JEPIN
(Jurnal Edukasi dan Penelitian Informatika)
ISSN(e): 2548-9364 / ISSN(p) : 2460-0741

Vol. 11
No. 2
Agustus
2025

Analisis Komparasi Metode *Gaussian Naïve Bayes* dan *Bernoulli Naïve Bayes* dalam Klasifikasi Pemilihan Program Peminatan Siswa SMA

Podang Binuryan[#1], Ni Made Satianingsih[#2], Chatarina Enny Murwaningtyas[#3]

[#]*Magister Pendidikan Matematika, Universitas Sanata Dharma*
Kampus III Universitas Sanata Dharma Paingan, Maguwoharjo, Depok, Yogyakarta
[1]1heaven.bassilica@gmail.com
[2]agustinanimadesatianingsih@gmail.com
[3]enny@usd.ac.id

Abstrak— **Penerapan Kurikulum Merdeka memberikan fleksibilitas kepada siswa dalam memilih program peminatan sesuai minat, bakat, dan kemampuan, namun menimbulkan tantangan dalam pemberian arahan yang tepat. Penelitian ini bertujuan membandingkan performa *Gaussian Naïve Bayes* dan Bernoulli Naïve Bayes dalam merekomendasikan program peminatan siswa di SMAS Katolik Santo Yoseph Denpasar. Model klasifikasi dibangun berdasarkan parameter seperti nilai akademik, jenis kelamin, preferensi pribadi, rencana jurusan, dan pengaruh orang tua. Dataset terdiri dari 449 baris data yang mencakup variabel numerik dan kategorikal, dengan variabel target berupa kelas terpilih (program peminatan siswa). *Gaussian Naïve Bayes* digunakan untuk data numerik, sementara Bernoulli Naïve Bayes diterapkan pada data yang telah dibinerisasi. Hasil studi menunjukkan bahwa *Gaussian Naïve Bayes* menghasilkan akurasi tertinggi (99%) untuk data numerik, tetapi performanya menurun saat ditambahkan variabel kategorikal. Sebaliknya, Bernoulli Naïve Bayes menunjukkan kinerja lebih stabil dengan akurasi 97%. Analisis learning curve menunjukkan bahwa kedua model memiliki kemampuan generalisasi yang baik, dengan Bernoulli lebih adaptif terhadap atribut biner. Penelitian ini menegaskan pentingnya pemilihan model yang sesuai dengan karakteristik data untuk mendukung pengambilan keputusan berbasis data di lingkungan pendidikan.**

Kata kunci— **Kurikulum Merdeka, *Gaussian Naïve Bayes*, *Bernoulli Naïve Bayes*, Program Peminatan Siswa SMA.**

I. PENDAHULUAN

Desain Kurikulum Merdeka memberikan fleksibilitas kepada peserta didik untuk memilih mata pelajaran berdasarkan minat, bakat, dan kemampuan mereka [1]. Kebijakan ini bertujuan mendukung pengembangan potensi individu secara optimal sekaligus mempersiapkan peserta didik menghadapi dunia kerja dan pendidikan tinggi [2]. Salah satu implementasi penting dari kebijakan ini adalah penghapusan sistem penjurusan yang sebelumnya mengarahkan siswa pada jalur IPA, IPS, atau Bahasa [3]. Sebagai gantinya, siswa diberikan kebebasan untuk memilih program peminatan yang sesuai dengan kebutuhan mereka [4]. Namun, kebijakan ini juga memunculkan tantangan baru, seperti kurangnya arahan dan bimbingan yang memadai bagi siswa untuk menentukan pilihan yang selaras dengan potensi mereka.

Keputusan peminatan yang tepat memiliki dampak signifikan terhadap kesuksesan jangka panjang siswa. Menurut Ghaleb [5], keselarasan antara pilihan karier dengan karakteristik individu, seperti minat, kemampuan, dan tujuan, meningkatkan kepuasan karier serta keberhasilan jangka panjang. Sebaliknya, ketidaksesuaian pilihan dapat memicu ketidakpuasan, perubahan karier, dan ketidakstabilan emosional. Oleh karena itu, penting untuk memberikan bimbingan yang strategis agar siswa dapat memilih program yang sesuai dengan potensi mereka.

Eksplorasi karier menjadi salah satu elemen penting dalam proses pemilihan peminatan. Sinring dan Umar [6] menekankan bahwa faktor seperti preferensi pribadi, pengaruh teman sebaya, dan media memainkan peran besar dalam membentuk identitas karier siswa generasi Z. Lingkungan sekolah yang mendukung dapat membantu siswa mengidentifikasi potensi diri mereka [7]. Dalam hal ini, Johnson dkk. [8] menambahkan bahwa kemampuan siswa untuk membuat keputusan secara mandiri, tanpa pengaruh emosional berlebihan, sangat penting dalam membangun kesadaran yang jelas tentang arah karier mereka. Sekolah memiliki peran strategis untuk mendukung siswa melalui proses ini dengan menyediakan bimbingan dan pendampingan yang relevan.

Meski demikian, praktik di lapangan menunjukkan bahwa kebebasan memilih mata pelajaran tidak selalu diiringi kematangan dalam pengambilan keputusan. Sebagian siswa masih cenderung memilih program peminatan hanya untuk menghindari mata pelajaran yang dianggap sulit, atau sekadar mengikuti tren teman sebaya. Di sisi lain, tekanan dari orang tua untuk memilih jalur tertentu juga menjadi kendala bagi siswa dalam

mengekspresikan minat dan bakat sebenarnya. Hal ini menimbulkan kesenjangan antara potensi akademik dan aspirasi karier siswa, sehingga diperlukan pendekatan komprehensif agar pemilihan jurusan benar-benar mempertimbangkan berbagai aspek, termasuk minat, bakat, dan daya dukung lingkungan.

Untuk menjawab tantangan tersebut, *Educational Data Mining* (EDM) menjadi relevan karena memungkinkan sekolah, guru, atau konselor menganalisis data siswa secara sistematis [9]. Ragam teknik seperti *clustering* dan klasifikasi membantu mengungkap pola-pola tersembunyi dalam data pendidikan, misalnya pola belajar, kinerja akademik, atau kecenderungan minat siswa [10]. Metode klasifikasi sering dimanfaatkan untuk memetakan siswa ke dalam kategori tertentu: mulai dari prediksi prestasi [11] identifikasi siswa berisiko putus sekolah [12], hingga pemberian intervensi khusus [13]. Dalam hal penjurusan, Roghib dkk. [14] menggunakan algoritma C4.5 untuk klasifikasi siswa di SMK Plus Al-Hilal Arjawinangun dan berhasil mencapai akurasi 98,02%. Hasil serupa ditemukan oleh Amanda dkk. [15] yang menerapkan algoritma CART di SMA Negeri 2 Perbaungan untuk memisahkan siswa ke jurusan IPA atau IPS, sehingga analisis data menjadi lebih efisien dan meminimalkan kesalahan manusia.

Sementara itu, Naïve Bayes menjadi salah satu algoritma berbasis probabilitas yang mulai banyak diaplikasikan dalam penjurusan siswa. Khafifah dkk. [16] menunjukkan bahwa pendekatan ini memiliki sensitivitas 75% dalam menentukan IPA, IPS, atau Bahasa berdasarkan nilai akademik, meski akurasi keseluruhan mencapai 55,55%. Ramadhani [17] menambahkan fungsi Gaussian pada Naïve Bayes untuk menangani atribut numerik seperti nilai Matematika, Fisika, Biologi, dan Kimia, yang terbukti meningkatkan ketepatan prediksi dalam memetakan siswa ke jurusan IPA dan IPS. Namun, penelitian terdahulu umumnya hanya menggunakan satu varian Naïve Bayes (Gaussian, Bernoulli, atau Multinomial), dan belum banyak yang secara khusus membandingkan performa Gaussian vs. Bernoulli dalam membantu pemilihan program peminatan siswa, terlebih di lingkungan Kurikulum Merdeka yang lebih fleksibel dan kompleks.

Berangkat dari pemaparan di atas, kesenjangan penelitian dapat dipetakan sebagai berikut. Pertama, masih terbatasnya studi yang secara eksplisit menyesuaikan proses klasifikasi peminatan dengan dinamika Kurikulum Merdeka berdasarkan tinjauan literatur yang relevan, terutama dalam konteks penggunaan data dari sekolah swasta dan pemodelan berbasis kombinasi atribut numerik dan kategorik. Kedua, kajian perbandingan antara Gaussian Naïve Bayes dan Bernoulli Naïve Bayes masih terbatas, padahal kedua varian ini berbeda dalam hal penanganan tipe atribut. Gaussian cocok untuk data numerik, sementara Bernoulli lebih efektif untuk data biner atau kategorik sederhana. Ketiga, belum ada penelitian serupa yang menyoroti SMAS Katolik Santo Yoseph Denpasar, yang tentu memiliki karakteristik siswa dan situasi sekolah swasta yang mungkin berbeda dari sekolah negeri atau SMK.

Oleh karena itu, penelitian ini berupaya membandingkan performa Gaussian Naïve Bayes dan Bernoulli Naïve Bayes dalam memberikan rekomendasi program peminatan di SMAS Katolik Santo Yoseph Denpasar. Pendekatan ini diharapkan dapat menjawab kebutuhan spesifik Kurikulum Merdeka di mana siswa membutuhkan arahan yang berbasis data, sekaligus mempertimbangkan perbedaan atribut numerik (misalnya nilai akademik) dan atribut kategorik/biner (misalnya minat, preferensi jurusan, atau indikator bakat). Diharapkan hasil penelitian ini dapat memperlihatkan metode mana yang lebih akurat dan efisien dalam memfasilitasi pemilihan peminatan, sehingga dapat mendukung pengambilan keputusan yang lebih matang oleh siswa, guru, maupun konselor. Dengan demikian, kebaruan yang diusung terletak pada perbandingan langsung antara dua varian Naïve Bayes di tengah konteks penerapan Kurikulum Merdeka, serta penerapannya pada sekolah swasta yang memiliki keragaman potensi dan tujuan akademik siswa.

II. Landasan Teori

A. Educational Data Mining (EDM)

Educational Data Mining (EDM) adalah proses penggalian informasi dan pola tersembunyi dari data pendidikan untuk mendukung pengambilan keputusan berbasis data. EDM sering digunakan untuk memahami perilaku siswa, meningkatkan efektivitas pembelajaran, dan mendukung bimbingan akademik. Dalam konteks peminatan siswa, EDM membantu menganalisis data akademik dan preferensi siswa untuk memberikan rekomendasi program peminatan yang sesuai.

Menurut Riyanto dan Ompusunggu [18], EDM terdiri dari berbagai tahapan, seperti *preprocessing data*, pemodelan, dan evaluasi. Tahapan ini memastikan data yang digunakan bersih, relevan, dan siap untuk dianalisis. Salah satu metode yang sering digunakan dalam EDM adalah Naïve Bayes, yang memungkinkan pengklasifikasian data siswa berdasarkan atribut tertentu untuk mendukung keputusan pendidikan. Supangat dan Giovanni [19] juga menyoroti peran EDM dalam seleksi penerimaan mahasiswa baru, yang menunjukkan bagaimana *data mining* dapat digunakan untuk memberikan hasil yang akurat dalam waktu yang singkat.

Dalam studi Asmoro dkk. [20], algoritma *data mining* seperti pohon keputusan digunakan untuk mengevaluasi kinerja akademik siswa. Studi ini relevan karena menunjukkan bagaimana teknik analitik dapat mendukung penentuan jalur pendidikan siswa. Di sisi lain, Apsari [21] menunjukkan bagaimana algoritma Naïve Bayes diterapkan untuk memprediksi prestasi siswa berdasarkan data historis mereka, mendukung penggunaan algoritma probabilistik dalam pendidikan.

Aplikasi EDM melibatkan berbagai metode seperti klasifikasi, *clustering*, dan analisis pola. Klasifikasi menjadi salah satu metode yang sering digunakan untuk memprediksi keputusan siswa, terutama melalui algoritma seperti Naïve Bayes. EDM memungkinkan integrasi data

akademik dan preferensi siswa sehingga hasil analisis menjadi lebih relevan dalam konteks pendidikan.

B. Konsep Naïve Bayes

Naïve Bayes adalah algoritma klasifikasi berbasis probabilitas yang didasarkan pada Teorema Bayes. Algoritma ini mengasumsikan independensi antar fitur dalam *dataset*. Rumus dasar untuk probabilitas posterior adalah sebagai berikut:

$$P(C_k \mid X) = \frac{P(X \mid C_k)P(C_k)}{P(X)} \quad (1)$$

dengan

$P(C_k \mid X)$:Probabilitas posterior kelas C_k dengan fitur X

$P(X \mid C_k)$:Probabilitas likelihood fitur X diberikan kelas C_k

$P(C_k)$: Probabilitas prior dari kelas C_k .

$P(X)$: Probabilitas total dari fitur X .

Naïve Bayes mengasumsikan bahwa setiap fitur X_i dalam X *independen* satu sama lain, sehingga probabilitas likelihood dapat ditulis sebagai:

$$P(X \mid C_k) = \prod_{i=1}^{n} P(X_i \mid C_k) \quad (2)$$

Naïve Bayes dikenal karena efisiensinya, terutama dalam menangani *dataset* besar. Dengan asumsi independensi, perhitungan probabilitas menjadi sederhana tetapi tetap efektif untuk banyak kasus praktis.

C. Gaussian Naïve Bayes

Gaussian Naïve Bayes adalah varian dari Naïve Bayes yang dirancang untuk menangani data numerik. Algoritma ini mengasumsikan bahwa fitur numerik mengikuti distribusi normal (Gaussian). Probabilitas likelihood untuk setiap fitur numerik x_i diberikan oleh:

$$P(x_i \mid C_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{(x_i - \mu_k)^2}{2\sigma_k^2}\right) \quad (3)$$

dengan

μ_k : Rata-rata nilai fitur x_i pada kelas C_k ,

σ_k^2 : Variansi nilai fitur x_i pada kelas C_k .

Dalam konteks penelitian ini, Gaussian Naïve Bayes digunakan untuk memproses nilai akademik siswa yang berupa data numerik. Model ini membantu mengidentifikasi pola performa siswa dan memetakan mereka ke program peminatan berdasarkan kemampuan akademik.

D. Bernoulli Naïve Bayes

Bernoulli Naïve Bayes adalah varian lain dari Naïve Bayes yang digunakan untuk data biner atau boolean. Algoritma ini menghitung probabilitas berdasarkan keberadaan atau ketiadaan fitur tertentu. Rumus likelihood untuk fitur biner x_j adalah:

$$P(x_j \mid C_k) = p_k^{x_j}(1 - p_k)^{1-x_j} \quad (4)$$

dengan

p_k : Probabilitas keberadaan fitur x_j dalam kelas C_k ,

x_j : Nilai biner fitur (0 atau 1).

Dalam penelitian ini, Bernoulli Naïve Bayes diterapkan pada fitur kategorikal seperti jenis kelamin, preferensi peminatan siswa dan pilihan orang tua/wali. Data kategorikal tersebut dikonversi menjadi format biner menggunakan metode *one-hot encoding* untuk memungkinkan analisis probabilistik.

E. Implementasi Gaussian dan Bernoulli Naïve Bayes dalam Klasifikasi Peminatan Siswa

Penelitian ini bertujuan untuk mengevaluasi performa dua metode Naïve Bayes, yaitu Gaussian dan Bernoulli, dalam mengklasifikasikan peminatan siswa berdasarkan data numerik dan kategorikal. Kedua metode digunakan secara independen, dengan asumsi bahwa data nilai akademik siswa memegang peranan utama dalam menentukan kelas peminatan, sedangkan data kategorikal, seperti jenis kelamin, preferensi siswa dan pilihan orang tua, memberikan konteks tambahan.

Gaussian Naïve Bayes diterapkan untuk data numerik seperti nilai akademik siswa. Algoritma ini mengasumsikan bahwa nilai akademik mengikuti distribusi normal, sehingga memungkinkan model untuk memanfaatkan pola data numerik secara optimal. Dalam penelitian Ramadhani dkk. [17], Gaussian Naïve Bayes terbukti efektif dalam memetakan siswa ke jurusan IPA atau IPS berdasarkan nilai akademik, yang menunjukkan relevansi pendekatan ini dalam pendidikan.

Bernoulli Naïve Bayes, di sisi lain, digunakan untuk data kategorikal yang dikonversi menjadi format biner. Karena variabel kategorikal tidak bersifat ordinal, metode *one-hot encoding* digunakan untuk memastikan setiap kategori direpresentasikan secara independen. Selain itu, penelitian ini mengaplikasikan pendekatan inovatif dengan mendiskritisasi data numerik berdasarkan rata-rata nilai untuk digunakan dalam Bernoulli Naïve Bayes. Proses ini menghasilkan data biner yang merepresentasi-kan apakah nilai siswa berada di atas atau di bawah rata-rata kelas.

Hasil dari kedua metode dibandingkan menggunakan metrik evaluasi seperti akurasi, presisi, *recall*, dan f1-score. Perbandingan ini bertujuan untuk mengevaluasi efektivitas masing-masing metode dalam mendukung pengambilan keputusan berbasis data di lingkungan pendidikan, khususnya dalam pemilihan program peminatan siswa.

F. Relevansi Teori dengan Penelitian

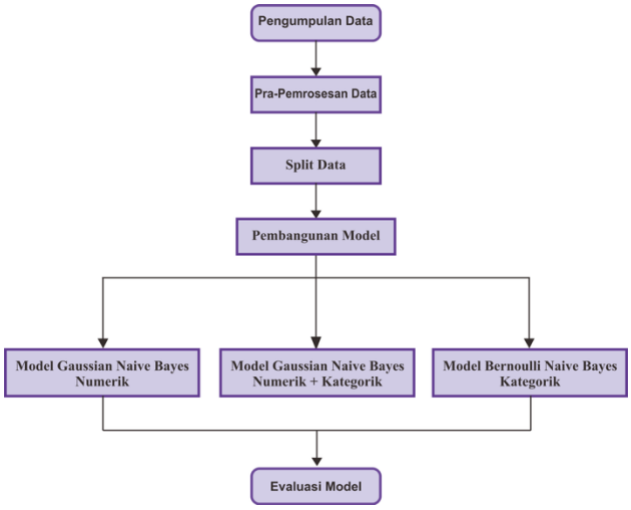
Landasan teori ini mendukung penelitian dengan menjelaskan bagaimana Gaussian dan Bernoulli Naïve Bayes digunakan dalam klasifikasi program peminatan siswa. Naïve Bayes, dengan pendekatan probabilistiknya, memungkinkan pengolahan data numerik dan biner secara efisien. Gaussian Naïve Bayes digunakan untuk memahami pola nilai akademik siswa, sedangkan Bernoulli Naïve Bayes digunakan untuk menganalisis preferensi siswa dan

pengaruh orang tua dalam pengambilan keputusan peminatan.

Penelitian ini memberikan kontribusi praktis dalam pengambilan keputusan berbasis data di sekolah, khususnya dalam mendukung siswa untuk memilih program peminatan yang sesuai dengan potensi dan minat mereka. Pendekatan ini tidak hanya meningkatkan akurasi klasifikasi tetapi juga membantu sekolah dalam merancang strategi pendidikan yang lebih efektif.

III. Metode Penelitian

Penelitian ini menggunakan algoritma Naïve Bayes untuk membangun model klasifikasi program peminatan siswa. Tahapan penelitian mencakup pengumpulan data, pra-pemrosesan data, pembangunan model, evaluasi performa, dan visualisasi hasil. Secara keseluruhan, tahapan metode penelitian tergambar dalam Gambar 1, yang memberikan alur sistematis dari awal hingga akhir penelitian.



Gambar. 1 Tahapan penelitian

A. *Pengumpulan Data*

Dataset yang digunakan dalam penelitian ini berasal dari data akademik siswa kelas X di SMAS Katolik Santo Yoseph Denpasar untuk tahun ajaran 2022–2023. *Dataset* ini mencakup berbagai variabel yang relevan dengan proses pemilihan program peminatan siswa. Data dikumpulkan dari catatan akademik sekolah dan kuesioner yang diisi oleh siswa serta orang tua/wali, mencakup nilai akademik, preferensi peminatan, dan informasi gender siswa. Secara keseluruhan, terdapat 449 data siswa yang tercatat di dalam *dataset*.

Enam kolom utama disertakan dalam *dataset* ini, yaitu nilai rata-rata support kelas (A–G), jenis kelamin, rencana program studi, pilihan kelas peserta didik, pilihan kelas orang tua/wali, dan kelas terpilih. Variabel target dalam penelitian ini adalah kelas terpilih, yang merefleksikan program peminatan atau jurusan yang dipilih oleh sistem berdasarkan analisis data prediktor. Variabel target tersebut bertujuan membantu memahami dan mengklasifikasikan minat akademik peserta didik sesuai berbagai faktor yang mempengaruhi keputusan mereka, sehingga mampu memberikan rekomendasi yang selaras dengan profil akademik dan minat mereka.

Kelas terpilih diukur sebagai variabel kategorikal yang mencerminkan jurusan yang tersedia di SMAS Katolik Santo Yoseph Denpasar, yaitu Kelas XI-A & XI-B (tenaga medis), Kelas XI-C & XI-D (teknik), Kelas XI-E & XI-F (akuntansi), serta Kelas XI-G (hubungan internasional). Adapun variabel prediktor yang dipakai meliputi nilai tiap mata pelajaran per semester, berjenis data kuantitatif, yang mencerminkan nilai rata-rata dan performa siswa dalam berbagai kategori kelas. Nilai rata-rata support kelas A hingga G diambil dari mata pelajaran sains untuk kelas A–C, mata pelajaran sosial untuk kelas D–F, dan mata pelajaran bahasa untuk kelas G.

Selain itu, rencana kelas program studi menjadi variabel fitur yang berjenis data kategorikal, berisi pilihan kelas jurusan yang mendukung cita-cita siswa ketika memasuki perguruan tinggi, digolongkan menjadi paket sains, sosial, dan komunikasi. Jurusan pada paket sains A dapat menunjang siswa yang memiliki rencana kuliah di kedokteran umum, kebidanan, teknologi pangan, kedokteran gigi, keperawatan, kimia, biologi, kelautan, farmasi, dokter hewan, arsitek, astronomi, perikanan, bioteknologi, ilmu kesehatan masyarakat, psikologi, teknik, serta sekolah kedinasan. Jurusan pada paket sains B dapat mendukung arsitek, sains kebumian, matematika, kedokteran gigi, kedokteran hewan, gizi, aktuaria, statistika, dan fisioterapi. Jurusan pada paket sains C dapat mengarahkan siswa ke desain, biofisika, informatika, geofisika, serta pendidikan guru matematika dan sains.

Paket sosial D ditujukan bagi siswa yang hendak mengambil administrasi bisnis, logistik, psikologi, hukum, ataupun bisnis digital. Paket sosial E ditujukan untuk filsafat, hukum, ilmu militer, akuntansi, dan manajemen. Paket sosial F mendukung humaniora, pendidikan guru ilmu sosial, ekonomi, serta ekonomi pembangunan. Sementara itu, paket komunikasi G mengarah ke bahasa Inggris, sastra Indonesia, linguistik, bahasa Mandarin, bahasa Jepang, pariwisata, DKV, hubungan internasional, ilmu komunikasi, dan seni.

Variabel jenis kelamin digunakan untuk mengidentifikasi apakah faktor gender memengaruhi pemilihan program peminatan. Pilihan kelas oleh peserta didik merupakan variabel kategori yang menggambarkan preferensi mereka berdasarkan minat dan bakat, sementara pilihan kelas oleh orang tua/wali mencerminkan preferensi keluarga terhadap masa depan pendidikan serta karier anak. Keseluruhan informasi tersebut kemudian digabungkan menjadi satu *dataset* terintegrasi.

B. *Pra-Pemrosesan Data*

Tahap pra-pemrosesan data dalam penelitian ini dirancang untuk menghasilkan *dataset* yang bersih, konsisten, dan siap digunakan dalam proses pemodelan. Langkah-langkahnya melibatkan penggabungan data,

pembersihan, transformasi, dan *encoding* yang disesuaikan untuk model Gaussian Naïve Bayes dan Bernoulli Naïve Bayes. Proses ini dilakukan secara hati-hati agar setiap langkah memastikan kualitas data dan meminimalkan risiko *data leakage* melalui pendekatan *pipeline*.

Dataset awal diperoleh dari catatan akademik siswa SMAS Katolik Santo Yoseph Denpasar tahun ajaran 2022–2023. Data ini mencakup informasi akademik, demografis, dan preferensi peminatan siswa, yang digabungkan dari berbagai sumber, seperti catatan nilai sekolah dan kuesioner yang diisi oleh siswa serta orang tua/wali. Setelah data digabungkan, dilakukan pemeriksa-an terhadap struktur dan kualitas data. Tidak ditemukan data duplikat atau entri yang perlu dihapus. Jumlah data tetap 449 baris seperti data awal, tanpa pengurangan observasi. Yang dihapus hanyalah beberapa kolom yang tidak relevan untuk proses pemodelan. Dengan demikian, pra-pemrosesan hanya mengubah tipe dan representasi variabel, bukan jumlah data.

Pada tahap berikutnya, *encoding* dilakukan untuk mengubah variabel kategorikal menjadi format numerik agar dapat digunakan dalam proses pemodelan. Variabel jenis kelamin, yang memiliki dua kategori (laki-laki dan perempuan), diubah menggunakan label *encoding*, dengan nilai 0 untuk laki-laki dan 1 untuk perempuan. Variabel non-ordinal seperti rencana jurusan, pilihan siswa, dan pilihan orang tua diubah menggunakan *one-hot encoding*, yang menciptakan kolom biner terpisah untuk setiap kategori. Teknik ini memastikan bahwa model tidak mengasumsikan hubungan ordinal antar kategori.

Khusus untuk model Bernoulli Naïve Bayes, variabel numerik seperti Nilai Support A–G diubah menjadi kategori biner berdasarkan rata-rata nilai. Nilai di atas rata-rata diklasifikasikan sebagai 1, sementara nilai di bawah atau sama dengan rata-rata diklasifikasikan sebagai 0. Proses binarisasi ini dilakukan untuk memastikan bahwa variabel numerik sesuai dengan format yang diperlukan oleh model Bernoulli.

Setelah langkah pembersihan dan *encoding*, *dataset* dibagi menjadi dua bagian: 70% untuk data latih dan 30% untuk data uji. Pembagian dilakukan menggunakan teknik *stratified splitting* untuk menjaga distribusi kelas target tetap seimbang antara data latih dan data uji. Setelah pembagian data, *pipeline* digunakan untuk menerapkan transformasi tambahan, seperti standarisasi variabel numerik menggunakan *Robust Scaler* untuk mengurangi pengaruh *outlier*. Selanjutnya, transformasi Yeo-Johnson diterapkan pada variabel numerik untuk mendekatkan distribusi data ke bentuk normal, yang sangat penting untuk mendukung performa optimal model Gaussian Naïve Bayes.

Penggunaan *pipeline* dalam proses pra-pemrosesan memberikan beberapa manfaat utama. *Pipeline* memastikan bahwa transformasi hanya dilakukan berdasarkan data latih selama *fitting*, sehingga mengurangi risiko data *leakage* yang dapat mengganggu validitas model. Selain itu, *pipeline* juga meningkatkan efisiensi dan mempermudah reproduksi proses, menjadikannya alat yang ideal untuk penelitian berbasis pembelajaran mesin. Dengan langkah-langkah ini, *dataset* yang telah diproses siap digunakan dalam pembangunan model klasifikasi Naïve Bayes.

C. Pembangunan Model

Pembangunan model dalam penelitian ini bertujuan untuk mengklasifikasikan pemilihan program peminatan siswa berdasarkan variabel prediktor yang telah diproses. Penelitian ini menggunakan algoritma Naïve Bayes, yang mencakup Gaussian Naïve Bayes dan Bernoulli Naïve Bayes, dengan mempertimbangkan karakteristik variabel prediktor yang tersedia. Meskipun terdapat pelanggaran asumsi distribusi Gaussian pada model tertentu, pendekatan ini tetap dilakukan untuk memperoleh wawasan tentang pengaruh berbagai variabel terhadap akurasi klasifikasi.

Model Gaussian Naïve Bayes diterapkan melalui dua pendekatan. Pada pendekatan pertama, hanya variabel numerik, yaitu Nilai Support A–G, yang digunakan sebagai prediktor. Variabel numerik tersebut sebelumnya telah melalui proses standarisasi menggunakan *Robust Scaler* dan transformasi Yeo-Johnson untuk mendekatkan distribusinya ke bentuk normal, sesuai asumsi algoritma Gaussian Naïve Bayes. Pendekatan ini bertujuan untuk mengevaluasi performa algoritma ketika hanya data numerik yang digunakan. Pada pendekatan kedua, model Gaussian melibatkan semua variabel, baik numerik maupun kategorikal. Variabel kategorikal seperti jenis kelamin, rencana jurusan, pilihan siswa, dan pilihan orang tua, telah di-*encode* sebelum dimasukkan ke dalam model. Meskipun pelibatan variabel kategorikal melanggar asumsi distribusi normal, model ini tetap diuji untuk memahami bagaimana variabel kategorikal dapat memengaruhi hasil prediksi.

Sementara itu, model Bernoulli Naïve Bayes dirancang untuk menangani data biner. Oleh karena itu, variabel numerik seperti Nilai Support A–G dibinarisasi berdasarkan rata-rata nilai pada data latih. Nilai di atas rata-rata diklasifikasikan sebagai 1, sedangkan nilai di bawah atau sama dengan rata-rata diklasifikasikan sebagai 0. Variabel kategorikal seperti rencana jurusan, pilihan siswa, dan pilihan orang tua, yang telah melalui proses *encoding*, secara langsung sesuai dengan format data yang diperlukan untuk model ini. Dengan menggunakan probabilitas kehadiran atau ketiadaan fitur, model Bernoulli memanfaatkan pendekatan probabilistik untuk memprediksi kelas target.

D. Evaluasi Model

Evaluasi ketiga model dilakukan menggunakan empat metrik utama, yaitu akurasi, presisi, *recall*, dan skor F1. Model Gaussian yang menggunakan variabel numerik, Gaussian yang melibatkan semua variabel, dan Bernoulli yang memanfaatkan data biner, dibandingkan untuk menentukan model yang memberikan hasil terbaik dalam memprediksi kelas terpilih siswa. Hasil evaluasi disajikan dalam bentuk *output* program python dan diagram batang

untuk memvisualisasikan perbandingan performa antar model.

Sebagai langkah tambahan, validasi model dilakukan menggunakan teknik k-*fold cross-validation*, yang membagi data latih menjadi beberapa lipatan (*folds*) untuk digunakan secara bergantian sebagai data latih dan validasi. Teknik ini memberikan metrik evaluasi yang lebih stabil dengan meminimalkan variabilitas akibat pembagian *dataset* tunggal. Selain itu, grafik Kurva pembelajaran (*learning curve*) dibuat untuk mengevaluasi performa model terhadap ukuran data latih. Grafik ini menunjukkan skor akurasi pada data latih dan validasi, memberikan wawasan tentang kemungkinan terjadinya *overfitting* atau *underfitting*.

Setiap model juga dievaluasi melalui matriks kebingungan (*confusion matrix*), yang divisualisasikan untuk menunjukkan distribusi prediksi terhadap kelas target sebenarnya. Matriks ini dilengkapi dengan anotasi nilai numerik pada setiap sel dan skala warna untuk mempermudah interpretasi hasil klasifikasi. Proses pembangunan model dilakukan menggunakan pendekatan *pipeline* untuk memastikan bahwa semua langkah transformasi, pelatihan, dan evaluasi dilakukan secara sistematis dan hanya berdasarkan data latih selama proses *fitting*. Pendekatan *pipeline* ini mengurangi risiko data *leakage*, meningkatkan efisiensi, serta mempermudah replikasi.

Dengan mempertimbangkan asumsi yang berlaku pada setiap algoritma, pembangunan model ini memberikan wawasan yang mendalam tentang kontribusi variabel numerik dan kategorikal dalam proses klasifikasi. Meskipun model Gaussian dengan semua variabel tidak sepenuhnya memenuhi asumsi distribusi Gaussian, analisis terhadap model ini tetap dilakukan untuk memahami bagaimana variabel kategorikal memengaruhi performa klasifikasi. Hasil dari analisis ini memberikan landasan yang kuat untuk memahami proses pemilihan program peminatan siswa di SMAS Katolik Santo Yoseph Denpasar.

IV. HASIL DAN PEMBAHASAN

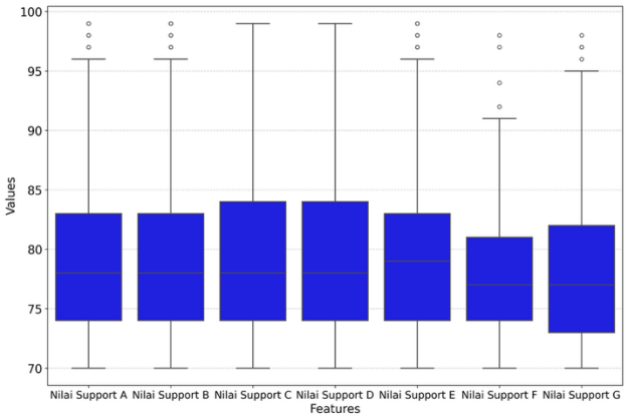
Dataset penelitian ini terdiri dari 448 baris data yang mencakup variabel numerik dan kategorikal, masing-masing memberikan gambaran faktor yang memengaruhi pemilihan kelas di SMAS Katolik Santo Yoseph Denpasar. Gambar 2 menampilkan sebagian *dataset*, yang terdiri dari kolom-kolom seperti nilai rata-rata (Nilai Support A–G), jenis kelamin, rencana jurusan yang akan dipilih di perguruan tinggi, pilihan siswa, pilihan orang tua, dan kelas terpilih sebagai variabel target. Variabel target ini mewakili program peminatan siswa, sedangkan variabel prediktor meliputi kombinasi data akademik dan preferensi siswa maupun orang tua.

Distribusi nilai rata-rata untuk variabel numerik ditunjukkan pada Gambar 3 melalui visualisasi boxplot. Secara keseluruhan, nilai rata-rata pada variabel Nilai Support A–G tersebar dalam rentang 70 hingga 100 dengan median yang berkisar di sekitar angka 80. Hal ini mencerminkan performa akademik siswa yang relatif konsisten di berbagai kategori mata pelajaran. Meskipun ditemukan beberapa *outlier*, terutama pada nilai di atas 95, keberadaan *outlier* ini lebih merepresentasikan siswa dengan performa akademik yang sangat tinggi dibandingkan indikasi adanya kesalahan data. Distribusi nilai ini juga divisualisasikan dalam histogram pada Gambar 4, yang menunjukkan bahwa sebagian besar nilai berkumpul pada rentang 75–85. Pola distribusi pada hampir semua variabel menunjukkan adanya *skewness*, dengan nilai-nilai yang cenderung terkonsentrasi pada rentang yang lebih rendah. Skewness ini terlihat pada hampir semua variabel yang menunjukkan pola distribusi yang sedikit miring ke kiri. Secara keseluruhan, pola distribusi ini memberikan gambaran penting tentang variasi dan kecenderungan nilai akademik siswa dalam berbagai kategori mata pelajaran.

	Sex	Rencana Jurusan	Pilihan Siswa	Pilihan Orang Tua	Nilai Support A	Nilai Support B	Nilai Support C	Nilai Support D	Nilai Support E	Nilai Support F	Nilai Support G	Kelas Terpilih
0	Perempuan	Sosial D1	XI-D	XI-D	73	78	71	98	71	73	77	XI-D
1	Laki-laki	Sosial D3	XI-D	XI-D	73	75	75	96	82	78	72	XI-D
2	Perempuan	Sains A2	XI-A	XI-A	99	71	79	74	75	84	72	XI-A
3	Perempuan	Sosial D2	XI-D	XI-D	75	82	74	99	73	81	84	XI-D
4	Perempuan	Sosial D1	XI-D	XI-D	78	72	84	86	74	81	82	XI-D
...	...	...	...	...	...	...	...	...	...	...	...	...
444	Laki-laki	Sains C1	XI-C	XI-C	81	80	92	76	70	72	72	XI-C
445	Perempuan	Sosial F3	XI-F	XI-F	81	75	79	82	82	89	81	XI-F
446	Perempuan	Sains A3	XI-A	XI-A	85	79	80	77	75	83	75	XI-A
447	Laki-laki	Komunikasi G3	XI-G	XI-F	78	75	79	76	81	85	82	XI-F
448	Laki-laki	Sains A3	XI-A	XI-A	99	78	80	80	79	80	73	XI-A

448 rows × 12 columns

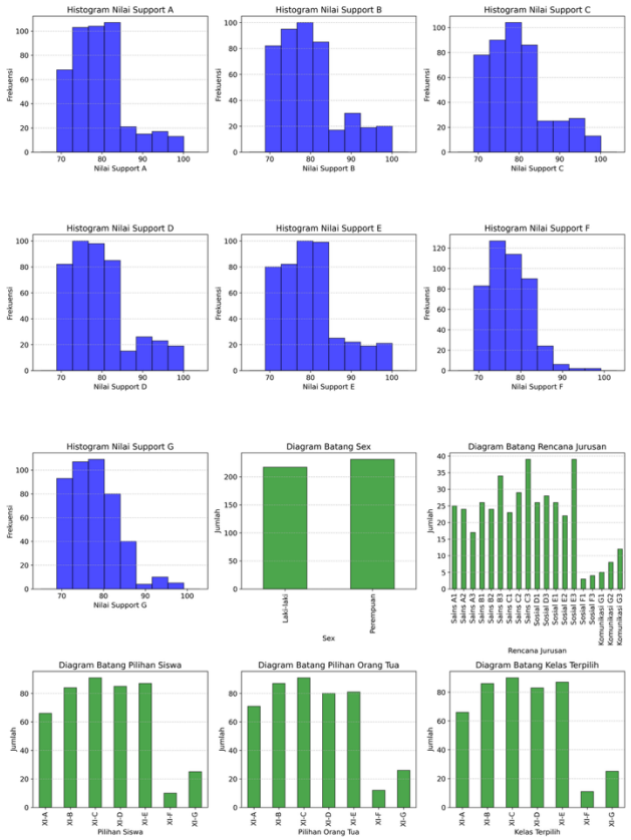
Gambar. 2 Cuplikan *dataset* penelitian



Gambar. 3 *Boxplot* nilai rata-rata untuk variabel numerik

Variabel kategorikal, sebagaimana divisualisasikan dalam Gambar 4, menunjukkan pola distribusi yang beragam. Proporsi jenis kelamin siswa hampir seimbang antara laki-laki dan perempuan, memberikan dasar yang netral untuk mengevaluasi apakah gender memengaruhi pemilihan kelas. Pada variabel rencana jurusan, terlihat bahwa kategori Sains C3 dan Sosial E3 memiliki jumlah siswa terbanyak, menunjukkan konsentrasi tertentu pada pilihan-pilihan tersebut.

Perbedaan pola juga terlihat pada variabel pilihan siswa dan pilihan orang tua. Dua kelas yang paling banyak dipilih oleh siswa adalah XI-C dan XI-E, sementara pilihan orang tua lebih banyak terkonsentrasi pada kelas XI-C dan XI-B. Perbedaan ini mengindikasikan adanya dinamika antara preferensi siswa dan pengaruh keluarga dalam pemilihan kelas.



Gambar. 4 Histogram variable numerik dan diagram batang variabel kategorik

Adapun variabel target, yaitu kelas terpilih, merepresentasikan kenyataan siswa ditempatkan di kelas mana berdasarkan data prediktor. Dua kelas dengan jumlah siswa terbanyak pada variabel ini adalah XI-B dan XI-C, mencerminkan hasil akhir dari proses penentuan kelas.

Dalam penelitian ini, dilakukan analisis menggunakan tiga pendekatan model Naïve Bayes: Gaussian Naive Bayes dengan variabel numerik saja, Gaussian Naive Bayes dengan semua variabel, dan Bernoulli Naive Bayes dengan variabel numerik yang dibinerisasi. Pendekatan ini dirancang untuk memahami peran masing-masing jenis data (numerik dan kategorikal) terhadap kemampuan model dalam mengklasifikasikan pemilihan kelas siswa. Evaluasi dilakukan dengan metrik akurasi, presisi, *recall*, F1-score, matriks kebingungan, dan kurva pembelajaran.

Gaussian Naive Bayes dirancang untuk menangani data numerik dengan asumsi distribusi normal. Dalam pendekatan ini, hanya variabel numerik (nilai rata-rata mata pelajaran kategori A–G) yang digunakan. Sebelum pemodelan, dilakukan transformasi data menggunakan *RobustScaler* untuk mengurangi pengaruh *outlier* dan *PowerTransformer* untuk menormalkan distribusi. Langkah-langkah ini bertujuan untuk memastikan bahwa asumsi Gaussian terpenuhi.

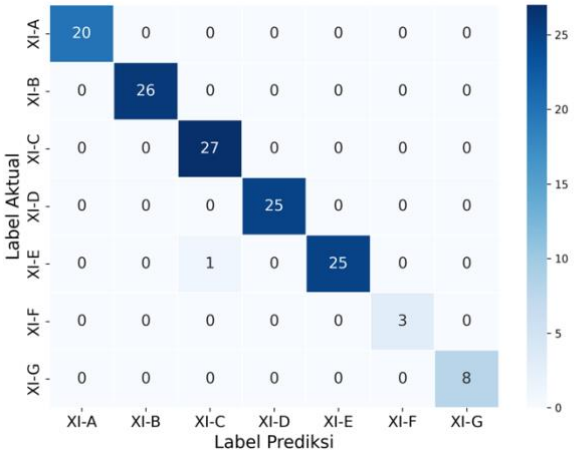
Model Gaussian Naive Bayes diterapkan pada data numerik untuk mengevaluasi akurasi prediksi berdasarkan nilai akademik siswa. Hasil laporan klasifikasi, sebagaimana ditunjukkan pada Gambar 5, menunjukkan performa model yang sangat baik. Model ini berhasil mencapai presisi, *recall*, dan *F1-score* sebesar 1.00 untuk sebagian besar kelas, termasuk XI-A, XI-B, XI-D, XI-E, XI-F, dan XI-G. Meskipun terdapat sedikit penurunan presisi pada kelas XI-C (0.96), nilai *recall* tetap sempurna pada semua kelas kecuali kelas XI-E, sehingga memastikan bahwa model memiliki kemampuan tinggi dalam mengenali semua data aktual untuk masing-masing kelas.

```
=== Classification Report: Gaussian Naive Bayes (Numerik) ===
              precision    recall  f1-score   support

        XI-A       1.00      1.00      1.00        20
        XI-B       1.00      1.00      1.00        26
        XI-C       0.96      1.00      0.98        27
        XI-D       1.00      1.00      1.00        25
        XI-E       1.00      0.96      0.98        26
        XI-F       1.00      1.00      1.00         3
        XI-G       1.00      1.00      1.00         8

    accuracy                           0.99       135
   macro avg       0.99      0.99      0.99       135
weighted avg       0.99      0.99      0.99       135
```

Gambar. 5 Laporan klasifikasi *gaussian naive bayes* (numerik)

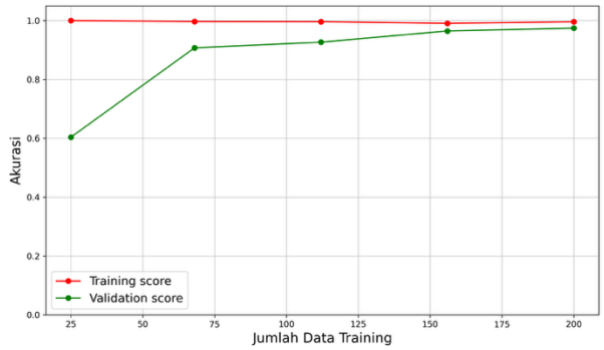


Gambar. 6 Matriks kebingungan *gaussian naive bayes* (numerik)

Selanjutnya, matriks kebingungan pada Gambar 6 memberikan representasi visual dari hasil prediksi model ini. Sebagian besar data pada setiap kelas berhasil diprediksi dengan benar tanpa adanya kesalahan klasifikasi, kecuali satu kasus pada kelas XI-E yang salah diprediksi sebagai XI-C. Hal ini menunjukkan bahwa model sangat mampu membedakan pola distribusi data numerik untuk sebagian besar kelas, tetapi terdapat tantangan dalam memisahkan kelas XI-E dari XI-C. Hal ini mungkin disebabkan oleh kemiripan distribusi data numerik antara kedua kelas tersebut.

Kurva pembelajaran untuk Model Gaussian Naive Bayes pada data numerik, seperti terlihat pada Gambar 7, menunjukkan kinerja yang sangat baik dengan akurasi pelatihan stabil di angka 1.00 dan akurasi validasi yang meningkat secara konsisten seiring bertambahnya data pelatihan, hingga hampir menyamai akurasi pelatihan pada titik akhir. Pola ini menunjukkan bahwa model tidak mengalami *overfitting* dan berhasil menjaga keseimbang-an antara akurasi pelatihan dan validasi, mencerminkan kemampuan generalisasi yang kuat. Hasil ini

mengindikasikan bahwa asumsi distribusi Gaussian sangat sesuai untuk data numerik dalam *dataset* ini, memungkinkan prediksi yang akurat berdasarkan pola distribusi nilai akademik siswa.



Gambar. 7 Kurva pembelajaran *gaussian naive bayes* (numerik)

Meskipun model Gaussian Naive Bayes dengan hanya variabel numerik menghasilkan performa terbaik, pendekatan ini hanya mempertimbangkan data akademik siswa, tanpa memperhitungkan preferensi atau pandangan siswa dan orang tua. Untuk memberikan gambaran yang lebih komprehensif tentang faktor-faktor yang memengaruhi pemilihan kelas, model Gaussian Naive Bayes dengan semua variabel, termasuk kategori, diuji untuk mengevaluasi pengaruh variabel non-numerik seperti jenis kelamin, rencana jurusan, pilihan siswa, dan pilihan orang tua.

Model Gaussian Naive Bayes yang menggunakan semua variabel (numerik dan kategorikal) menunjukkan performa yang cukup baik meskipun mengalami penurunan akurasi dibandingkan model yang hanya menggunakan data numerik. Penurunan ini terjadi karena asumsi distribusi Gaussian pada variabel kategorikal tidak dapat terpenuhi dengan baik. Variabel kategorikal tidak memiliki distribusi kontinu, sehingga menyulitkan model untuk secara akurat menginterpretasikan hubungan antara kategori dan target kelas.

Hasil laporan klasifikasi pada Gambar 8 menunjukkan bahwa model ini mencapai akurasi keseluruhan sebesar 96%, dengan nilai precision dan *recall* rata-rata berturut-turut sebesar 0,95 dan 0,91. *F1-score* yang dihasilkan juga cukup baik, dengan nilai rata-rata tertimbang sebesar 0,96, mencerminkan kemampuan model untuk menangani ketidakseimbangan data pada beberapa kelas. Namun, analisis presisi dan *recall* menunjukkan bahwa kelas XI-F dan XI-G memiliki performa yang relatif lebih rendah dibandingkan kelas lainnya. Sebagai contoh, presisi kelas XI-F hanya mencapai 1,00, tetapi *recall* sebesar 0,67 menunjukkan bahwa model cenderung gagal mengenali sebagian besar siswa pada kelas ini.

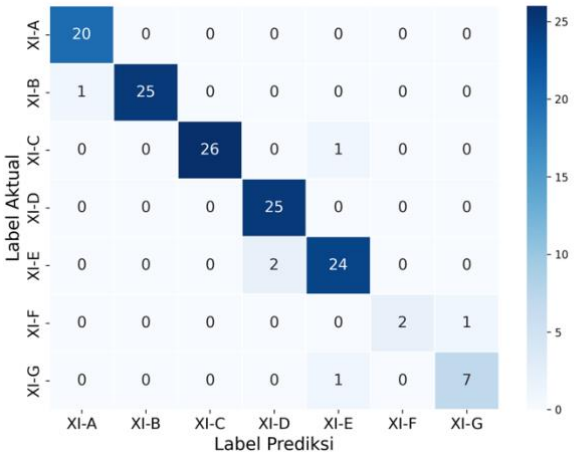
```
=== Classification Report: Gaussian Naive Bayes (Semua Variabel) ===
              precision    recall  f1-score   support

        XI-A       0.95      1.00      0.98        20
        XI-B       1.00      0.96      0.98        26
        XI-C       1.00      0.96      0.98        27
        XI-D       0.93      1.00      0.96        25
        XI-E       0.92      0.92      0.92        26
        XI-F       1.00      0.67      0.80         3
        XI-G       0.88      0.88      0.88         8

    accuracy                           0.96       135
   macro avg       0.95      0.91      0.93       135
weighted avg       0.96      0.96      0.96       135
```

Gambar. 8 Laporan klasifikasi *gaussian naive bayes* (semua variabel)

Pada Gambar 9, matriks kebingungan memberikan wawasan lebih rinci tentang distribusi prediksi model. Sebagian besar prediksi model sesuai dengan label aktual, tetapi terdapat beberapa kesalahan klasifikasi, terutama pada kelas XI-E dan XI-F. Misalnya, dua siswa dari kelas XI-E salah diklasifikasikan sebagai kelas XI-D. Hal ini menunjukkan bahwa karakteristik siswa pada kedua kelas ini cenderung serupa, sehingga menyulitkan model untuk membedakannya. Sebaliknya, kelas XI-A dan XI-D menunjukkan hasil prediksi yang sangat baik, dengan hampir tidak ada kesalahan klasifikasi.



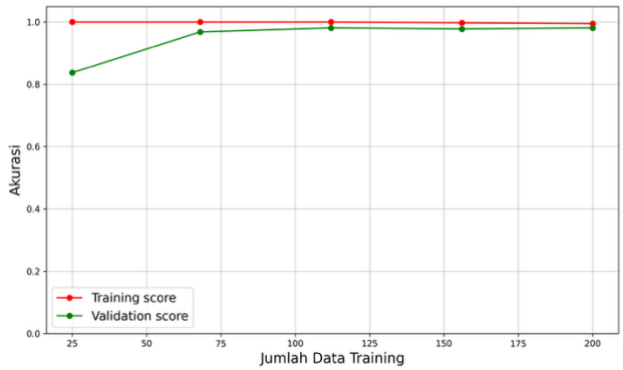
Gambar. 9 Matriks kebingungan *gaussian naive bayes* (semua variabel)

Namun, pentingnya melibatkan variabel kategorikal dalam analisis tidak dapat diabaikan. Preferensi siswa dan orang tua mencerminkan faktor-faktor non-akademik, seperti jenis kelamin, cita-cita, minat pribadi, dan pengaruh keluarga. Walaupun performa model sedikit menurun, melibatkan variabel kategorikal memungkinkan analisis yang lebih inklusif, yang mencerminkan perspektif manusiawi dalam pengambilan keputusan peminatan siswa.

Sebagai contoh, preferensi siswa dan orang tua sering kali mencerminkan tujuan pendidikan jangka panjang yang tidak selalu berkorelasi langsung dengan nilai akademik siswa. Oleh karena itu, meskipun data numerik memberikan prediksi yang sangat akurat, mempertimbangkan variabel kategorikal memberikan wawasan tambahan yang penting, terutama untuk memahami bagaimana preferensi ini memengaruhi keputusan akhir.

Gambar 10 menunjukkan kurva pembelajaran model Gaussian Naive Bayes dengan semua variabel. Pada

awalnya, terdapat celah antara skor pelatihan dan validasi akibat keterbatasan data pelatihan. Namun, seiring meningkatnya jumlah data hingga 150, skor validasi mendekati skor pelatihan, menunjukkan peningkatan generalisasi model. Setelah jumlah data mencapai ambang batas, kurva validasi mulai mendatar, menandakan bahwa penambahan data tidak lagi memberikan peningkatan signifikan.



Gambar. 10 Kurva pembelajaran *gaussian naive bayes* (semua variabel)

Kurva pembelajaran pada Gambar 7 dan Gambar 10 menunjukkan perbedaan signifikan dalam kemampuan generalisasi model. Model dengan hanya variabel numerik menunjukkan akurasi validasi yang meningkat secara konsisten hingga mendekati 100%, mencerminkan kecocokan data numerik dengan asumsi distribusi normal yang mendasari Gaussian Naive Bayes. Sebaliknya, model dengan semua variabel, meskipun akurasi pelatihannya tetap tinggi, mengalami penurunan akurasi validasi yang stabil pada kisaran 91–94%. Penurunan ini disebabkan oleh variabel kategorikal yang distribusinya tidak sesuai dengan asumsi Gaussian, yang pada akhirnya memengaruhi performa. Namun, model dengan semua variabel memberikan perspektif yang lebih luas, karena melibatkan faktor non-akademik seperti preferensi siswa dan orang tua, yang sangat relevan dalam konteks pengambilan keputusan peminatan.

Meskipun model Gaussian Naive Bayes dengan semua variabel memberikan wawasan lebih komprehensif, performanya tetap lebih rendah dibandingkan model numerik karena keterbatasan asumsi distribusi normal pada data kategorikal. Untuk mengatasi kelemahan ini, diperlukan pendekatan yang dapat menangani data numerik dan kategorikal secara bersamaan tanpa melanggar asumsi distribusi. Salah satu pendekatan yang digunakan adalah Bernoulli Naive Bayes, yang mengubah data numerik menjadi data biner berdasarkan ambang batas nilai rata-rata. Pendekatan ini dirancang untuk menyelaraskan sifat data sehingga lebih sesuai dengan model, sekaligus mempertahankan relevansi variabel kategorikal dalam analisis.

Laporan klasifikasi dari model ini, sebagaimana ditampilkan pada Gambar 11, memperlihatkan bahwa akurasi keseluruhan model mencapai 97%, menandakan kemampuan model yang sangat baik dalam mengklasifikasikan siswa ke dalam kelas yang tepat. Metrik presisi, *recall*, dan *F1-score* memberikan wawasan lebih lanjut tentang performa model untuk masing-masing kelas. Presisi tertinggi dicapai pada kelas XI-B, XI-C, dan XI-F, masing-masing dengan nilai 100%, yang mengindikasikan bahwa model hampir tidak membuat kesalahan dalam memprediksi data siswa ke kelas-kelas tersebut. Namun, *recall* untuk kelas XI-F hanya berada di angka 67%, menunjukkan bahwa meskipun prediksi untuk kelas ini sangat akurat, model gagal menangkap semua siswa yang sebenarnya berada di kelas XI-F. Sebaliknya, kelas XI-G memiliki *recall* sempurna sebesar 100%, namun nilai presisinya lebih rendah, yaitu 89%, yang menunjukkan bahwa beberapa siswa yang diprediksi berada di kelas XI-G sebenarnya berasal dari kelas lain. F1-score, yang merupakan metrik harmonisasi antara presisi dan *recall*, memberikan hasil rata-rata yang cukup baik untuk semua kelas. Nilai *F1-score* tertinggi dicapai oleh kelas XI-C dan XI-D, yang menunjukkan keseimbangan optimal dalam identifikasi siswa yang benar secara prediktif dan aktual. Secara makro, rata-rata *F1-score* mencapai 95%, sedangkan rata-rata berbobot menunjukkan hasil yang sedikit lebih baik di angka 97%.

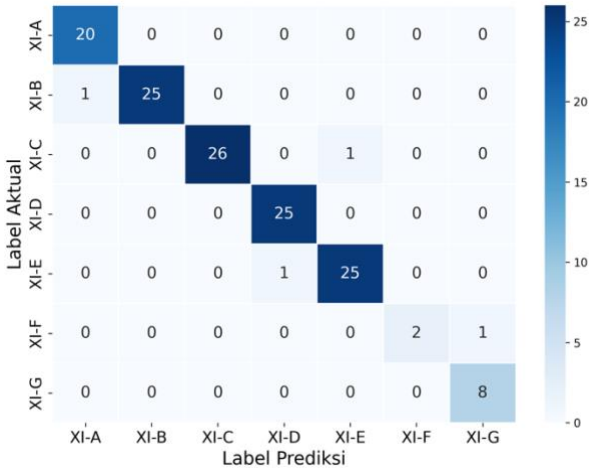
```
=== Classification Report: Bernoulli Naive Bayes (Numerik Biner) ===
              precision    recall  f1-score   support

        XI-A       0.95      1.00      0.98        20
        XI-B       1.00      0.96      0.98        26
        XI-C       1.00      0.96      0.98        27
        XI-D       0.96      1.00      0.98        25
        XI-E       0.96      0.96      0.96        26
        XI-F       1.00      0.67      0.80         3
        XI-G       0.89      1.00      0.94         8

    accuracy                           0.97       135
   macro avg       0.97      0.94      0.95       135
weighted avg       0.97      0.97      0.97       135
```

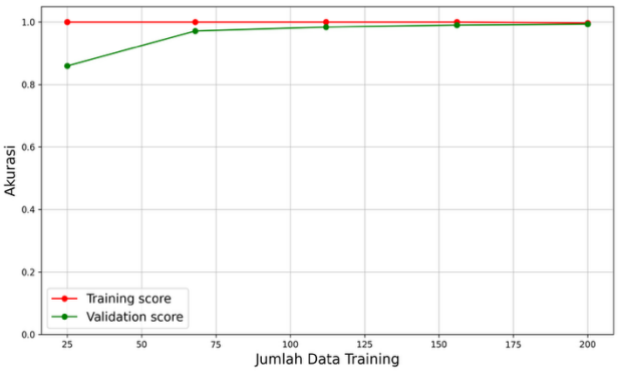
Gambar. 11 Laporan klasifikasi *bernoulli naive bayes*

Matriks kebingungan untuk model ini, sebagaimana ditampilkan pada Gambar 12, menunjukkan pola prediksi yang mendekati sempurna untuk sebagian besar kelas. Sebagian besar prediksi berada di diagonal utama, yang menunjukkan bahwa model mampu mengenali dengan baik data siswa yang sesuai dengan kelas target mereka. Kesalahan prediksi yang kecil, seperti salah klasifikasi siswa dari kelas XI-F ke kelas XI-G, dapat terjadi karena distribusi data yang tidak seimbang atau fitur biner yang kurang merepresentasikan karakteristik unik beberapa kelas.



Gambar. 12 Matriks kebingungan *bernoulli naive bayes*

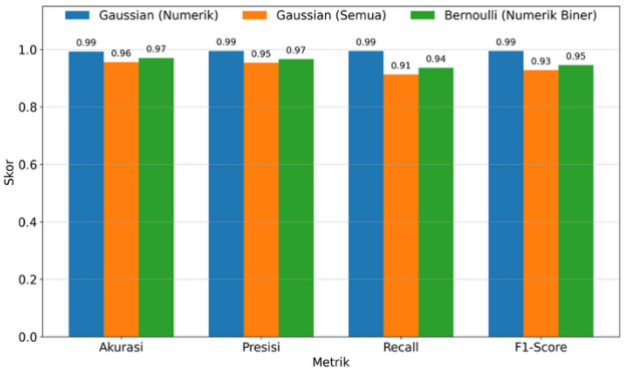
Kurva pembelajaran dari model ini, sebagaimana diperlihatkan pada Gambar 13, memberikan gambaran tentang stabilitas model seiring dengan bertambahnya jumlah data pelatihan. Kurva *training* dan validasi menunjukkan pola konvergensi, dengan akurasi mendekati 100% saat jumlah data pelatihan meningkat. Pola ini menunjukkan bahwa model tidak mengalami *overfitting* atau *underfitting* yang signifikan, yang menandakan bahwa pendekatan binerisasi data berhasil menjaga keseimbangan antara kompleksitas model dan generalisasi terhadap data baru.



Gambar. 13 Kurva pembelajaran *bernoulli naive bayes*

Model Bernoulli Naive Bayes berhasil menawarkan alternatif yang efektif untuk menangani asumsi Gaussian yang mungkin tidak dipenuhi oleh data asli. Dengan membinerisasi data numerik, model ini mampu menangkap pola yang lebih sederhana namun tetap relevan untuk pengambilan keputusan. Walaupun performa keseluruhan sangat baik, ada ruang untuk perbaikan, terutama dalam menangani kelas dengan jumlah data yang lebih sedikit, seperti XI-F, yang *recall*-nya masih dapat ditingkatkan melalui pengayaan data atau pendekatan pemodelan yang lebih kompleks.

Gambar 14 menunjukkan perbandingan performa ketiga model: Gaussian Naive Bayes dengan data numerik saja, Gaussian Naive Bayes dengan semua variabel, dan Bernoulli Naive Bayes dengan numerik biner. Model Gaussian Naive Bayes dengan data numerik memiliki performa tertinggi pada semua metrik, dengan akurasi, presisi, *recall*, dan *F1-score* masing-masing mencapai 0,99. Penambahan variabel kategorikal pada model Gaussian Naive Bayes sedikit menurunkan performa menjadi 0,96 untuk akurasi dan F1-score, dan 0,93 untuk *recall*, menunjukkan bahwa variabel kategorikal memperkenalkan sedikit ketidakpastian dalam prediksi.



Gambar. 14 Perbandingan metrik untuk model *naive bayes*

Model Bernoulli Naive Bayes dengan numerik biner memiliki performa yang lebih seimbang dibandingkan Gaussian dengan semua variabel, dengan nilai presisi dan F1-score masing-masing mencapai 0,97 dan akurasi tetap 0,97. Pendekatan ini menunjukkan bahwa binerisasi data numerik membantu mengatasi pelanggaran asumsi distribusi Gaussian, tetapi tetap mempertahankan pengaruh variabel kategorikal. Temuan ini signifikan karena menunjukkan bahwa modifikasi sederhana terhadap struktur data dapat menghasilkan model klasifikasi yang lebih stabil dan kompatibel dalam konteks data pendidikan yang kompleks dan tidak selalu memenuhi asumsi statistik tertentu.

Secara keseluruhan, Gaussian Naive Bayes dengan data numerik memiliki performa terbaik, tetapi Bernoulli Naive Bayes memberikan alternatif yang stabil dengan metrik yang cukup kompetitif, terutama dalam menangani kombinasi variabel numerik dan kategorikal. Kontribusi utama dari penelitian ini adalah membuktikan bahwa pemilihan model yang tepat dan penyesuaian tipe data (binerisasi) sangat berperan dalam meningkatkan keandalan sistem rekomendasi peminatan siswa. Hal ini penting terutama dalam lingkungan sekolah dengan karakteristik data terbatas namun beragam, seperti pada SMAS Katolik Santo Yoseph Denpasar. Hasil kinerja model yang cenderung mendekati sempurna juga menjadi catatan penting, karena bisa menjadi indikasi anomali yang perlu diuji lebih lanjut dengan menggunakan data yang belum memiliki label sama sekali (*unsupervised*) untuk menilai kemampuan generalisasi yang sesungguhnya. Dengan demikian, penelitian ini memberikan dasar kuat untuk pengembangan sistem rekomendasi peminatan berbasis data yang adaptif dan akurat.

V. KESIMPULAN

Penelitian ini membandingkan performa Gaussian Naïve Bayes dan Bernoulli Naïve Bayes dalam membantu siswa memilih program peminatan di SMAS Katolik Santo Yoseph Denpasar, dalam konteks penerapan Kurikulum Merdeka yang memberikan kebebasan lebih kepada siswa untuk memilih mata pelajaran berdasarkan minat, bakat, dan kemampuan. Berdasarkan hasil analisis, Gaussian Naïve Bayes menunjukkan performa yang sangat baik untuk data numerik, dengan akurasi, presisi, *recall*, dan *F1-score* yang konsisten tinggi di semua kelas. Namun, ketika variabel kategorik ditambahkan ke dalam model, performa Gaussian sedikit menurun, menunjukkan bahwa model ini kurang optimal dalam menangani kombinasi data numerik dan kategorik secara bersamaan.

Sebaliknya, Bernoulli Naïve Bayes, yang mengubah data numerik menjadi biner, menunjukkan kinerja yang lebih stabil dalam memprediksi beberapa kategori, meskipun tidak setinggi Gaussian Naïve Bayes pada data numerik saja. Model ini efektif dalam menangkap pola dari atribut biner yang mewakili preferensi siswa dan pengaruh lingkungan, namun masih memiliki kelemahan dalam beberapa kategori dengan frekuensi rendah. Kurva pembelajaran (*Learning curve*) juga menunjukkan bahwa Bernoulli Naïve Bayes memiliki kemampuan generalisasi yang baik dengan kurva validasi yang mendekati kurva pelatihan seiring bertambahnya data.

Hasil ini menegaskan bahwa penggunaan Gaussian Naïve Bayes sangat cocok untuk analisis berbasis data numerik, seperti nilai akademik siswa. Namun, dalam konteks Kurikulum Merdeka yang menekankan pada minat dan preferensi siswa, penggunaan Bernoulli Naïve Bayes menjadi relevan untuk melibatkan atribut kategorik dan biner. Kombinasi kedua pendekatan ini dapat menjadi solusi strategis untuk mendukung pengambilan keputusan yang lebih komprehensif bagi siswa, guru, dan konselor.

Saran untuk implementasi di masa depan, integrasi kedua model dapat dipertimbangkan, di mana Gaussian Naïve Bayes digunakan untuk menganalisis data numerik, sementara Bernoulli Naïve Bayes mengolah atribut biner dan kategorik. Selain itu, penambahan fitur baru yang lebih representatif terhadap preferensi siswa, seperti hasil tes minat bakat atau wawancara, dapat meningkatkan akurasi model secara keseluruhan. Hal ini akan memastikan bahwa pemilihan program peminatan benar-benar mencerminkan potensi dan kebutuhan individu siswa.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada SMAS Katolik Santo Yoseph Denpasar atas dukungan, kerja sama, dan pemberian data yang diperlukan sehingga penelitian ini dapat diselesaikan dengan baik.

REFERENSI

[1] R. P. Arwitaningsih, B. F. Dewi, E. M. Rahmawati, dan K. Khuriyah, "Konsep dan Implementasi Kurikulum Merdeka pada Ranah Rumpun Mata Pelajaran Pendidikan Islam di Sekolah Dasar Islam Terpadu Al Hadi Mojolaban Sukoharjo," *Modeling: Jurnal Program Studi PGMI*, vol. 10, no. 2, Art. no. 2, Jun 2023, doi: 10.69896/modeling.v10i2.1752.

[2] G. Napitupulu, M. Silalahi, dan S. Gultom, "Implementasi Manajemen Kurikulum Merdeka Belajar dalam Peningkatan Mutu Pendidikan di SMA Negeri 1 Bandar," *Journal on Education*, vol. 6, no. 1, Art. no. 1, Jun 2023, doi: 10.31004/joe.v6i1.3722.

[3] S. Mahaly, J. O. Papilaya, dan Jumail ', "Analisis Pemilihan Minat Mata Pelajaran Pilihan Siswa SMA Laboratorium Universitas Pattimura," *Pedagogika: Jurnal Pedagogik dan Dinamika Pendidikan*, vol. 12, no. 1, Art. no. 1, Apr 2024, doi: 10.30598/pedagogikavol12issue1page101-108.

[4] I. Nugraheni, A. C. Budiati, dan Nurhadi, "Analisis Strategi Sekolah Dalam Melaksanakan Kebijakan Penghapusan Penjurusan di SMA Negeri 3 Surakarta," *Pendas : Jurnal Ilmiah Pendidikan Dasar*, vol. 9, no. 3, Art. no. 3, Agu 2024, doi: 10.23969/jp.v9i3.17440.

[5] B. D. S. Ghaleb, "The Relationship between Career Selection and Career Satisfaction," *J. Bus. Manag. Econ. Development.*, vol. 2, no. 03, hlm. 1045–1056, Jun 2024, doi: 10.59653/jbmed.v2i03.851.

[6] A. Sinring dan N. F. Umar, "The influence of Different Types of Career Exploration on Achievement Career Identity Among Z Generation," *j. of. sci. technol.*, vol. 9, no. 1, hlm. 8, Apr 2023, doi: 10.26858/est.v9i1.43326.

[7] H. Umar dan E. Masnawati, "Peran Lingkungan Sekolah Dalam Pembentukan Identitas Remaja," *Jurnal Kajian Pendidikan Islam*, hlm. 191–204, Jul 2024, doi: 10.58561/jkpi.v3i2.137.

[8] P. Johnson, T. D. Schamuhn, D. B. Nelson, dan W. C. Buboltz, "Differentiation Levels of College Students: Effects on Vocational Identity and Career Decision Making," *The Career Development Quart*, vol. 62, no. 1, hlm. 70–80, Mar 2014, doi: 10.1002/j.2161-0045.2014.00071.x.

[9] V. Kumari, A. F. Meghji, R. Qadir, U. Gianchand, dan F. B. Shaikh, "Predicting Student Performance Using Educational Data Mining: A Review," *KJCIS*, vol. 7, no. 1, Okt 2024, doi: 10.51153/kjcis.v7i1.212.

[10] I. D. Setiawan dan A. Triayudi, "Penerapan Data Mining Dengan Menggunakan Algoritma Clustering K-Means Untuk Pembagian Jurusan Pada Sekolah Menengah Atas," *JoSYC*, vol. 5, no. 2, hlm. 380–392, Feb 2024, doi: 10.47065/josyc.v5i2.4970.

[11] R. Sovia, E. P. W. Mandala, dan S. Mardhiah, "Algoritma K-Means dalam Pemilihan Siswa Berprestasi dan Metode SAW untuk Prediksi Penerima Beasiswa Berprestasi," *JEPIN*, vol. 6, no. 2, hlm. 181, Agu 2020, doi: 10.26418/jp.v6i2.37759.

[12] N. Nurajijah, D. A. Ningtyas, dan M. Wahyudi, "Klasifikasi Siswa Smk Berpotensi Putus Sekolah Menggunakan Algoritma Decision Tree, Support Vector Machine dan Naive Bayes," *JKI*, vol. 7, no. 2, Des 2019, doi: 10.31294/jki.v7i2.6839.

[13] I. M. D. Priyatama dan R. Ridwansyah, "Klasifikasi Anak Berkebutuhan Khusus Tunagrahita Menggunakan Metode Algoritma C4.5," *Jurnal Informatika dan Komputer*, vol. 24, no. 1, hlm. 90–95, Mar 2022, doi: 10.31294/paradigma.v24i1.1087.

[14] Moh. Roghib, N. Rahaningsih, dan R. Danar Dana, "Penerapan Algoritma C4.5 Untuk Seleksi Penjurusan Siswa Baru Pada Sekolah Menengah Kejuruan (Studi Kasus: SMK Plus Al-Hilal Arjawinangun)," *jati*, vol. 8, no. 1, hlm. 861–866, Mar 2024, doi: 10.36040/jati.v8i1.8436.

[15] K. Amanda, D. Saripurna, dan Mhd. Z. Siambaton, "Penerapan Algoritma Cart dalam Penentuan Jurusan Siswa di SMA: Studi Kasus SMA Negeri 2 Perbaungan," *hello world j. ilmu komp'ût.*, vol. 2, no. 4, hlm. 169–177, Mar 2024, doi: 10.56211/helloworld.v2i4.404.

[16] N. Khafifah, G. N. A. Wibawa, Arman, W. Somayasa, Ruslan, dan L. Gubu, "Penerapan Metode Naïve Bayes Untuk Penentuan Jurusan Siswa-Siswi Pada SMAN 9 Kendari," *Jurnal Matematika Komputasi dan Statistika*, vol. 4, no. 1, Art. no. 1, Jul 2024, doi: 10.33772/jmks.v4i1.73.

[17] N. Ramadhani, Z. Effendy, dan I. Darmawan, "Penerapan Algoritma Naïve Bayes Classifier dan Fungsi Gaussian Untuk Penentuan Penjurusan Siswa Kelas X," *SMARTICS Journal*, vol. 8, no. 1, Art. no. 1, Apr 2022, doi: 10.21067/smartics.v8i1.6996.

[18] A. Riyanto dan E. S. Ompusunggu, "Implementasi Data Mining Untuk Mengklasifikasi Hasil Belajar Siswa/i Dengan Metode Naïve Bayes," *Jurnal Teknologi Dan Ilmu Komputer Prima*, vol. 7, no. 2, Art. no. 2, Okt 2024, doi: 10.34012/jutikomp.v7i2.5237.

[19] Supangat dan R. Giovanni, "Evaluasi Tingkat Persaingan Siswa dalam Seleksi Nasional Masuk Perguruan Tinggi Negeri Menggunakan Algoritma Naive Bayes," *Journal of Scientech Research and Development*, vol. 6, no. 1, Art. no. 1, Jul 2024, doi: 10.56670/jsrd.v6i1.337.

[20] A. S. B. Asmoro, W. S. G. Irianto, dan U. Pujianto, "Perbandingan Kinerja Hasil Seleksi Fitur pada Prediksi Kinerja Akademik Siswa Berbasis Pohon Keputusan," *JEPIN*, vol. 4, no. 2, hlm. 84, Des 2018, doi: 10.26418/jp.v4i2.29294.

[21] R. M. Apsari, "Penerapan Metode Naïve bayes dalam Memprediksi Prestasi Siswa," *Jurnal Pustaka AI (Pusat Akses Kajian Teknologi Artificial Intelligence)*, vol. 4, no. 2, Art. no. 2, Agu 2024, doi: 10.55382/jurnalpustakaai.v3i3.760.