuluan

Penulis tentunya berterima kasih kepada berbagai pihak dalam dukungan serta doa dalam penyelesaian skripsi. Penulis juga mengucapkan terima kasih kepada:

1. Bapak Prof. Dr. Agussani, M.AP., Rektor Universitas Muhammadiyah Sumatera Utara (UMSU)
2. Bapak Dr. Al-Khowarizmi, S.Kom., M.Kom. Dekan Fakultas Ilmu Komputer dan Teknologi Informasi (FIKTI) UMSU.
3. Ibu Fatma Sari Hutagalung, S.Kom., M.Kom, selaku Ketua Program Studi Teknologi Informasi yang telah memberikan izin kepada penulis untuk menyusun skripsi.
4. Bapak Mhd Basri, S.Si., M.Kom, selaku Sekretaris Program Studi Teknologi Informasi.
5. Bapak Dr. Al-Khowarizmi, S.Kom., M.Kom., selaku Dosen Pembimbing yang telah banyak meluangkan waktunya untuk memberikan bimbingan, arahan dan saran dalam penelitian skripsi ini.
6. Orang Tua dan abang penulis, Ibunda Tri Julianti dan Abangda Fredy Wandana yang sudah memberikan doa dan dukungan baik secara material maupun non-material.
7. Wahyu, Rizki, Rio, Atif yang selalu memberi motivasi dan dukungan kepada saya.
8. Team Kost Petarung yang selalu memberi dukungan selama perkuliahan saya.
9. Semua pihak yang terlibat langsung ataupun tidak langsung yang tidak dapat penulis ucapkan satu-persatu yang telah membantu penyelesaian skripsi ini.

PERBANDINGAN ALGORITMA C4.5 DAN NAÏVE BAYES
UNTUK MEMPREDIKSI PRESTASI SISWA/SISWI

ABSTRAK

Penelitian ini bertujuan untuk menganalisis dan memprediksi prestasi siswa menggunakan teknik data mining dengan metode C4.5 dan Naive Bayes. Data yang digunakan meliputi berbagai faktor yang mempengaruhi prestasi akademik siswa, seperti nilai sebelumnya, kehadiran, dan penghasilan orang tua. Metode C4.5, yang merupakan algoritma pohon keputusan, digunakan untuk mengidentifikasi pola-pola dalam data dan membuat keputusan berbasis aturan. Sementara itu, Naive Bayes, yang merupakan teknik klasifikasi probabilistik, digunakan untuk menghitung kemungkinan prestasi berdasarkan distribusi fitur yang ada. Model algoritma C4.5 menunjukkan performa yang sangat baik dalam mengklasifikasikan siswa ke dalam kategori "Kurang Berprestasi" dan "Berprestasi," dengan akurasi dan F1-Score yang sempurna untuk kedua kelas. Di sisi lain, model Naive Bayes menunjukkan hasil yang kurang optimal, terutama dalam mengenali siswa "Berprestasi." Meskipun model Naive Bayes berhasil memprediksi semua siswa "Kurang Berprestasi" dengan benar, model tersebut gagal sepenuhnya dalam mendeteksi siswa "Berprestasi," yang terlihat dari F1-Score yang nol untuk kelas tersebut.

Kata Kunci: Data mining, prediksi prestasi siswa, metode C4.5, Naive Bayes, klasifikasi.

COMPARISON OF C4.5 AND NAÏVE BAYES ALGORITHMS TO PREDICT STUDENT ACHIEVEMENT

ABSTRACT

This research aims to analyze and predict student achievement using data mining techniques with the C4.5 and Naive Bayes methods. The data used includes various factors that affect students' academic performance, such as previous grades, attendance, and parents' income. The C4.5 method, which is a decision tree algorithm, is used to identify patterns in the data and make rule-based decisions. Meanwhile, Naive Bayes, which is a probabilistic classification technique, is used to calculate the probability of achievement based on the distribution of features. The C4.5 algorithm model showed excellent performance in classifying students into the categories of "Underachieving" and "Achieving," with perfect accuracy and F1-Score for both classes. On the other hand, the Naive Bayes model showed less than optimal results, especially in recognizing "Outstanding" students. Although the Naive Bayes model managed to correctly predict all the "Underachieving" students, it failed completely in detecting the "Achieving" students, as seen from the zero F1-Score for the class.

Keywords: Data mining, student achievement prediction, C4.5 method, Naive Bayes, classification.

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

BAB I

PENDAHULUAN

1.1. Latar Belakang Masalah

Teknologi adalah perkembangan baik dalam perangkat keras (hardware) maupun perangkat lunak (software) yang didasari oleh ilmu pengetahuan dan kebutuhan pengguna yang terus berkembang seiring waktu. Kemajuan teknologi telah mengubah cara kita bekerja, yang sebelumnya dilakukan secara manual, seperti dalam pengiriman surat dan pembuatan laporan keuangan, kini dapat dilakukan dengan SMS (Short Message Service) atau aplikasi komputer untuk laporan keuangan (Karim et al., 2021). Kemajuan teknologi informasi dalam beberapa tahun terakhir berkembang dengan sangat pesat, sehingga mengubah cara masyarakat dalam mencari informasi, yang kini tidak hanya dari surat kabar, audio visual, dan media elektronik, tetapi juga melalui internet. Salah satu bidang yang sangat terpengaruh oleh kemajuan teknologi ini adalah pendidikan, yang pada dasarnya merupakan proses komunikasi informasi antara pendidik dan peserta didik untuk menyampaikan materi pembelajaran (Kusumawati, 2023).

Penelitian ini membandingkan algoritma C4.5 dan Naïve Bayes dalam memprediksi prestasi siswa. Hasilnya menunjukkan bahwa algoritma pohon keputusan, seperti C4.5, memiliki prediksi yang lebih baik. Selain itu, beberapa penelitian terdahulu juga banyak menggunakan Naïve Bayes dan C4.5 dalam menggali pengetahuan baru. Algoritma C4.5 sendiri merupakan teknik klasifikasi data yang menggunakan metode pohon keputusan, di mana atribut paling atas menjadi akar, sementara bagian bawahnya disebut daun (Romli & Zy, 2020).

Algoritma Naïve Bayes didasarkan pada asumsi sederhana bahwa nilai-nilai atribut bersifat independen jika diberikan nilai output tertentu. Dengan kata

lain, probabilitas gabungan dari beberapa atribut diperoleh dari hasil kali probabilitas individu atribut-atribut tersebut. Kelebihan dari Naïve Bayes adalah bahwa algoritma ini hanya membutuhkan data pelatihan yang relatif sedikit untuk memperkirakan parameter yang diperlukan dalam proses klasifikasi. Naïve Bayes sering kali bekerja lebih baik daripada yang diperkirakan dalam situasi kompleks di dunia nyata (Kawani, 2019).

Penelitian terdahulu juga menunjukkan bahwa algoritma Naïve Bayes dapat digunakan untuk memprediksi prestasi siswa dan dibandingkan dengan Neural Network, di mana hasilnya menunjukkan akurasi Neural Network yang lebih tinggi. Selain itu, prediksi prestasi siswa juga telah dilakukan menggunakan algoritma Support Vector Machine dan sistem pendukung keputusan hybrid (Rovidatul et al., 2023).

Untuk menangani permasalahan tersebut, penelitian ini bertujuan membandingkan akurasi dua algoritma dari data mining, yaitu algoritma C4.5 dan Naïve Bayes, pada berbagai dataset. Perbandingan ini dilakukan untuk mengetahui algoritma mana yang memiliki akurasi prediksi prestasi siswa yang lebih tinggi (Rahmayanti et al., 2022).

Pendidikan adalah upaya sadar dan terencana untuk menciptakan suasana belajar yang memungkinkan peserta didik mengembangkan potensi diri, baik spiritual, kepribadian, kecerdasan, maupun keterampilan yang berguna bagi diri sendiri, masyarakat, bangsa, dan negara. Berdasarkan UU No. 20 Tahun 2003 tentang Sistem Pendidikan Nasional Pasal 3, tujuan pendidikan nasional adalah mengembangkan potensi peserta didik agar beriman, bertaqwa, berakhlak mulia, sehat, berilmu, cakap, kreatif, mandiri, dan menjadi warga negara yang bertanggung jawab. Dengan demikian, kualitas dan manajemen pembelajaran di

sekolah perlu ditingkatkan, yang dapat dilihat dari prestasi belajar siswa (Noviriandini & Nurajijah, 2019).

Prestasi siswa dapat dilihat dari pengelompokan kelas sesuai kemampuan individu masing-masing. Inilah alasan penelitian ini dilakukan di SMA Asuhan Daya, dengan mempertimbangkan berbagai variabel yang mempengaruhi prestasi siswa. Dalam hal ini, metode pengambilan keputusan di sekolah ini masih tradisional, yang menyebabkan kesulitan dalam menentukan prestasi siswa. Oleh karena itu, dibutuhkan perubahan yang dapat membantu sekolah dalam meningkatkan kualitas pendidikan melalui pemahaman yang lebih baik tentang prestasi siswa. Untuk itu, sistem perhitungan berbasis data mining diperlukan untuk mengelompokkan siswa berdasarkan prestasi mereka (Br Sembiring et al., 2022).

Berdasarkan latar belakang tersebut, penulis tertarik untuk melakukan penelitian berjudul "Perbandingan Algoritma C4.5 dan Naïve Bayes untuk Memprediksi Prestasi Siswa."

1.2. Rumusan Masalah

Rumusan masalah dari penelitian ini adalah perbandingan algoritma yang lebih baik antara algoritma C4.5 dan algoritma Naive Bayes dalam memprediksi prestasi siswa/siswi di SMA Asuhan Daya Medan, akurasi masing-masing algoritma dalam memprediksi prestasi siswa/siswi di SMA Asuhan Daya Medan

1.3. Batasan Masalah

Batasan masalah sangat berguna agar pembahasan yang dilakukan penulis dapat terarah sesuai dengan tujuan penulisan maka batasan masalah sebagai berikut:

1. Metode yang digunakan pada penelitian ini fokus pada dua algoritma, yaitu C4.5 dan Naive Bayes.

2. Data yang digunakan pada penelitian ini merupakan data yang diambil dari SMA Asuhan Daya.

1.4. Tujuan Penelitian

Adapun tujuan yang hendak di capai dari penelitian ini yaitu :

1. Untuk mengetahui hasil prediksi prestasi murid dengan menggunakan algoritma C4.5 dan Naive Bayes.
2. Untuk mengetahui hasil implementasi data mining dalam memprediksi prestasi murid dengan hasil data yang lebih akurat.

1.5. Manfaat Penelitian

Manfaat penelitian ini yaitu dengan adanya Implementasi data mining yang bisa memprediksi prestasi diharapkan bisa memberikan gambaran ilmiah tentang proses prediksi prestasi siswa/siswi menggunakan Algoritma C4.5 dan Naive Bayes terhadap data yang diolah.

BAB II

LANDASAN TEORI

2.1. Data Mining

Penambangan data adalah proses penemuan teknik analisis untuk menelusuri data dalam jumlah besar dari berbagai basis data, seperti data relasional, data berorientasi objek, dan data transaksi, guna menemukan informasi baru yang terdapat dalam basis data tersebut (Andini et al., 2022). Perkembangan pesat dalam teknologi informasi telah memungkinkan akumulasi data dalam jumlah yang sangat besar, yang juga mempercepat pertumbuhan penambangan data (Marlina & Bakri, 2021). Namun, pesatnya akumulasi data sering kali menciptakan kondisi “kaya data tetapi miskin informasi” karena data yang terkumpul tidak selalu dapat diolah menjadi informasi yang berguna, bahkan terkadang dibiarkan begitu saja hingga membentuk "kuburan data."

2.2. C4.5

Algoritma ID3 dikembangkan lebih lanjut menjadi Algoritma C4.5 oleh Quinlan, yang juga merupakan pengembang Algoritma C4.5 ini. Algoritma C4.5 adalah teknik klasifikasi yang sering digunakan oleh para peneliti. Hasil dari penghitungan algoritma ini adalah pohon keputusan atau decision tree (Hana, 2020).

Beberapa rumus yang digunakan dalam pengolahan data menggunakan algoritma ini adalah sebagai berikut:

a. Entropy

Entropi (S) merupakan estimasi jumlah bit yang dibutuhkan untuk mengidentifikasi suatu kelas (+ atau -) dari sekumpulan data acak dalam ruang sampel S. Entropi dapat dianggap sebagai kebutuhan bit untuk

merepresentasikan suatu kelas. Fungsi entropi ini digunakan untuk mengukur ketidakmurnian S. Perhitungan entropi dilakukan sebagai berikut:

$$Entropy(S) = -\sum_{i}^{n} \text{pi} * \log 2 \text{ pi}$$

Keterangan:

Entropy(S): Mengukur ketidakpastian atau kemurnian dari himpunan S.

$\sum_{i}^{n}$: Simbol sigma yang menunjukkan penjumlahan dari i=1 sampai n.

pi: Probabilitas kelas ke-i dalam himpunan S.

$\log_2$ pi: Logaritma basis 2 dari probabilitas pi.

b. Gain

Gain (S, A) adalah nilai informasi yang diperoleh dari atribut A terhadap output data S. Perolehan informasi ini didapatkan dari output atau variabel dependen S yang dikelompokkan berdasarkan atribut A, dan dinotasikan sebagai gain (S, A). Atribut yang dipilih untuk pemrosesan akan didasarkan pada nilai gain tertinggi dari seluruh atribut yang tersedia. Rumusnya adalah sebagai berikut:

$$Gain\ (S,\ A) = Entropy(S) - \sum(\ |\text{Si}|/|\text{S}|\ *\ \text{Entropy(Si)}\ )$$

Keterangan:

Gain(S, A): Nilai gain yang diperoleh dengan membagi dataset S berdasarkan kata atau frasa A.

Entropy(S): Entropi dari seluruh dataset S.

$\sum$: Simbol sigma yang menunjukkan penjumlahan.

|Si|/|S|: Proporsi dokumen dalam subset Si terhadap total dokumen.

Entropy(Si): Entropi dari subset Si.

c. Split Info

Split Info adalah rumus yang digunakan untuk memisahkan atau membagi atribut data berdasarkan nilai gain yang telah dihasilkan. Perhitungan ini membantu menentukan informasi tambahan yang dihasilkan dari pembagian atribut tersebut. Rumus Split Info adalah sebagai berikut:

$$SplitInfo(A) = -p(A = 1|s) * \log2(\text{p}(\text{A} = 1|\text{s})) - p(A = 0|s) * log2(p(A = 0|s))$$

Keterangan:

A: Keberadaan suatu kata.

p(A=1|s): Probabilitas suatu kata muncul dalam suatu dokumen dari himpunan data s.

p(A=0|s): Probabilitas suatu kata tidak muncul dalam suatu dokumen dari himpunan data s.

d. Gain rasio

Gain rasio digunakan sebagai metode untuk menentukan atribut terbaik dengan menghitung ulang nilai gain yang diperoleh pada setiap tahapannya. Rumusnya adalah sebagai berikut:

$$GainRatio(S, j) = Gain(S, j) \; / \; SplitInfo(S, j)$$

Keterangan:

S: Himpunan data (dokumen).

j: Atribut (kata).

Gain(S, j): Hitung menggunakan rumus Information Gain yang telah dijelaskan sebelumnya.

SplitInfo(S, j): Hitung menggunakan rumus SplitInfo yang telah

dijelaskan sebelumnya.

2.3. Naïve Bayes

Naive Bayes adalah metode yang digunakan dalam pengambilan keputusan, di mana metode ini menentukan klasifikasi dengan mencari probabilitas tertinggi dari data latih. Algoritma ini hanya memerlukan sedikit data latih untuk memperkirakan parameter yang dibutuhkan (Indahsari et al., 2021). Metode Bayes sendiri merupakan pendekatan statistik yang digunakan untuk melakukan inferensi induktif dalam masalah klasifikasi. Prosesnya dimulai dengan memahami konsep dasar dan definisi Teorema Bayes, yang kemudian digunakan untuk klasifikasi dalam analisis data. Teorema Bayes memiliki bentuk umum sebagai berikut:

$$P(H|X) = P(X|H).P(H) \ / \ P(X)$$

Keterangan

X : Data dengan class yang belum diketahui

H : Hipotesis data merupakan suatu class spesifik

P(H|X): Probabilitas posterior. Ini adalah probabilitas dari hipotesis H (sesuatu yang ingin kita ketahui) setelah kita mengamati bukti X. Ini adalah hasil yang ingin kita hitung.

P(X|H): Likelihood. Ini adalah probabilitas mengamati bukti X jika hipotesis H benar. P(H): Probabilitas prior. Ini adalah probabilitas awal kita tentang kebenaran hipotesis H sebelum kita melihat bukti apa pun.

P(X): Probabilitas bukti. Ini adalah probabilitas mengamati bukti X, terlepas dari kebenaran hipotesis H.

Tahapan proses Naive Bayes, yaitu:

a. Menghitung jumlah kelas / label

b. Menghitung Jumlah Kasus Per Kelas

c. Kalikan Semua Variable Kelas

d. Bandingkan Hasil Per Kelas

2.4. Prediksi

Prediksi atau peramalan adalah metode yang digunakan untuk memproyeksikan atau memperkirakan sesuatu yang belum terjadi. Prediksi ini adalah proses memproyeksikan secara terstruktur tentang kejadian yang mungkin di masa mendatang, berdasarkan data dan informasi dari masa lalu serta sekarang, sehingga tingkat kesalahan dapat diminimalkan. Berdasarkan pengertian ini, prediksi adalah kegiatan memperkirakan situasi di masa depan dengan menggunakan data masa lalu untuk meminimalkan kemungkinan kesalahan (Dewi et al., 2022).

2.5. Prestasi

Prestasi belajar adalah istilah yang terdiri dari dua kata, yakni "prestasi" dan "belajar," yang memiliki makna berbeda. Maka, sebelum membahas pengertian prestasi belajar, penting untuk lebih dahulu memahami arti masing-masing kata tersebut. Pendekatan ini akan membantu dalam memahami lebih mendalam makna "prestasi" dan "belajar" (Rambe, 2019).

2.6. Google Colab

Google Colab, singkatan dari Google Colaboratory, adalah platform berbasis cloud dari Google yang memungkinkan pengguna untuk menulis dan menjalankan kode Python melalui peramban web tanpa memerlukan konfigurasi tambahan. Platform ini memanfaatkan infrastruktur cloud Google dan menyediakan akses gratis ke GPU dan TPU (unit pemrosesan tensor). Dengan fitur kolaboratifnya, pengguna dapat berbagi notebook dan bekerja bersama secara real-time, menjadikan Google Colab alat yang sangat populer di kalangan pengembang, peneliti, dan pelajar untuk proyek Python. Platform ini dapat diakses melalui peramban tanpa perlu menginstal perangkat lunak tambahan, dan kemudahan berbagi notebook membuatnya ideal bagi pemula dan profesional yang ingin bekerja secara kolaboratif (Nazar, 2024).

Gambar 2.1 Google Colab

Sumber : https://www.pngwing.com/id/free-png-arnny/download

2.7. Penelitian Terdahulu

Tabel 2.1 Penelitian Terdahulu

No.	Judul	Metode	Hasil Penelitian	Tahun
1.	Prediksi Penerimaan SNMPTN Menggunakan Algoritma C4.5 dan Naïve Bayes	C4.5 dan Naïve Bayes	Penelitian ini menggunakan data mining klasifikasi dengan algoritma Naive Bayes dan C4.5 untuk memprediksi penerimaan SNMPTN berdasarkan 268 data siswa 31 atribut nilai siswa semester 1 sampai 5 dan atribut label yaitu Hasil. Dengan memanfaatkan software rapid miner 9.1 untuk eksperimen dan pengukuran menunjukkan bahwa hasil performansi algoritma C4.5 lebih baik dengan akurasi 85,09% dan AUC 0,873 sedangkan performansi algoritma Naive Bayes menghasilkan akurasi 63,01% dan AUC 0,665.	2023
2.	Pemanfaatan Algoritma Naïve Bayes dan K-Nearest Neighbor Untuk Klasifikasi Jurusan Siswa Kelas XI	Algoritma Naïve Bayes dan K-Nearest Neighbor	Hasil penelitian ini menggunakan algoritma Naïve bayes memberikan hasil accuracy sebesar 81.82% dan algoritma K-Nearest Neighbor memberikan hasil accuracy sebesar 92.73%, artinya jumlah data yang diprediksi memilih jurusan IPA dan pada kenyataaannya memilih IPA menghasilkan nilai yang signifikan atau dapat disebut sebagai True Positif.	2023

No.	Judul	Metode	Hasil Penelitian	Tahun
3.	Perbandingan Algoritma Naive Bayes, Decision Tree dan Random Forest dalam Klasifikasi Gaya Belajar Mahasiswa Universitas Kristen Indonesia Toraja	Algoritma Naive Bayes, Decision Tree dan Random Forest	Pemodelan algoritma dengan nilai akurasi tertinggi pada algoritma Naive Bayes dengan nilai akurasi 75% dan terendah pada algoritma Random Forest nilai 59% ini menunjukan bahwa algoritma Naive Bayes lebih cocok digunakan untuk klasifikasi gaya belajar mahasiswa . Hasil penelitian ini dapat diinterpretasikan sebagai pentingnya memahami gaya belajar mahasiswa, karena hal ini dapat membantu para pendidik dalam merancang pembelajaran yang lebih efektif dan efisien.	2023
4	Perbandingan Metode Cost Sensitive pada Decision Tree dan Naïve Bayes untuk Klasifikasi Data Multiclass	Cost sensitive decision tree C4.5 dan Cost sensitive naïve bayes	Berdasarkan hasil pengujian dan evaluasi dapat disimpulkan bahwa pengujian dengan menggunakan metode cost sensitive decision tree C4.5 memiliki nilai accuracy yang lebih baik dari pada menggunakan metode cost sensitive naïve bayes pada dataset glass, lympografi, vehicle dan wine berturut-turut 70.09%, 83.33%, 76.86% dan 97.62%. Sedangkan dengan menggunakan metode cost sensitive naïve bayes memiliki nilai accuracy yang lebih baik dari pada cost sensitive decision tree C4.5 pada dataset thyroid sebesar 97.67%.	2020

No.	Judul	Metode	Hasil Penelitian	Tahun
5	Perbandingan Algoritma C4.5 dan Naive Bayes Dalam Mendeteksi Hipertensi di Puskesmas Banyubiru	Algoritma C4.5 dan Naive Bayes	Hasil penelitian menunjukan bahwa model C4.5 memiliki akurasi yang lebih baik dengan 74,00 % dibandingkan model Naive Bayes dengan akurasi 67,00 %.	2023

BAB III

METODOLOGI PENELITIAN

3.1 Analisis Sistem Yang Berjalan

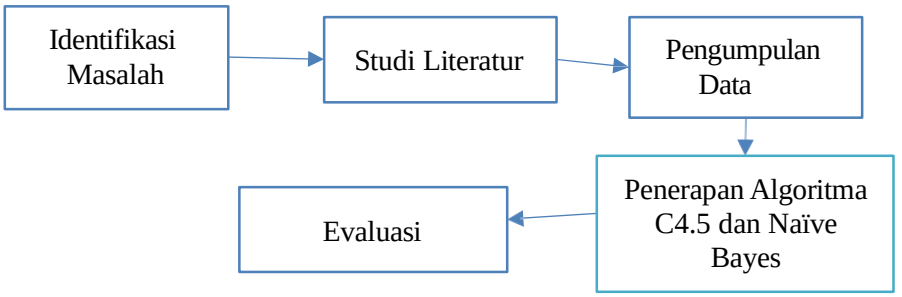
Upaya untuk melakukan prediksi dini terhadap siswa yang kemungkinan berpotensi kurang berprestasi bertujuan agar sekolah dapat mengambil tindakan pencegahan atau langkah antisipatif guna menghindari terjadinya tidak naik kelas atau dikeluarkannya siswa dari sekolah. Dengan mengenali siswa yang berpotensi mengalami kesulitan dalam prestasi akademik atau tidak naik kelas, sekolah dapat memberikan pendampingan khusus kepada siswa tersebut. Tujuannya adalah agar semua siswa, dengan beragam faktor latar belakang, dapat mengoptimalkan prestasi belajar mereka. Beberapa faktor yang dapat memengaruhi prestasi belajar siswa SMA meliputi kondisi sosial ekonomi, yang seringkali berkaitan dengan penghasilan orangtua, fasilitas belajar yang disediakan sekolah, tingkat kehadiran atau absensi, serta partisipasi siswa dalam kegiatan ekstrakurikuler. Namun, sistem yang saat ini berjalan belum mampu memprediksi prestasi siswa secara spesifik berdasarkan kedisiplinan dan status sosial mereka.

3.2 Analisis Kebutuhan Sistem

Dalam sistem yang sedang berjalan, terdapat beberapa aspek yang perlu dipenuhi, yaitu sistem yang mampu memprediksi prestasi siswa berdasarkan kedisiplinan dan status sosial di SMA Asuhan Daya Medan. Selain itu, diperlukan perangkat lunak dan perangkat keras yang mendukung kinerja sistem yang dikembangkan serta data atau sampel yang memadai untuk memaksimalkan proses kerja sistem tersebut.

3.3 Kerangka Penelitian

Dalam melaksanakan penelitian, diperlukan prosedur yang sistematis agar penelitian dapat berjalan dengan lancar. Prosedur penelitian ini bertujuan untuk membandingkan algoritma C4.5 dan Naïve Bayes dalam memprediksi prestasi akademik siswa.



Gambar 3.1 Diagram Prosedur Penelitian

Berikut adalah penjelasan dari masing-masing tahap dalam diagram prosedur penelitian.

1. Identifikasi Masalah

   Pada tahap ini, dilakukan pengidentifikasian masalah yang relevan dengan bidang studi. Masalah yang diangkat dalam penelitian ini adalah ketiadaan sistem yang mampu memprediksi prestasi siswa berdasarkan tingkat kehadiran mereka.

2. Studi Literatur

   Pada tahap ini, pencarian referensi dilakukan untuk mendukung topik penelitian, baik berupa buku maupun artikel jurnal. Pencarian literatur ini bertujuan untuk menemukan solusi yang dapat membantu dalam menyelesaikan permasalahan penelitian.

3. Pengumpulan Data

   a. Observasi

Metode ini digunakan untuk mengumpulkan data melalui pengamatan langsung, dengan tujuan memperoleh data yang relevan. Data penelitian diambil dari catatan prestasi siswa yang disediakan oleh pihak sekolah.

b. Wawancara

Wawancara dilakukan sebagai cara sistematis untuk mendapatkan informasi yang diperlukan melalui pertanyaan yang diajukan kepada pihak sekolah yang memiliki data prestasi siswa. Tujuannya adalah memperoleh informasi yang lebih lengkap dan akurat untuk pengembangan sistem baru sesuai kebutuhan penelitian.

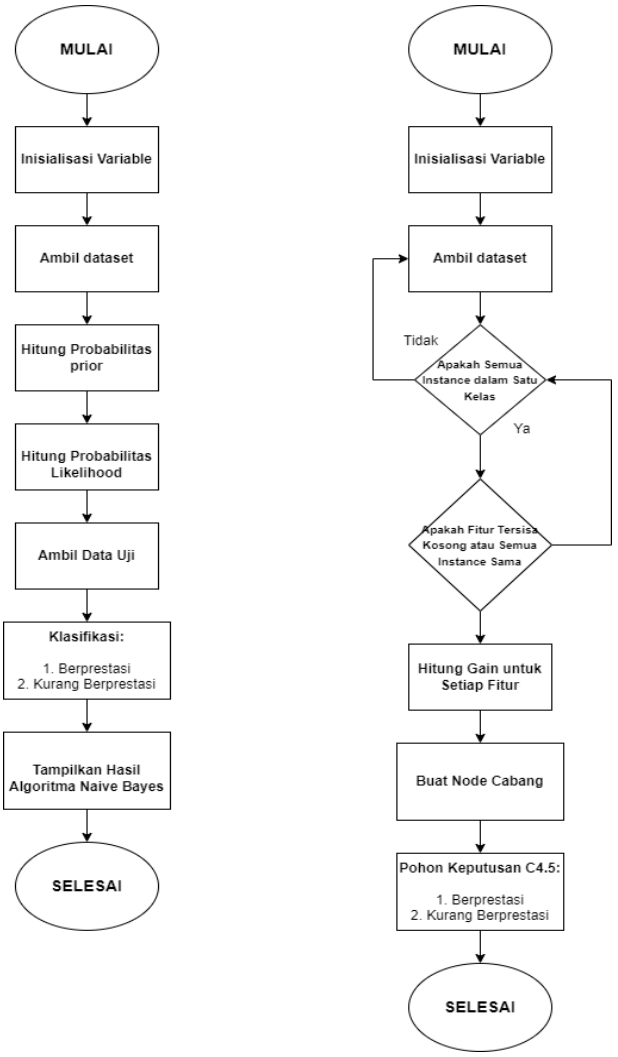
4. Penerapan Algoritma C4.5 dan Naïve Bayes

Algoritma C4.5 dan Naïve Bayes diterapkan untuk memprediksi prestasi siswa, sehingga hasil akhir yang diperoleh dapat digunakan sebagai prediksi pencapaian siswa.

5. Evaluasi

Tahap evaluasi dilakukan untuk mengukur akurasi sistem yang dikembangkan, guna memastikan prediksi yang dihasilkan sesuai dengan tujuan penelitian.

3.4 Flowchart C.45 dan Naive Bayes



Gambar 3.2 Flowchart Penelitian

Keterangan :

1. Algoritma Naive Bayes

   Bagian kiri diagram mewakili algoritma Naive Bayes. Algoritma ini bekerja berdasarkan prinsip probabilitas dan asumsi bahwa fitur-fitur dalam data adalah saling independen.

   - Inisialisasi Variabel: Menyiapkan variabel-variabel yang akan digunakan dalam perhitungan.
   - Ambil Dataset: Memuat data yang akan digunakan untuk melatih model.

- Hitung Probabilitas Prior: Menghitung probabilitas awal dari setiap kelas.
- Hitung Probabilitas Likelihood: Menghitung probabilitas kemunculan suatu fitur pada suatu kelas.
- Ambil Data Uji: Memuat data baru yang belum pernah dilihat oleh model untuk diprediksi kelasnya.
- Klasifikasi: Mengklasifikasikan data uji berdasarkan perhitungan probabilitas yang telah dilakukan.
- Tampilkan Hasil: Menampilkan hasil klasifikasi.

2. Algoritma C4.5

Bagian kanan dari diagram mewakili algoritma C4.5, yang merupakan salah satu metode pohon keputusan. Algoritma ini membangun pohon keputusan dengan cara membagi data secara rekursif berdasarkan fitur yang paling informatif.

- Inisialisasi Variabel: Menyiapkan variabel-variabel yang akan digunakan dalam perhitungan.
- Ambil Dataset: Memuat data yang akan digunakan untuk melatih model.
- Hitung Gain untuk Setiap Fitur: Menghitung informasi gain dari setiap fitur untuk menentukan fitur terbaik untuk membuat pemisahan data.
- Buat Node Cabang: Membagi data menjadi beberapa cabang berdasarkan nilai fitur yang terpilih.
- Ulangi proses hingga semua data terklasifikasi atau mencapai kriteria penghentian lainnya.

3.5 Lokasi dan Waktu Penelitian

3.5.1 Lokasi Penelitian

Penelitian ini dilakukan di SMA Asuhan Daya Medan, yang berlokasi di Kota Medan. Lokasi penelitian dipilih karena belum ada studi sebelumnya yang membandingkan algoritma C4.5 dan Naive Bayes untuk memprediksi prestasi siswa di sekolah tersebut, sehingga penelitian ini diharapkan dapat memberikan kontribusi baru.

3.5.2 Waktu Penelitian

Proses penelitian ini membutuhkan waktu selama 6 bulan dimulai dari Februari sampai dengan Juli 2024.

3.6 Instrumen Penelitian

Beberapa perangkat yang digunakan untuk mengerjakan tugas akhir ini adalah sebagai berikut :

1. Perangkat Lunak

   Dalam melakukan penelitian, peneliti menggunakan beberapa perangkat lunak berikut :

   a. Sistem operasi yang digunakan adalah Microsoft Windows 10 Profesional
   b. Aplikasi google colab untuk memproses data dan menulis code program.

2. Perangkat Keras

   Beberapa perangkat keras yang dibutuhkan peneliti dalam melakukan penelitian adalah sebagai berikut :

   a. Prosesor yang digunakan adalah Intel Core i3 2.0 Ghz
   b. RAM dengan ukuran 4GB

c. Solid State Drive 128GB

3. Perangkat Keras

Beberapa perangkat keras yang dibutuhkan peneliti dalam melakukan penelitian adalah sebagai berikut :

a. Prosesor yang digunakan adalah Intel Core i3 2.0 Ghz
b. RAM dengan ukuran 4GB
c. Solid State Drive 128GB

BAB IV
PEMBAHASAN

4.1 Pengumpulan Data

Data yang digunakan diambil dari nilai rapor semester satu sebanyak 28 siswa, yang terbagi dalam dua sheet pada file Excel: "NILAI SEM-1" dan "SMA IPA".

Tabel 4.1 Contoh Dataset Nilai Semester Satu

No.	Nama Kolom	Deskripsi
1.	Urut	Nomor urut siswa SMA Asuhan Daya
2.	NIS	Nomor induk siswa atau siswi di sekolah SMA Asuhan Daya
3.	NISN	Nomor induk siswa atau siswi di sekolah SMA Asuhan Daya
4.	Nama Siswa	Nama siswa atau siswi di SMA Asuhan Daya
5.	L/P	Jenis kelamin siswa atau siswi SMA Asuhan Daya
6.	Nilai Rapot	Nilai rapot siswa atau siswi di SMA Asuhan Daya yang terdiri dari nilai mata pelajaran PAI dan Budi Pekerti, PKN, Bahasa Indonesia, Matematika, Sejarah Indonesia, Bahasa Inggris, Seni Budaya, PJOK, Prakarya, Conversation, Qir'ah Qur'an, Matematika (Peminatan), Biologi, Fisika, Kimia, Nilai Rata-rata.

Pada Tabel 4.1 adalah contoh dari dataset pada nama sheet NILAI SEM-1 yang terdiri dari enam nama kolom.

Tabel 4.2 Contoh Dataset SMA IPA

No.	Nama Kolom	Deskripsi
1.	No	Nomor urutan pada dataset
2.	Nama	Nama siswa atau siswi di SMA Asuhan Daya
3.	Tempat/Tanggal Lahir	Tempat atau tanggal lahir siswa atau siswi SMA Asuhan Daya
4.	NIS	Nomor induk siswa atau siswi di SMA Asuhan Daya
5.	NISN	Nomor induk siswa atau siswi nasional

No.	Nama Kolom	Deskripsi
6.	Orang Tua	Nama orang tua siswa atau siswi SMA Asuhan Daya
7.	Pekerjaan Orang Tua	Profesi atau pekerjaan orang tua siswa atau siswi SMA Asuhan Daya
8.	Penghasilan Orang Tua	Jumlah penghasilan orang tua siswa atau siswi di SMA Asuhan Daya
9.	Absensi	Jumlah absensi atau kehadiran siswa atau siswi SMA Asuhan Daya

Pada Tabel 4.2 tersebut merupakan contoh dari dataset pada nama *sheet* SMA IPA dengan jumlah kolom sebanyak sembilan kolom.

4.2 Penerapan Algoritma C4.5 dan Naive Bayes

4.2.1 *Preprocessing* Data

Sebelum menerapkan algoritma C4.5 dan Naive Bayes, data perlu melalui tahap preprocessing untuk memastikan kualitas dan kesiapan data dalam analisis dan pemodelan. Pada tahap ini, data dari file Excel dimuat dan diperiksa untuk integritasnya. Dataset yang akan diprepros memiliki beberapa kolom dari sheet "SMA IPA," seperti NIS, penghasilan orang tua, dan jumlah kehadiran, sementara kolom nilai diambil dari sheet "NILAI SEM-1," sebagaimana terlihat pada Tabel 4.3.

Tabel 4.3 Dataset Siap di *Preprocessing*

No.	Nama	NIS	Penghasilan Orang Tua	Jumlah Kehadiran	Nilai
1	AMANDA DEFINA	977	Rp. 1,000,000 - Rp. 1,999,999	26	84,53
2	CINDY CELCEA	981	Rp. 1,000,000 - Rp. 1,999,999	26	83,73
3	DWI APRILIA SUNDARI	989	Rp. 1,000,000 - Rp. 1,999,999	26	84,80
4	EKA FITRIANA	990	Rp. 1,000,000 - Rp. 1,999,999	24	84,40

No.	Nama	NIS	Penghasilan Orang Tua	Jumlah Kehadiran	Nilai
5	FADLY SURYA PRANATA	991	Rp. 500,000 - Rp. 999,999	26	82,40
6	FAJAR	992	Rp. 1,000,000 - Rp. 1,999,999	26	82,53
7	FANY RAVINA	993	Rp. 500,000 - Rp. 999,999	25	83,20
8	FARA HAMIDAH	994	Rp. 500,000 - Rp. 999,999	26	84,13
9	FEBRI ANSYAH	995	Rp. 2,000,000 - Rp. 4,999,999	25	83,27
10	FIKA SONTRIANI LUBIS	996	Rp. 1,000,000 - Rp. 1,999,999	26	84,87
11	FITRI HANDAYANI LUBIS	997	Rp. 500,000 - Rp. 999,999	26	84,87
12	IRDA OKTAVIA	999	Tidak Berpenghasilan	26	84,00
13	LATIFAH UMA	1000	Rp. 1,000,000 - Rp. 1,999,999	26	83,80
14	MAULINDA APRIANI	1003	Rp. 2,000,000 - Rp. 4,999,999	26	83,67
15	MUHAMMAD SYAHPUTRA	1005	Rp. 1,000,000 - Rp. 1,999,999	25	83,47
16	MONA APRILIA	1007	Rp. 1,000,000 - Rp. 1,999,999	25	84,27
17	M. ILHAM FAHRUDIN	1010	Kurang dari Rp. 500,000	26	83,47
18	MHD. SYAHPUTRA	1010	Rp. 1,000,000 - Rp. 1,999,999	26	82,93
19	RESVI AULIA	1020	Rp. 2,000,000 - Rp. 4,999,999	26	83,67
20	RIONALDO FEBRIANSYAH	1021	Rp. 500,000 - Rp. 999,999	25	81,67
21	VINA LESTARI	1024	Rp. 1,000,000 - Rp. 1,999,999	26	84,00
22	WINDA KHAIRANI NST	1026	Rp. 500,000 - Rp. 999,999	26	86,20
23	ARINI FEBIOLA	1079	Rp. 1,000,000 - Rp. 1,999,999	26	84,07
24	MHD. ILHAM SAPUTRA	1080	Rp. 1,000,000 - Rp. 1,999,999	26	83,80
25	AHMAD HINDRA BERUTU	1083	Rp. 1,000,000 - Rp. 1,999,999	25	81,07

No.	Nama	NIS	Penghasilan Orang Tua	Jumlah Kehadiran	Nilai
26	SHEIRLLA CHANTIQA	1084	Rp. 500,000 - Rp. 999,999	25	85,60
27	NISA ULFITRI	1085	Rp. 500,000 - Rp. 999,999	26	85,07
28	STEFANI NERTIANA R. HUTAPEA	1134	Rp. 500,000 - Rp. 999,999	23	80,80

Pada Tabel 4.3 untuk kolom penghasilan orang tua yang pertama memiliki nilai penghasilan Rp. 1,000,000 - Rp. 1,999,999 dengan jumlah empat belas orang tua siswa, penghasilan kedua yaitu Rp. 500,000 - Rp. 999,999 dengan jumlah sembilan orang tua siswa, penghasilan ketiga Rp. 2,000,000 - Rp. 4,999,999 dengan jumlah tiga orang tua siswa, penghasilan keempat yaitu Tidak Berpenghasilan dengan jumlah satu orang tua siswa, penghasilan kelima yaitu memiliki nilai Kurang dari Rp. 500,000 dengan jumlah satu orang tua siswa.

Tabel 4.4 Mengubah *Unique Value* Kolom Penghasilan Orang Tua

Sebelum Diubah Menjadi Label	Sesudah Diubah Menjadi Label
Tidak Berpenghasilan	0
Kurang dari Rp. 500,000	1
Rp. 500,000 - Rp. 999,999	2
Rp. 1,000,000 - Rp. 1,999,999	3
Rp. 2,000,000 - Rp. 4,999,999	4

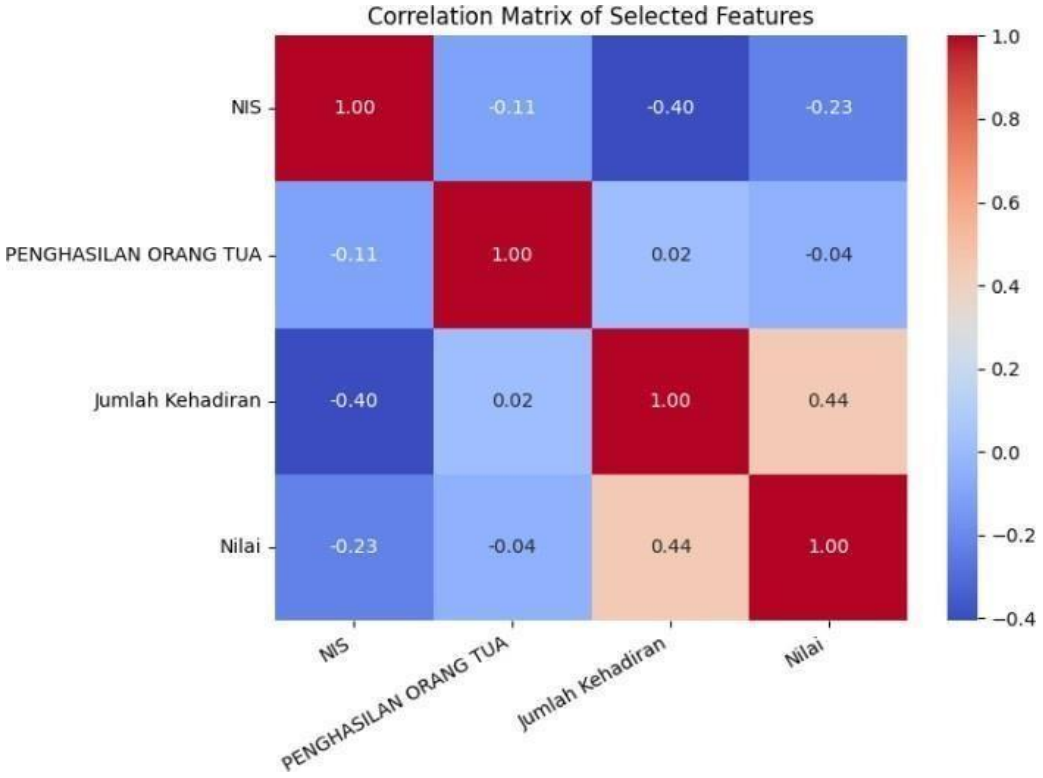
Pada Tabel 4.4 adalah proses mengubah *unique* value menjadi label pada kolom penghasilan orang tua.

Tabel 4.5 Pelabelan Siswa Berprestasi

Siswa Berprestasi	Jumlah
False	25

True	3

Pada Tabel 4.5 adalah proses pelabelan untuk siswa atau siswi berprestasi. Pelabelan dilakukan dengan syarat jika nilai rata-rata siswa atau siswi lebih dari 85 maka siswa atau siswi tersebut berprestasi. Nilai *False* artinya siswa atau siswi tersebut kurang berprestasi dengan jumlah siswa sebanyak 25 orang, kemudian nilai *True* adalah siswa atau siswi yang berprestasi dengan jumlah siswa atau siswi sebanyak tiga orang.



Gambar 4.1 *Correlation Matrix*

Pada Gambar 4.1 adalah matriks korelasi atau hubungan antar variabel atau kolom seperti NIS, penghasilan orang tua, jumlah kehadiran, dan nilai. Jika nilai korelasi matriks nya positif mendekati satu maka variabel tersebut memiliki keterkaitan dengan variabel lainnya. Jika nilai variabelnya negatif mendekati satu maka tidak ada keterkaitan dengan variabel lainnya. Berikut adalah nilai korelasi matriks untuk siswa atau siswi berprestasi:

1. Jumlah Kehadiran dan Nilai memiliki korelasi positif terkuat (0,44), yang

berarti bahwa kehadiran berkaitan dengan nilai yang lebih baik.

2. NIS dan Jumlah Kehadiran memiliki korelasi negatif sedang (-0,40), menunjukkan bahwa variabel NIS memiliki hubungan terbalik dengan kehadiran.
3. Penghasilan orang tua menunjukkan korelasi yang sangat lemah atau tidak ada dengan variabel lain, menunjukkan bahwa penghasilan orang tua tidak banyak mempengaruhi kehadiran atau nilai siswa dalam dataset ini.

4.2.2 *Splitting* Dataset

Setelah tahap *preprocessing* data selesai, dataset kemudian dibagi menjadi dua bagian utama yaitu data pelatihan *(training set)* dan data pengujian *(test set).* Pembagian ini dilakukan untuk memastikan bahwa model yang akan dibangun dapat dievaluasi secara objektif dan tidak *overfitting* terhadap data pelatihan. Proses *splitting* ini menggunakan fungsi *train_test_split* dari *library scikit-learn*, dengan proporsi 70% untuk data pelatihan dan 30% untuk data pengujian.

Pembagian dataset dilakukan dengan parameter *random_state*=42 untuk memastikan konsistensi hasil, di mana pemisahan data dapat diulang dengan hasil yang sama setiap kali kode dijalankan. Data pelatihan digunakan untuk melatih model prediksi, sementara data pengujian digunakan untuk mengevaluasi performa model tersebut pada data yang belum pernah dilihat sebelumnya.

Setelah preprocessing, dataset dibagi menjadi data pelatihan (training set) dan data pengujian (test set) untuk menghindari overfitting dan memastikan evaluasi objektif. Pembagian ini dilakukan menggunakan fungsi train_test_split dari library scikit-learn dengan proporsi 70% untuk pelatihan dan 30% untuk pengujian, serta parameter random_state=42 untuk hasil yang konsisten. Data pelatihan mencakup 70% dari total dataset (19 data) untuk melatih model,

sementara 30% sisanya (9 data) digunakan sebagai data pengujian.

Proporsi 70/30 ini dipilih untuk menjaga keseimbangan data pelatihan dan pengujian, sehingga model memiliki cukup informasi untuk belajar dan generalisasi. Pembagian ini penting agar model dapat memberikan prediksi yang akurat dan andal pada data baru yang tidak pernah dilihat selama pelatihan, sekaligus memungkinkan evaluasi model secara objektif.

4.2.3 Penerapan Algoritma C4.5

Untuk membangun model prediksi siswa berprestasi di SMA Asuhan Daya, fitur-fitur yang relevan untuk algoritma C4.5 adalah nilai, jumlah kehadiran, dan penghasilan orang tua. Algoritma ini dipilih karena kemampuannya menangani data dengan fitur numerik dan kategorikal, serta kemudahannya menghasilkan pohon keputusan yang dapat diinterpretasikan. Berikut adalah langkah-langkah penerapan algoritma C4.5 dalam memprediksi siswa berprestasi di SMA Asuhan Daya:

a. Pembentukan Label Target

Langkah pertama dalam penerapan C4.5 adalah menentukan label target, yaitu "siswa berprestasi". Dalam konteks ini, siswa dianggap berprestasi jika memiliki nilai akhir di atas ambang batas tertentu, misalnya 85. Siswa yang memenuhi kriteria ini diberi label 1 (Berprestasi), sedangkan siswa lainnya diberi label 0 (Kurang Berprestasi).

b. Pemilihan Fitur

Algoritma C4.5 kemudian mengevaluasi fitur-fitur yang tersedia untuk menentukan fitur mana yang paling informatif dalam memisahkan kelas target. Fitur-fitur seperti "Nilai", "Jumlah Kehadiran", dan "Penghasilan Orang Tua" dianalisis untuk mengukur sejauh mana mereka dapat mengurangi ketidakpastian

(entropy) dalam dataset.

c. Perhitungan Entropy dan Information Gain

C4.5 menghitung entropi awal dataset, yang mencerminkan tingkat ketidakpastian atau keragaman kelas dalam data. Selanjutnya, algoritma menghitung information gain untuk setiap fitur, yaitu sejauh mana fitur tersebut dapat mengurangi entropi ketika dataset dibagi berdasarkan nilai fitur tersebut. Fitur dengan information gain tertinggi akan dipilih sebagai node akar (root node) dari pohon keputusan.

a). Menghitung entropi awal

Rumus entropi

$$\text{Entropy}(S) = -\left(p_{\text{Kurang Berprestasi}} \log_2(p_{\text{Kurang Berprestasi}}) + p_{\text{Berprestasi}} \log_2(p_{\text{Berprestasi}})\right)$$

Keterangan:

- $p_{\text{Kurang Berprestasi}} = \frac{18}{19}$
- $p_{\text{Berprestasi}} = \frac{1}{19}$

Perhitungan:

$$p_{\text{Kurang Berprestasi}} = \frac{18}{19} \approx 0.947$$

$$p_{\text{Berprestasi}} = \frac{1}{19} \approx 0.053$$

$$\text{Entropy}(S) = -\left(0.947 \times \log_2(0.947) + 0.053 \times \log_2(0.053)\right)$$

$$\log_2(0.947) \approx -0.080$$

$$\log_2(0.053) \approx -4.247$$

$$\text{Entropy}(S) = -\left(0.947 \times -0.080 + 0.053 \times -4.247\right)$$

$$\text{Entropy}(S) = -\left(-0.07576 + -0.22508\right) \approx 0.297$$

b). Pembagian dataset berdasarkan nilai

Dataset dibagi berdasarkan threshold ≤ 84.967. threshold tersebut menghasilkan dua subset yaitu subset kiri (Nilai ≤ 84.967) dengan 18

sampel siswa atau siswi kurang berprestasi. Kemudian subset kanan (Nilai > 84.967) dengan 1 sampel siswa atau siswi berprestasi.

c). Menghitung entropi untuk masing-masing subset

Subset kiri (Nilai ≤ 84.967):

$$\text{Entropy}(S_1) = 0$$

Karena pada subset kiri semua sampel yang berjumlah 18 sampel tidak ada keragaman dalam subset ini (semua sampel dari kelas yang sama), entropinya adalah 0.

Subset kanan (Nilai > 84.967):

$$\text{Entropy}(S_2) = 0$$

Pada subset kanan yang hanya memiliki satu sampel saja dan tidak ada keragaman maka entropinya adalah 0.

d). Menghitung information gain

Rumus information gain:

$$\text{IG}(S, \text{Nilai}) = \text{Entropy}(S) - \left(\frac{|S_1|}{|S|} \times \text{Entropy}(S_1) + \frac{|S_2|}{|S|} \times \text{Entropy}(S_2) \right)$$

- $|S_1| = 18$ (jumlah sampel dalam subset kiri)
- $|S_2| = 1$ (jumlah sampel dalam subset kanan)
- $|S| = 19$ (total sampel)

$$\text{Weighted Entropy} = \left(\frac{18}{19} \times 0 \right) + \left(\frac{1}{19} \times 0 \right) = 0$$

Maka information gain yang didapat adalah

$$\text{IG(S, Nilai)} = 0.297 - 0 = 0.297$$

e). Menghitung split information

Meskipun information gain sudah memberikan gambaran tentang efektivitas pembagian. Sebelum memasuki langkah perhitungan gain ratio yaitu mempertimbangkan split information.

$$\text{SplitInfo}(\text{Nilai}) = -\left(\frac{|S_1|}{|S|}\log_2\left(\frac{|S_1|}{|S|}\right) + \frac{|S_2|}{|S|}\log_2\left(\frac{|S_2|}{|S|}\right)\right)$$

$$\text{SplitInfo}(\text{Nilai}) = -\left(\frac{18}{19}\log_2\left(\frac{18}{19}\right) + \frac{1}{19}\log_2\left(\frac{1}{19}\right)\right)$$

$$\log_2\left(\frac{18}{19}\right) \approx -0.074$$

$$\log_2\left(\frac{1}{19}\right) \approx -4.247$$

$$\text{SplitInfo}(\text{Nilai}) = -\left(\frac{18}{19} \times -0.074 + \frac{1}{19} \times -4.247\right)$$

$$\text{SplitInfo}(\text{Nilai}) \approx -\left(0.070 + 0.224\right) \approx 0.294$$

Hasil perhitungan untuk split informationnya adalah 0.294.

d. Perhitungan Gain Ratio

Untuk mengatasi bias terhadap fitur dengan banyak nilai unik, C4.5 menggunakan gain ratio, yaitu rasio antara information gain dan split information. Split information mengukur seberapa merata pembagian dataset oleh fitur tertentu. Fitur dengan gain ratio tertinggi dipilih sebagai node utama dalam

$$\text{GainRatio}(S, \text{Nilai}) = \frac{\text{IG}(S, \text{Nilai})}{\text{SplitInfo}(\text{Nilai})} = \frac{0.297}{0.294} \approx 1.01$$

pohon keputusan.

e. Pembentukan Pohon Keputusan

Berdasarkan fitur yang terpilih, pohon keputusan dibentuk secara bertahap dengan memisahkan dataset di sepanjang node sampai semua sampel dalam node akhir (leaf nodes) memiliki kelas yang sama, atau tidak ada lagi gain ratio yang signifikan. Setiap node dalam pohon keputusan mewakili keputusan berdasarkan fitur tertentu, dan cabang-cabangnya mewakili hasil dari keputusan tersebut.

Gambar 4.2 Pohon Keputusan Algoritma C4.5

4.2.4 Penerapan Naive Bayes

Naive bayes bekerja berdasarkan Teorema Bayes, yang menggabungkan probabilitas prior (sebelum melihat data) dengan probabilitas kondisional (berdasarkan data yang diamati). Meskipun algoritma ini mengasumsikan bahwa semua fitur bersifat independen, yang jarang terjadi dalam praktik, asumsi ini menyederhanakan perhitungan dan membuat algoritma ini efisien. Berikut adalah langkah-langkah penerapan naïve bayes untuk prediksi siswa atau siswi berprestasi:

a. Menghitung Probabilitas Prior (Prior Probability)

Probabilitas prior dihitung sebagai proporsi dari jumlah siswa dalam setiap kategori. Misalkan dalam dataset terdapat 9 siswa, di mana 2 siswa adalah

$$P(\text{Berprestasi}) = \frac{2}{9} \approx 0.222$$

$$P(\text{Kurang Berprestasi}) = \frac{7}{9} \approx 0.778$$

"Berprestasi" dan 7 siswa adalah "Kurang Berprestasi".

b. Menghitung Probabilitas Kondisional (Likelihood)

Probabilitas kondisional untuk setiap fitur dihitung menggunakan distribusi Gaussian, yang memerlukan perhitungan rata-rata (μ) dan variansi (σ^2) untuk setiap kelas dengan nilai 85 dan kehadiran 24 maka perhitungan probabilitas

kondisionalnya adalah:

- Nilai rata-rata (μ) berprestasi = 88, variansi (σ^2) berprestasi = 2, kehadiran rata-rata (μ) berprestasi = 25, variansi (σ^2) kehadiran berprestasi = 1.

  Probabilitas kondisionalnya adalah:

$$P(\text{Nilai} = 85|\text{Berprestasi}) = \frac{1}{\sqrt{2\pi \times 2}} e^{-\frac{(85-88)^2}{2\times 2}}$$

$$P(\text{Nilai} = 85|\text{Berprestasi}) = \frac{1}{\sqrt{4\pi}} e^{-\frac{9}{4}} = \frac{1}{\sqrt{4\pi}} e^{-2.25}$$

$$P(\text{Nilai} = 85|\text{Berprestasi}) \approx \frac{1}{\sqrt{12.566}} \times 0.1054 \approx \frac{1}{3.5449} \times 0.1054 \approx 0.0297$$

$$P(\text{Kehadiran} = 24|\text{Berprestasi}) = \frac{1}{\sqrt{2\pi \times 1}} e^{-\frac{(24-25)^2}{2\times 1}}$$

$$P(\text{Kehadiran} = 24|\text{Berprestasi}) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}}$$

$$P(\text{Kehadiran} = 24|\text{Berprestasi}) \approx \frac{1}{2.5066} \times 0.6065 \approx 0.2419$$

- Nilai rata-rata (μ) kurang berprestasi = 83, variansi (σ2) berprestasi = 10, kehadiran rata-rata (μ) kurang berprestasi = 24, variansi (σ2) kehadiran kurang berprestasi = 5.

  Probabilitas kondisionalnya adalah:

$$P(\text{Nilai} = 85|\text{Kurang Berprestasi}) = \frac{1}{\sqrt{2\pi \times 10}} e^{-\frac{(85-83)^2}{2\times 10}}$$

$$P(\text{Nilai} = 85|\text{Kurang Berprestasi}) = \frac{1}{\sqrt{20\pi}} e^{-\frac{4}{20}} = \frac{1}{\sqrt{20\pi}} e^{-0.2}$$

$$P(\text{Nilai} = 85|\text{Kurang Berprestasi}) \approx \frac{1}{7.9267} \times 0.8187 \approx 0.1033$$

$$P(\text{Kehadiran} = 24|\text{Kurang Berprestasi}) = \frac{1}{\sqrt{2\pi \times 5}} e^{-\frac{(24-24)^2}{2\times 5}}$$

$$P(\text{Kehadiran} = 24|\text{Kurang Berprestasi}) = \frac{1}{\sqrt{10\pi}} e^{0}$$

$$P(\text{Kehadiran} = 24|\text{Kurang Berprestasi}) = \frac{1}{\sqrt{10\pi}} \approx \frac{1}{5.60499} \approx 0.1784$$

c. Menghitung Probabilitas Posterior

Probabilitas posterior dihitung dengan mengalikan probabilitas prior dengan probabilitas kondisional untuk semua fitur yang diberikan kelas tertentu. Berikut adalah perhitungan probabilitas posterior:

- Untuk label berprestasi:

  Rumus probabilitas posterior:

  P (Berprestasi | Nilai = 85, Kehadiran = 24) = P (Berprestasi) × P (Nilai = 85 | Berprestasi) × P (Kehadiran = 24 | Berprestasi)

  Perhitungannya adalah:

  P (Berprestasi | Nilai = 85, Kehadiran = 24) ≈ 0.222 × 0.0297 × 0.2419 ≈ 0.00159

- Untuk label kurang berprestasi:

  Rumus probabilitas posterior:

  P (Kurang Berprestasi | Nilai = 85, Kehadiran = 24) = P (Kurang Berprestasi) × P (Nilai = 85 | Kurang Berprestasi) × P (Kehadiran = 24 | Kurang Berprestasi)

  Perhitungannya adalah:

  P (Kurang Berprestasi | Nilai = 85, Kehadiran = 24) ≈ 0.778 × 0.1033 × 0.1784 ≈ 0.0143

d. Menentukan Kelas dengan Probabilitas Tertinggi

Setelah menghitung probabilitas posterior untuk kedua kelas didapatkan hasil:

- Probabilitas untuk kelas "Berprestasi": 0.00159
- Probabilitas untuk kelas "Kurang Berprestasi": 0.0143

Dengan demikian,

P (Kurang Berprestasi | Nilai = 85, Kehadiran = 24) > *P* (Berprestasi | Nilai = 5, Kehadiran = 24), maka prediksi untuk siswa tersebut adalah "Kurang Berprestasi".

4.3 Evaluasi

4.3.1 Evaluasi Model Algoritma C4.5

Setelah pohon keputusan terbentuk, model dievaluasi menggunakan data pengujian. Kinerja model diukur berdasarkan akurasi, precision, recall, dan F1-score untuk memastikan bahwa model dapat secara akurat mengklasifikasikan siswa ke dalam kategori "Berprestasi" dan "Kurang Berprestasi". Hasil evaluasi menunjukkan bagaimana model dapat diandalkan dalam memprediksi kinerja siswa berdasarkan fitur-fitur yang tersedia.

Tabel 4.6 Evaluasi F1-Score

	Precision	Recall	F1-score	Support
Kurang berprestasi	1.00	1.00	1.00	7
Berprestasi	1.00	1.00	1.00	2
Accuracy			1.00	9
Macro avg	1.00	1.00	1.00	9
Weighted avg	1.00	1.00	1.00	9

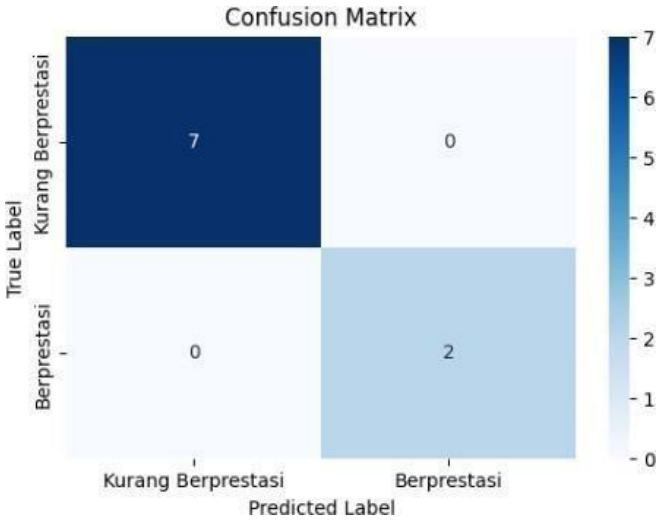
Penjelasan Tabel 4.6 untuk evaluasi dari F1-Score:

a). Accuracy: Model memiliki akurasi 1.00, yang berarti semua prediksi yang dilakukan oleh model adalah benar.

b). Macro Avg dan Weighted Avg: Kedua metrik ini juga bernilai 1.00, menunjukkan bahwa model memberikan performa yang sama baiknya pada kedua kelas, baik dalam rata-rata sederhana (macro avg) maupun dalam rata-rata yang memperhitungkan jumlah sampel di setiap kelas (weighted avg).

c). F1-Score menggabungkan precision dan recall menjadi satu metrik, dan dalam hal ini, nilai 1.00 menunjukkan bahwa model bekerja dengan sempurna, tidak melakukan kesalahan dalam prediksi untuk kedua kelas (Kurang Berprestasi dan Berprestasi).

d). Dengan nilai F1-Score yang sempurna untuk kedua kelas, ini berarti model C4.5 yang digunakan sangat efektif dalam memisahkan dan mengklasifikasikan siswa ke dalam kategori "Kurang Berprestasi" dan "Berprestasi" tanpa adanya kesalahan.



Gambar 4.3 Confusion Matrix Algoritma C4.5

Pada Gambar 4.3 adalah penjelasan dari hasil confusion matrix algoritma C4.5 yaitu untuk label sebenarnya dari “Kurang Berprestasi di prediksi benar oleh model sebanyak tujuh kali, sedangkan untuk label sebenarnya dari “Berprestasi” di prediksi benar oleh model sebanyak dua kali. Ini menunjukan hasil dari akurasi untuk prediksi siswa atau siswi di SMA Asuhan daya yang sempurna.

4.3.2 Evaluasi Model Naive Bayes

Setelah perhitungan selesai maka langkah selanjutnya adalah evaluasi model naïve bayes menggunakan f1-score, precision, recall, support.

Tabel 4.7 F1-Score naïve bayes

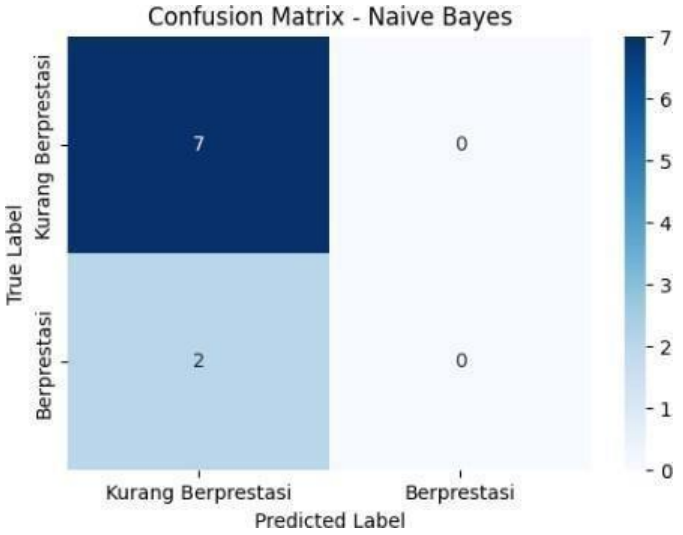
	Precision	Recall	F1-score	Support
Kurang berprestasi	0.78	1.00	0.88	7
Berprestasi	0.00	0.00	0.00	2
Accuracy			0.78	9
Macro avg	0.39	0.50	0.44	9
Weighted avg	0.60	0.78	0.68	9

Penjelasan Tabel 4.7 untuk evaluasi f1-score naïve bayes:

- Accuracy: Model memiliki akurasi 0.78, yang berarti 78% dari prediksi yang dilakukan oleh model adalah benar. Ini menunjukkan bahwa sebagian besar siswa diklasifikasikan dengan benar sebagai "Kurang Berprestasi" atau "Berprestasi", meskipun terdapat kelemahan dalam mendeteksi siswa yang "Berprestasi".
- Macro Avg: Precision 0.39, Recall 0.50, dan F1-Score 0.44 menunjukkan

bahwa ketika menghitung rata-rata sederhana dari metrik ini untuk kedua kelas, model memiliki performa yang kurang optimal, terutama pada kelas "Berprestasi" yang memiliki precision dan recall sangat rendah.

- Weighted Avg: Precision 0.60, Recall 0.78, dan F1-Score 0.68 menunjukkan bahwa ketika mempertimbangkan proporsi jumlah sampel di setiap kelas, model lebih cenderung memberikan performa yang lebih baik pada kelas dengan jumlah sampel yang lebih besar, yaitu "Kurang Berprestasi".
- F1-Score menggabungkan precision dan recall menjadi satu metrik. Dalam hal ini, F1-Score untuk kelas "Kurang Berprestasi" adalah 0.88, yang menunjukkan bahwa model cukup baik dalam mendeteksi siswa "Kurang Berprestasi". Namun, F1-Score untuk kelas "Berprestasi" adalah 0.00, menunjukkan bahwa model tidak mampu mengklasifikasikan siswa dalam kategori ini dengan benar.
- Dengan F1-Score yang sempurna untuk "Kurang Berprestasi" tetapi nol untuk "Berprestasi", ini berarti model Naive Bayes yang digunakan sangat efektif dalam mendeteksi siswa "Kurang Berprestasi" namun gagal dalam mengenali siswa "Berprestasi". Hal ini mungkin terjadi karena ketidakseimbangan dalam jumlah sampel antara kedua kelas atau karena fitur yang digunakan tidak cukup kuat untuk membedakan kedua kelas dengan baik.



Gambar 4.4 Confusion Matrix Naive Bayes

Penjelasan pada Gambar 4.4 tentang confusion matrix naive bayes adalah:

a. jumlah siswa yang sebenarnya "Kurang Berprestasi" dan berhasil diprediksi dengan benar oleh model sebagai "Kurang Berprestasi". Dari total 7 siswa yang benar-benar "Kurang Berprestasi", model memprediksi semua dengan benar.
b. jumlah siswa yang sebenarnya "Berprestasi" dan berhasil diprediksi dengan benar oleh model sebagai "Berprestasi". Namun, dalam hal ini, tidak ada siswa yang diprediksi dengan benar sebagai "Berprestasi" oleh model.
c. jumlah siswa yang diprediksi oleh model sebagai "Berprestasi" tetapi sebenarnya "Kurang Berprestasi". Tidak ada kesalahan prediksi dalam hal ini.
d. jumlah siswa yang sebenarnya "Berprestasi" tetapi diprediksi oleh model sebagai "Kurang Berprestasi". Dari 2 siswa yang benar-benar "Berprestasi", model salah memprediksi keduanya sebagai "Kurang Berprestasi"

Model bekerja dengan sangat baik dalam mengklasifikasikan siswa "Kurang Berprestasi". Semua siswa yang termasuk dalam kategori ini diprediksi dengan benar, yang menunjukkan bahwa model cukup efektif untuk kelas ini. Akan tetapi model gagal sepenuhnya dalam mendeteksi siswa "Berprestasi". Kedua siswa

yang sebenarnya "Berprestasi" diklasifikasikan sebagai "Kurang Berprestasi". Ini menunjukkan kelemahan signifikan dalam kemampuan model untuk mengenali kelas "Berprestasi”.

Confusion matrix ini menggarisbawahi bias model terhadap kelas "Kurang Berprestasi", yang mungkin disebabkan oleh ketidakseimbangan kelas (lebih banyak siswa "Kurang Berprestasi" dalam dataset) atau oleh fitur yang kurang efektif dalam membedakan antara kedua kelas.

BAB V
KESIMPULAN

5.1 Kesimpulan

Kesimpulannya, model algoritma C4.5 menunjukkan performa yang sangat baik dalam mengklasifikasikan siswa ke dalam kategori "Kurang Berprestasi" dan "Berprestasi", dengan akurasi dan F1-Score yang sempurna untuk kedua kelas. Algoritma C4.5 mampu memisahkan data dengan efektif, menghasilkan pohon keputusan yang jelas dan dapat diinterpretasikan dengan baik, serta tidak mengalami kesalahan dalam prediksi. Model ini sangat efektif dalam memanfaatkan fitur yang ada untuk membedakan antara kedua kelas, yang membuatnya menjadi pilihan yang kuat untuk tugas klasifikasi dalam konteks ini.

Di sisi lain, model Naive Bayes menunjukkan hasil yang kurang optimal, terutama dalam mengenali siswa "Berprestasi". Meskipun model ini berhasil memprediksi semua siswa "Kurang Berprestasi" dengan benar, ia gagal total dalam mendeteksi siswa "Berprestasi", yang terlihat dari F1-Score yang nol untuk kelas tersebut. Hal ini mungkin disebabkan oleh asumsi independensi antar fitur yang tidak sepenuhnya berlaku dalam data ini, serta ketidakseimbangan jumlah siswa antara kedua kelas. Secara keseluruhan, sementara C4.5 menunjukkan kinerja yang sangat baik, Naive Bayes perlu dioptimalkan lebih lanjut untuk meningkatkan kemampuannya dalam mengklasifikasikan kedua kelas secara akurat.

5.2 Saran

Tingkatkan performa model dengan menangani ketidakseimbangan kelas, menambahkan fitur relevan, dan menggunakan oversampling atau SMOTE

untuk Naive Bayes. Sederhanakan pohon keputusan C4.5 melalui pruning dan validasi k-fold, serta pertimbangkan penambahan data latih untuk meningkatkan akurasi dan keandalan klasifikasi.

DAFTAR PUSTAKA

Alfarizi, M. R. S., Al-farish, M. Z., Taufiqurrahman, M., Ardiansah, G., & Elgar, M. (2023). Penggunaan Python Sebagai Bahasa Pemrograman untuk Machine Learning dan Deep Learning. *Karya Ilmiah Mahasiswa Bertauhid (KARIMAH TAUHID)*, *2*(1), 1–6.

Andini, Y., Hardinata, J. T., Purba, Y. P., Studi, P., Informasi, S., Utara, S., & Apriori, M. (2022). Penerapan Data Mining Terhadap Tata Letak Buku. *Jurnal Technology Informatics & Computer System*, *XI*(1), 9–15.

Br Sembiring, S. N., Winata, H., & Kusnasari, S. (2022). Pengelompokan Prestasi Siswa Menggunakan Algoritma K-Means. *Jurnal Sistem Informasi Triguna Dharma (JURSI TGD)*, *1*(1), 31. https://doi.org/10.53513/jursi.v1i1.4784

Dewi, S. P., Nurwati, N., & Rahayu, E. (2022). Penerapan Data Mining Untuk Prediksi Penjualan Produk Terlaris Menggunakan Metode K-Nearest Neighbor. *Building of Informatics, Technology and Science (BITS)*, *3*(4), 639–648. https://doi.org/10.47065/bits.v3i4.1408

Hana, F. M. (2020). Klasifikasi Penderita Penyakit Diabetes Menggunakan Algoritma Decision Tree C4.5. *Jurnal SISKOM-KB (Sistem Komputer Dan Kecerdasan Buatan)*, *4*(1), 32–39. https://doi.org/10.47970/siskom-kb.v4i1.173

Indahsari, G. J. F., Kasiliyani, A., & ... (2021). Sistem Pengambilan Keputusan Beban Kinerja Menggunakan Naive Bayes Studi Kasus Pdam Bandarmasih. *... Terapan Riset Inovatif …*, 571–581.

Karim, A., Darma, U. B., Purnama, I., Labuhanbatu, U., Harahap, S. Z., & Labuhanbatu, U. (2021). *OR* (Issue January).

Kawani, G. P. (2019). Implementasi Naive Bayes. *Journal of Informatics, Information System, Software Engineering and Applications (INISTA)*, *1*(2), 73–81. https://doi.org/10.20895/inista.v1i2.73

Kusumawati, K. (2023). Pemanfaatan Teknologi Informasi Dalam Pendidikan. *Jurnal Limits*, *5*(1), 7–14. https://doi.org/10.59134/jlmt.v5i1.311

Muharram, R. F., Suryadi, A., Raya, J., No, T., Gedong, K., Rebo, P., & Timur, J. (2022). Implementasi Artificial Intelligence untuk Deteksi Masker Secara Realtime dengan Tensorflow dan SSD MobileNet Berbasis Python. *Jurnal Widya*, *3*(2), 281–290. https://jurnal.amikwidyaloka.ac.id/index.php/awl

Noviriandini, A., & Nurajijah, N. (2019). Analisis Kinerja Algoritma C4.5 Dan Naïve Bayes Untuk Memprediksi Prestasi Siswa Sekolah Menengah Kejuruan. *JITK (Jurnal Ilmu Pengetahuan Dan Teknologi Komputer)*, *5*(1), 23–28.

https://doi.org/10.33480/jitk.v5i1.607

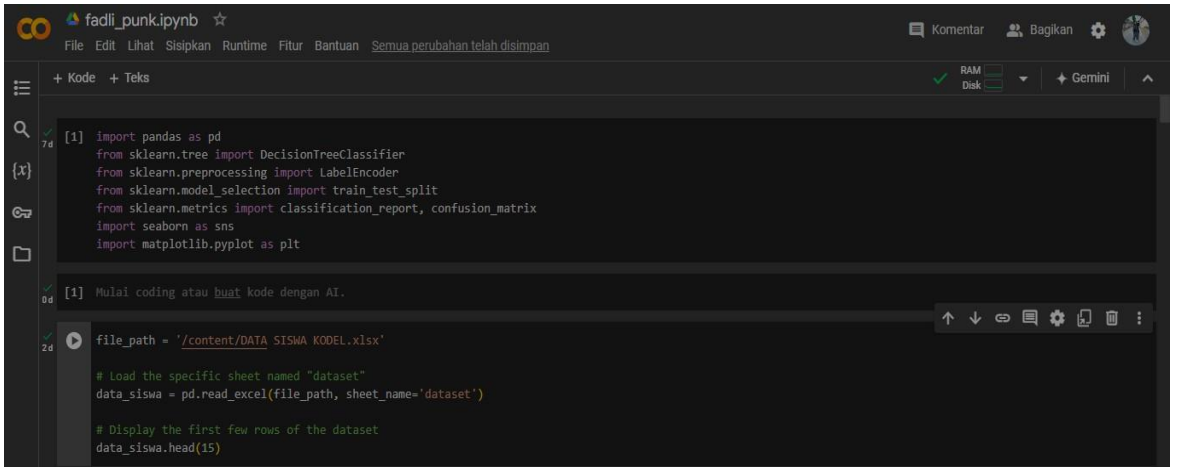
Rahmayanti, A., Rusdiana, L., & Suratno, S. (2022). Perbandingan Metode Algoritma C4.5 Dan Naïve Bayes Untuk Memprediksi Kelulusan Mahasiswa. *Walisongo Journal of Information Technology*, *4*(1), 11–22. https://doi.org/10.21580/wjit.2022.4.1.9654

Rambe, N. M. (2019). Peran Keluarga Dalam Meningkatkan Prestasi Belajar Siswa. *Prosiding Seminar Nasional Fakultas Ilmu Sosial Universitas Negeri Medan*, *3*, 930–934.

Romli, I., & Zy, A. T. (2020). Penentuan Jadwal Overtime Dengan Klasifikasi Data Karyawan Menggunakan Algoritma C4.5. *Jurnal Sains Komputer & Informatika (J-SAKTI*, *4*(2), 694–702.

Rovidatul, Yunus, Y., & Nurcahyo, G. W. (2023). Perbandingan algoritma c4.5 dan naive bayes dalam prediksi kelulusan mahasiswa. *Jurnal CoSciTech (Computer Science and Information Technology)*, *4*(1), 193–199. https://doi.org/10.37859/coscitech.v4i1.4755

LAMPIRAN


fadli_punk.ipynb
File Edit Lihat Sisipkan Runtime Fitur Bantuan Semua perubahan telah disimpan
Komentar
Bagikan
+ Kode + Teks
RAM
Disk
Gemini
[1] import pandas as pd
from sklearn.tree import DecisionTreeClassifier
from sklearn.preprocessing import LabelEncoder
from sklearn.model_selection import train_test_split
from sklearn.metrics import classification_report, confusion_matrix
import seaborn as sns
import matplotlib.pyplot as plt
[1] Mulai coding atau buat kode dengan AI.
file_path = '/content/DATA SISWA KODEL.xlsx'
# Load the specific sheet named "dataset"
data_siswa = pd.read_excel(file_path, sheet_name='dataset')
# Display the first few rows of the dataset
data_siswa.head(15)