

Implementasi Naive Bayes untuk Memprediksi Prestasi Belajar Siswa MTs Fathurrahman Padang Tualang

Alfin Noval Permana¹, Juwita Adinda^{2*}, Roberto Kaban³

¹⁻³Fakultas Sains dan Teknologi, Program Studi Sistem Informasi, Institut Teknologi dan Bisnis Indonesia

Email: alfinbaik68@gmail.com¹, juwitaadinda930@gmail.com², roberto.kaban@yahoo.com³

*Penulis Korespondensi: juwitaadinda930@gmail.com

Abstract. Student learning achievement is an important indicator in evaluating the success of the learning process in madrasah. This study aims to implement the Naive Bayes algorithm in predicting the learning achievement of Grade VII, VIII, and IX students at MTs Fathurrahman Padang Tualang. The research data uses 77 students from three grade levels; 38 students (49.35%) are classified as Achieving and 39 students (50.65%) as Underachieving. Model evaluation using the Hold-Out Split Data method (80% training, 20% testing) achieved Accuracy of 93.75%, Precision of 100.00%, and Recall of 87.50%, confirming the model's high reliability. The UAS variable is the strongest predictor with a mean difference of 13.59 points between classes ($\mu_{\text{Achieving}} = 80.92$ vs $\mu_{\text{Underachieving}} = 67.33$). This research proves that Naive Bayes is an effective and efficient classification algorithm for predicting student learning achievement across grade levels in madrasah tsanawiyah. The proposed model can support educators in identifying students with potential academic difficulties, enabling early intervention and more targeted learning strategies. Furthermore, the implementation of predictive analytics provides valuable insights for improving academic management and supporting data-driven decision-making in educational institutions.

Keywords: Academic Prediction; Learning Achievement; Madrasah Tsanawiyah; Naive Bayes; Student Classification.

Abstrak. Prestasi belajar siswa merupakan indikator penting dalam mengevaluasi keberhasilan proses pembelajaran di madrasah. Penelitian ini bertujuan untuk mengimplementasikan algoritma *Naive Bayes* dalam memprediksi prestasi belajar siswa kelas VII, VIII, dan IX MTs Fathurrahman Padang Tualang secara menyeluruh. Data penelitian menggunakan 77 siswa dari tiga jenjang kelas dengan variabel kehadiran, keaktifan, nilai UTS, dan nilai UAS sebagai atribut prediksi, serta status berprestasi dan kurang berprestasi sebagai kelas target. Distribusi data terdiri atas 25 siswa kelas VII, 32 siswa kelas VIII, dan 20 siswa kelas IX, dengan 38 siswa (49,35%) terklasifikasi Berprestasi dan 39 siswa (50,65%) Kurang Berprestasi. Evaluasi model menggunakan metode *Hold-Out Split Data* (80% training, 20% testing) menghasilkan Akurasi 93,75%, Presisi 100,00%, dan Recall 87,50%, yang mengonfirmasi keandalan tinggi model. Variabel UAS menjadi prediktor terkuat dengan selisih mean antar kelas sebesar 13,59 poin ($\mu_{\text{Berprestasi}} = 80,92$ vs $\mu_{\text{Kurang Berprestasi}} = 67,33$). Penelitian ini membuktikan bahwa *Naive Bayes* merupakan algoritma klasifikasi yang efektif dan efisien untuk memprediksi prestasi belajar siswa lintas jenjang kelas pada madrasah tsanawiyah.

Kata kunci: Algoritma Klasifikasi; Madrasah Tsanawiyah; Naive Bayes; Prediksi Prestasi; Prestasi Belajar Siswa.

1. LATAR BELAKANG

Prestasi belajar menjadi salah satu indikator yang digunakan sekolah untuk menilai perkembangan akademik siswa sekaligus mengevaluasi ketercapaian proses pembelajaran. Informasi mengenai kondisi akademik sebenarnya dapat diperoleh dari berbagai data yang telah dihimpun selama kegiatan belajar, seperti tingkat kehadiran, keaktifan siswa, nilai Ujian Tengah Semester (UTS), dan nilai Ujian Akhir Semester (UAS). Akan tetapi, data tersebut sering kali masih dimanfaatkan sebatas pencatatan dan pelaporan hasil belajar. Ketika jumlah siswa berasal dari beberapa jenjang kelas, proses identifikasi siswa yang berpotensi mengalami penurunan capaian akademik menjadi lebih kompleks apabila hanya dilakukan melalui

pemeriksaan manual. Kondisi ini memperlihatkan perlunya pendekatan analitis yang mampu mengolah data akademik menjadi informasi prediktif bagi sekolah (Sukmawati & Hermawan, 2023).

Pemanfaatan teknologi informasi dalam bidang pendidikan memberikan peluang untuk melakukan analisis terhadap pola capaian siswa melalui teknik *data mining*. Salah satu tugas utama dalam *data mining* adalah klasifikasi, yaitu proses membentuk pola dari data terdahulu untuk menentukan kategori pada data yang dianalisis. Pendekatan ini relevan bagi pengelolaan data pendidikan karena karakteristik akademik siswa dapat dipetakan ke dalam kelas tertentu berdasarkan atribut yang tersedia. Sejumlah metode telah digunakan untuk kebutuhan prediksi akademik, di antaranya *Decision Tree*, *k-Nearest Neighbor*, *Support Vector Machine*, dan Naive Bayes. Setiap metode mempunyai karakteristik berbeda, namun Naive Bayes menarik digunakan karena proses klasifikasinya relatif sederhana, berbasis probabilitas, dan dapat diterapkan pada data dengan jumlah terbatas (Ramadhani & Fitria, 2022; Septiani & Kurniawan, 2021).

Naive Bayes menentukan kelas suatu data melalui probabilitas posterior berdasarkan Teorema Bayes dengan asumsi independensi antar fitur. Pada data akademik yang berbentuk numerik kontinu, seperti persentase kehadiran dan nilai ujian, pendekatan Gaussian Naive Bayes dapat digunakan untuk memodelkan distribusi setiap atribut pada masing-masing kelas. Karakteristik tersebut membuat metode ini sesuai untuk mengolah data siswa yang memiliki variasi capaian akademik. Selain menghasilkan kategori prediksi, pendekatan probabilistik juga memungkinkan proses klasifikasi ditelusuri melalui nilai probabilitas yang mendasari keputusan model (Rish, 2021; Zhang, 2021).

Penerapan metode klasifikasi untuk memprediksi prestasi akademik telah dibahas dalam beberapa penelitian sebelumnya. Nugroho dan Prasetyo (2022) menerapkan Naive Bayes pada prediksi kelulusan mahasiswa berdasarkan data akademik dan memperoleh akurasi sebesar 84,7%. Pada tingkat sekolah dasar, Wijaya et al. (2023) melaporkan capaian akurasi 80,2% dalam prediksi prestasi belajar siswa. Penelitian Hasanah dan Mulyani (2024) menggunakan Gaussian Naive Bayes pada siswa sekolah menengah pertama dengan akurasi 82,5%. Studi lain oleh Santoso et al. (2024) membandingkan sejumlah algoritma *machine learning* untuk prediksi performa akademik siswa sekolah menengah di Indonesia dan menempatkan Naive Bayes sebagai salah satu metode dengan kinerja yang kompetitif. Sementara itu, Wahyuningsih et al. (2025) menunjukkan bahwa Gaussian Naive Bayes dapat diterapkan pada konteks siswa madrasah tsanawiyah berbasis data historis akademik.

Walaupun berbagai penelitian tersebut memperlihatkan potensi Naive Bayes dalam prediksi akademik, masih terdapat ruang kajian pada penerapan model yang menggabungkan siswa dari beberapa tingkat kelas dalam satu analisis pada lingkungan madrasah tsanawiyah. Sejumlah studi terdahulu berfokus pada jenjang atau kelompok pendidikan tertentu, sedangkan karakteristik siswa antarkelas dapat memiliki variasi yang berbeda. Selain itu, penerapan prediksi perlu disertai pengujian pada data yang dipisahkan dari data pelatihan agar kemampuan generalisasi model dapat dinilai secara lebih terukur. Celah tersebut menjadi dasar penelitian ini dengan menggabungkan data siswa kelas VII, VIII, dan IX serta melakukan evaluasi model melalui skema *Hold-Out Split Data*. Pendekatan ini diarahkan untuk memberikan gambaran prediksi prestasi belajar lintas jenjang berdasarkan karakteristik akademik siswa.

MTs Fathurrahman Padang Tualang memiliki data akademik siswa yang mencakup tiga jenjang aktif, yaitu kelas VII, VIII, dan IX. Data yang tersedia meliputi kehadiran, keaktifan, nilai UTS, dan nilai UAS, namun pemanfaatannya sebagai dasar prediksi prestasi belajar masih perlu dikembangkan. Penelitian ini menggunakan 77 data siswa yang terdiri atas 25 siswa kelas VII, 32 siswa kelas VIII, dan 20 siswa kelas IX. Berdasarkan kondisi tersebut, penelitian bertujuan mengimplementasikan algoritma Gaussian Naive Bayes untuk memprediksi prestasi belajar siswa MTs Fathurrahman Padang Tualang ke dalam kategori berprestasi dan kurang berprestasi. Kinerja model selanjutnya dievaluasi menggunakan *confusion matrix* melalui pengukuran akurasi, presisi, dan *recall*, agar kemampuan klasifikasi model dapat dinilai berdasarkan hasil prediksi pada data pengujian.

2. KAJIAN TEORITIS

Prestasi Belajar Siswa

Prestasi belajar siswa adalah gambaran capaian akademik yang diperoleh setelah mengikuti proses pembelajaran dalam periode tertentu dan dapat digunakan untuk menilai perkembangan penguasaan siswa terhadap materi yang telah dipelajari. Prestasi tersebut dapat diamati melalui data akademik yang merepresentasikan proses maupun hasil pembelajaran, seperti kehadiran, keaktifan, serta capaian evaluasi berupa nilai UTS dan UAS. Ramadhani dan Fitria (2022) menjelaskan bahwa data akademik dapat dimanfaatkan untuk mengenali pola prestasi siswa melalui pendekatan analitis, sedangkan Prasetya dan Wibowo (2023) menekankan penggunaan model prediktif dalam mengolah informasi akademik siswa pada berbagai jenjang pendidikan. Pada konteks madrasah, Mardiyanto et al. (2025) juga

menunjukkan bahwa data pendidikan dapat digunakan sebagai dasar untuk memperkirakan keberhasilan akademik melalui metode klasifikasi probabilistik.

Berdasarkan pandangan tersebut, penelitian ini memaknai prestasi belajar tidak hanya sebagai nilai akhir yang berdiri sendiri, tetapi sebagai kondisi akademik yang dapat dikenali melalui keterkaitan beberapa atribut siswa. Kehadiran dan keaktifan dipandang mewakili keterlibatan dalam pembelajaran, sedangkan nilai UTS dan UAS mencerminkan capaian evaluatif, sehingga kombinasi keempat atribut tersebut digunakan untuk membentuk dasar prediksi kategori prestasi belajar siswa MTs Fathurrahman Padang Tualang.

Data Mining dan Klasifikasi

Data mining merupakan proses analisis terhadap sekumpulan data untuk menemukan pola, hubungan, dan informasi yang dapat dimanfaatkan dalam mendukung pengambilan keputusan. Salah satu teknik yang digunakan dalam *data mining* adalah klasifikasi, yaitu proses mempelajari pola dari data yang telah memiliki label untuk menentukan kategori pada data lainnya berdasarkan karakteristik tertentu. Dalam bidang pendidikan, teknik ini dapat dimanfaatkan untuk mengolah data akademik menjadi informasi prediktif mengenai capaian siswa. Ramadhani dan Fitria (2022) menjelaskan bahwa pendekatan *data mining* dapat digunakan dalam prediksi prestasi akademik melalui pemanfaatan pola yang terkandung pada data pendidikan. Sejalan dengan hal tersebut, Septiani dan Kurniawan (2021) menunjukkan bahwa metode klasifikasi memiliki peran dalam memprediksi prestasi siswa berdasarkan karakteristik data yang dianalisis. Alsaaidah et al. (2022) juga mengaitkan *educational data mining* dengan prediksi performa akademik siswa melalui penerapan algoritma klasifikasi Naive Bayes.

Dari konsep tersebut, peneliti memandang bahwa data akademik siswa tidak seharusnya hanya berfungsi sebagai arsip administratif, tetapi dapat diolah menjadi informasi yang membantu mengenali kecenderungan prestasi belajar. Dalam penelitian ini, klasifikasi digunakan untuk mempelajari pola dari atribut kehadiran, keaktifan, nilai UTS, dan nilai UAS, kemudian memetakan siswa ke dalam kategori Berprestasi atau Kurang Berprestasi. Pemanfaatan beberapa atribut secara bersamaan diharapkan dapat memberikan dasar prediksi yang lebih sistematis dibandingkan penilaian yang hanya bergantung pada satu komponen akademik.

Algoritma Naive Bayes

Algoritma *Naive Bayes* merupakan metode klasifikasi probabilistik yang menggunakan Teorema Bayes untuk menentukan kemungkinan suatu data termasuk ke dalam kelas tertentu berdasarkan atribut yang dimilikinya. Proses klasifikasi dilakukan dengan mempertimbangkan

probabilitas awal (*prior*), peluang kemunculan atribut pada setiap kelas (*likelihood*), dan probabilitas akhir (*posterior*) sebagai dasar penentuan hasil prediksi. Metode ini menggunakan asumsi bahwa setiap atribut bersifat independen secara kondisional terhadap kelas target. Rish (2021) menjelaskan bahwa *Naive Bayes* memiliki mekanisme klasifikasi yang relatif sederhana karena memanfaatkan perhitungan probabilitas berdasarkan kontribusi setiap atribut. Zhang (2021) juga membahas karakteristik statistik *Naive Bayes* dan menunjukkan bahwa asumsi independensi yang digunakan tidak selalu menghilangkan kemampuan model dalam menghasilkan keputusan klasifikasi yang baik. Dalam konteks pendidikan, Alsaaidah et al. (2022) menerapkan *Naive Bayes classifier* untuk memprediksi performa akademik siswa melalui pengolahan data pendidikan, sedangkan Ardiansyah dan Suryana (2023) menggunakan algoritma yang sama untuk klasifikasi prestasi belajar siswa pada jenjang sekolah menengah pertama.

Dengan konsep tersebut, peneliti memandang bahwa *Naive Bayes* sesuai digunakan untuk memprediksi prestasi belajar karena mampu mengolah sejumlah karakteristik akademik ke dalam keputusan kelas berbasis probabilitas. Pada penelitian ini, atribut kehadiran, keaktifan, nilai UTS, dan nilai UAS digunakan sebagai informasi masukan untuk memperkirakan kemungkinan siswa termasuk kategori Berprestasi atau Kurang Berprestasi. Pendekatan tersebut dipilih agar proses prediksi tidak hanya bergantung pada satu indikator akademik, melainkan mempertimbangkan pola probabilitas dari keseluruhan atribut yang tersedia pada data siswa MTs Fathurrahman Padang Tualang.

Gaussian Naive Bayes

Gaussian Naive Bayes merupakan pengembangan dari algoritma *Naive Bayes* yang digunakan ketika atribut prediksi berbentuk data numerik kontinu. Metode ini mengasumsikan bahwa nilai setiap atribut pada masing-masing kelas mengikuti distribusi Gaussian atau distribusi normal, yang karakteristiknya ditentukan melalui nilai rata-rata (*mean*) dan standar deviasi. Dengan pendekatan tersebut, peluang kemunculan suatu nilai atribut terhadap kelas tertentu dapat dihitung menggunakan fungsi densitas probabilitas Gaussian sebelum digabungkan dengan probabilitas *prior* untuk memperoleh nilai *posterior*. Fitriani et al. (2024) menerapkan *Gaussian Naive Bayes* dalam klasifikasi prestasi akademik dengan memanfaatkan karakteristik data numerik, sedangkan Hasanah dan Mulyani (2024) menggunakan pendekatan serupa untuk memprediksi prestasi belajar siswa SMP berdasarkan data nilai akademik. Pada konteks pendidikan madrasah, Wahyuningsih et al. (2025) juga mengaitkan penerapan *Gaussian Naive Bayes* dengan klasifikasi potensi akademik siswa madrasah tsanawiyah berbasis data historis.

Dari karakteristik tersebut, peneliti memandang bahwa *Gaussian Naive Bayes* relevan dengan bentuk data yang digunakan dalam penelitian ini karena atribut kehadiran, keaktifan, nilai UTS, dan nilai UAS dinyatakan dalam bentuk numerik. Setiap atribut memiliki variasi nilai yang berbeda pada kelompok siswa Berprestasi dan Kurang Berprestasi, sehingga penggunaan distribusi Gaussian dapat menjadi dasar untuk memperkirakan peluang suatu nilai muncul pada masing-masing kelas. Melalui pendekatan ini, proses prediksi prestasi belajar siswa MTs Fathurrahman Padang Tualang dilakukan dengan mempertimbangkan pola distribusi data akademik, bukan hanya berdasarkan perbandingan nilai secara langsung.

Penelitian Terdahulu

Penerapan algoritma *Naive Bayes* untuk analisis dan prediksi capaian akademik telah dikaji pada berbagai jenjang pendidikan. Alsaaidah et al. (2022) menerapkan *Naive Bayes classifier* dalam sistem *educational data mining* untuk memprediksi performa akademik siswa, yang menunjukkan relevansi pendekatan probabilistik dalam pengolahan data pendidikan. Pada konteks pendidikan tinggi, Nugroho dan Prasetyo (2022) memanfaatkan data akademik untuk memprediksi kelulusan mahasiswa menggunakan *Naive Bayes*, sedangkan Fitriani et al. (2024) menerapkan *Gaussian Naive Bayes* untuk klasifikasi prestasi akademik mahasiswa dengan evaluasi *k-fold cross validation*. Penelitian pada tingkat sekolah juga menunjukkan perkembangan serupa. Ardiansyah dan Suryana (2023) mengkaji klasifikasi prestasi belajar siswa sekolah menengah pertama, Hasanah dan Mulyani (2024) menerapkan *Gaussian Naive Bayes* pada prediksi prestasi siswa SMP berdasarkan nilai akademik, dan Wahyuningsih et al. (2025) membawa pendekatan tersebut ke lingkungan madrasah tsanawiyah melalui pemanfaatan data historis akademik.

Tabel 1. Perbandingan Penelitian Terdahulu

Peneliti	Fokus Penelitian	Metode	Hasil/Kontribusi Utama	Keterkaitan dengan Penelitian Ini
Alsaaidah et al. (2022)	Prediksi performa akademik siswa melalui <i>educational data mining</i>	Naive Bayes	Menunjukkan pemanfaatan klasifikasi probabilistik pada data pendidikan	Menjadi dasar penerapan Naive Bayes untuk prediksi prestasi akademik
Nugroho & Prasetyo (2022)	Prediksi kelulusan mahasiswa berdasarkan data akademik	Naive Bayes	Menggunakan karakteristik akademik sebagai dasar prediksi	Memperkuat penggunaan data akademik untuk pembentukan kelas prediksi
Ardiansyah & Suryana (2023)	Klasifikasi prestasi belajar siswa SMP	Naive Bayes	Menerapkan klasifikasi pada konteks siswa sekolah menengah	Memiliki kedekatan objek, tetapi penelitian ini mencakup tiga jenjang kelas MTs

Hasanah & Mulyani (2024)	Prediksi prestasi siswa SMP berdasarkan nilai akademik	Gaussian Naive Bayes	Memfaatkan atribut numerik untuk prediksi prestasi	Mendukung pemilihan Gaussian Naive Bayes pada data akademik numerik
Fitriani et al. (2024)	Klasifikasi prestasi akademik mahasiswa	Gaussian Naive Bayes	Mengevaluasi model menggunakan <i>k-fold cross validation</i>	Menjadi pembanding dari sisi metode evaluasi model
Wahyuningsih et al. (2025)	Klasifikasi potensi akademik siswa MTs berbasis data historis	Gaussian Naive Bayes	Menunjukkan penerapan metode pada lingkungan madrasah tsanawiyah	Paling dekat dari sisi objek, sedangkan penelitian ini menggabungkan kelas VII, VIII, dan IX

Sumber: Data Penelitian 2026

Berdasarkan perbandingan tersebut, penelitian terdahulu memperlihatkan bahwa *Naive Bayes* dan *Gaussian Naive Bayes* telah digunakan untuk memprediksi capaian akademik pada tingkat mahasiswa, sekolah dasar, sekolah menengah, hingga madrasah. Meskipun demikian, fokus penelitian sebelumnya cenderung berada pada kelompok atau konteks pendidikan tertentu dan belum menempatkan data siswa kelas VII, VIII, dan IX secara bersamaan dalam satu proses klasifikasi pada MTs Fathurrahman Padang Tualang. Atas dasar itu, penelitian ini diarahkan pada penerapan *Gaussian Naive Bayes* terhadap 77 data siswa lintas jenjang dengan atribut kehadiran, keaktifan, nilai UTS, dan nilai UAS, kemudian mengevaluasi kemampuan model melalui skema *Hold-Out Split Data*. Posisi tersebut menjadi pembeda penelitian ini karena prediksi dilakukan pada gabungan tiga jenjang kelas dalam satu model klasifikasi, bukan hanya pada satu kelompok siswa tertentu.

3. METODE PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif dengan teknik *data mining* untuk mengklasifikasikan prestasi belajar siswa menggunakan algoritma *Gaussian Naive Bayes*. Penelitian dilakukan pada MTs Fathurrahman Padang Tualang dengan memanfaatkan data akademik tahun ajaran 2025/2026. Populasi penelitian mencakup seluruh siswa aktif kelas VII, VIII, dan IX sebanyak 77 siswa, terdiri atas 25 siswa kelas VII, 32 siswa kelas VIII, dan 20 siswa kelas IX. Seluruh populasi digunakan sebagai sampel melalui teknik *total sampling*. Data dikumpulkan dengan teknik dokumentasi dari arsip akademik sekolah. Atribut prediksi yang digunakan meliputi kehadiran, keaktifan, nilai Ujian Tengah Semester (UTS), dan nilai Ujian Akhir Semester (UAS), sedangkan kelas target dibedakan menjadi Berprestasi untuk rata-rata nilai ≥ 80 dan Kurang Berprestasi untuk rata-rata nilai < 80 .

Tahapan analisis dimulai dengan pemeriksaan data akademik, penentuan kelas target, pembagian data menjadi data *training* dan *testing*, pembentukan model *Gaussian Naive Bayes*, proses prediksi, serta evaluasi performa model. Secara konseptual, algoritma *Naive Bayes* menentukan probabilitas suatu kelas C berdasarkan sekumpulan atribut X menggunakan Teorema Bayes. Hubungan probabilitas tersebut dirumuskan sebagai berikut (Rish, 2021):

$$P(C | X) = P(X)P(X | C) \times P(C) \quad (1)$$

$P(C | X)$ merupakan probabilitas *posterior* kelas C berdasarkan data X , $P(X | C)$ adalah probabilitas data X pada kelas C atau *likelihood*, $P(C)$ menunjukkan probabilitas awal kelas atau *prior*, sedangkan $P(X)$ merupakan probabilitas data pengamatan. Karena nilai $P(X)$ bersifat konstan pada proses perbandingan antarkelas, persamaan klasifikasi dapat disederhanakan menjadi:

$$P(C | X) \propto P(C) \times P(X | C) = P(C) \times \prod_{i=1}^n P(x_i | C) \quad (2)$$

Data penelitian terdiri atas atribut numerik kontinu, sehingga perhitungan *likelihood* dilakukan menggunakan pendekatan Gaussian. Pada setiap kelas, distribusi atribut dibentuk berdasarkan nilai rata-rata dan standar deviasi. Fungsi densitas probabilitas Gaussian dirumuskan sebagai berikut (Zhang, 2021):

$$P(x_i | C) = \frac{1}{\sigma_C \sqrt{2\pi}} \exp\left(-\frac{(x_i - \mu_C)^2}{2\sigma_C^2}\right) \quad (3)$$

dengan nilai rata-rata dihitung menggunakan:

$$\mu = \frac{1}{n} \sum_{i=1}^n x_i \quad (4)$$

dan standar deviasi dihitung menggunakan:

$$\sigma = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \mu)^2} \quad (5)$$

Evaluasi model dilakukan menggunakan skema *Hold-Out Split Data* dengan rasio 80:20. Sebanyak 61 data siswa digunakan sebagai data *training* dan 16 data lainnya sebagai data *testing*. Hasil prediksi pada data pengujian dievaluasi menggunakan *confusion matrix* yang terdiri atas *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN). Performa model diukur melalui akurasi, presisi, dan *recall* (Prasetya & Wibowo, 2023). Akurasi dihitung menggunakan:

$$\text{Akurasi} = \frac{TP+TN}{TP+TN+FP+FN} \times 100\% \quad (6)$$

Presisi dihitung menggunakan:

$$\text{Presisi} = \frac{TP}{TP+FP} \times 100\% \quad (7)$$

sedangkan *recall* dihitung menggunakan:

$$\text{Recall} = \frac{TP}{TP+FN} \times 100\% \quad (8)$$

Nilai akurasi digunakan untuk mengukur proporsi keseluruhan prediksi yang benar, presisi menunjukkan ketepatan model dalam menghasilkan prediksi kelas positif, sedangkan *recall* mengukur kemampuan model dalam mengenali data positif yang sebenarnya. Rangkaian analisis tersebut digunakan untuk menilai kemampuan *Gaussian Naive Bayes* dalam memprediksi prestasi belajar siswa berdasarkan pola atribut akademik yang tersedia.

4. HASIL DAN PEMBAHASAN

Deskripsi Data Penelitian

Data penelitian mencakup 77 siswa aktif MTs Fathurrahman Padang Tualang yang berasal dari tiga jenjang kelas, yaitu kelas VII, VIII, dan IX. Berdasarkan komposisi data, kelas VII terdiri atas 25 siswa, kelas VIII sebanyak 32 siswa, dan kelas IX sebanyak 20 siswa. Setiap data siswa memuat empat atribut akademik berupa kehadiran, keaktifan, nilai Ujian Tengah Semester (UTS), dan nilai Ujian Akhir Semester (UAS). Keempat atribut tersebut digunakan sebagai masukan dalam proses klasifikasi, sedangkan status prestasi belajar ditetapkan sebagai kelas target yang terbagi menjadi Berprestasi dan Kurang Berprestasi. Distribusi siswa berdasarkan jenjang kelas dan status prestasi disajikan pada Tabel 2.

Tabel 2. Distribusi Siswa Berdasarkan Jenjang Kelas

Kelas	Jumlah Siswa	Berprestasi	Kurang Berprestasi
VII	25	12 (48,00%)	13 (52,00%)
VIII	32	16 (50,00%)	16 (50,00%)
IX	20	10 (50,00%)	10 (50,00%)
Total	77	38 (49,35%)	39 (50,65%)

Sumber: Data Akademik MTs Fathurrahman Padang Tualang, 2026

Berdasarkan Tabel 2, sebanyak 38 siswa atau 49,35% termasuk kategori Berprestasi, sedangkan 39 siswa atau 50,65% berada pada kategori Kurang Berprestasi. Pada kelas VII, jumlah siswa Kurang Berprestasi sedikit lebih tinggi, yaitu 13 siswa dibandingkan 12 siswa Berprestasi. Sementara itu, kelas VIII memiliki komposisi yang sama antara kedua kategori, masing-masing sebanyak 16 siswa, dan kelas IX juga menunjukkan pembagian seimbang dengan masing-masing 10 siswa. Komposisi keseluruhan memperlihatkan selisih hanya satu siswa antara kedua kelas target. Data dan persentase ini sama dengan distribusi yang tercantum pada tabel hasil di halaman 3 file sumber terbaru.

Perhitungan Probabilitas Prior

Probabilitas *prior* digunakan untuk menunjukkan peluang awal setiap kelas sebelum atribut akademik siswa diperhitungkan dalam proses klasifikasi. Berdasarkan 77 data siswa, sebanyak 38 siswa termasuk kategori Berprestasi, sedangkan 39 siswa berada pada kategori Kurang Berprestasi. Nilai probabilitas *prior* masing-masing kelas dihitung berdasarkan proporsi jumlah anggota kelas terhadap keseluruhan data. Perhitungan tersebut menghasilkan:

$$P(\text{Berprestasi}) = \frac{38}{77} = 0,4935 \quad (9)$$

$$P(\text{Kurang Berprestasi}) = \frac{39}{77} = 0,5065 \quad (10)$$

Hasil perhitungan menunjukkan bahwa probabilitas awal kelas Berprestasi sebesar 0,4935 atau 49,35%, sedangkan kelas Kurang Berprestasi sebesar 0,5065 atau 50,65%. Selisih probabilitas kedua kelas hanya sebesar 0,0130 atau 1,30 poin persentase, yang memperlihatkan bahwa komposisi kelas target pada dataset relatif berimbang. Kondisi tersebut membuat probabilitas awal kedua kelas memiliki bobot yang hampir sama dalam proses klasifikasi, sehingga penentuan hasil prediksi berikutnya lebih banyak dipengaruhi oleh nilai *likelihood* dari atribut kehadiran, keaktifan, UTS, dan UAS. Nilai *prior* ini tetap mengikuti hasil perhitungan pada file sumber terbaru.

Perhitungan Mean dan Standar Deviasi

Tahap berikutnya dilakukan dengan menghitung parameter distribusi Gaussian pada setiap atribut berdasarkan kelas Berprestasi dan Kurang Berprestasi. Parameter yang digunakan meliputi nilai rata-rata (*mean*) (μ) dan standar deviasi (σ), yang menjadi dasar dalam menentukan nilai *likelihood* setiap atribut pada proses klasifikasi *Gaussian Naive Bayes*. Perhitungan dilakukan terhadap empat atribut akademik, yaitu kehadiran, keaktifan, nilai UTS, dan nilai UAS. Hasil parameter distribusi masing-masing kelas disajikan pada Tabel 3 dan tetap mengikuti nilai pada file sumber.

Tabel 3. Mean dan Standar Deviasi Atribut Berdasarkan Kelas

Atribut	μ Berprestasi	σ Berprestasi	μ Kurang Berprestasi	σ Kurang Berprestasi
Kehadiran	83,45	3,57	78,77	3,00
Keaktifan	83,89	3,34	76,87	5,02
UTS	80,89	3,09	70,95	6,15
UAS	80,92	3,81	67,33	7,89

Sumber: Hasil Perhitungan Python, 2026

Berdasarkan Tabel 3, seluruh atribut pada kelas Berprestasi memiliki nilai *mean* lebih tinggi dibandingkan kelas Kurang Berprestasi. Pada atribut kehadiran, nilai rata-rata kedua kelas masing-masing sebesar 83,45 dan 78,77 dengan selisih 4,68 poin. Atribut keaktifan memiliki rata-rata 83,89 pada kelas Berprestasi dan 76,87 pada kelas Kurang Berprestasi, dengan perbedaan 7,02 poin. Selisih yang lebih besar terlihat pada nilai UTS, yaitu 80,89

berbanding 70,95 atau berbeda 9,94 poin. Perbedaan paling tinggi terdapat pada atribut UAS dengan rata-rata 80,92 untuk kelas Berprestasi dan 67,33 untuk kelas Kurang Berprestasi, menghasilkan selisih sebesar 13,59 poin. Pola tersebut menunjukkan bahwa UAS memberikan pemisahan nilai rata-rata paling besar di antara kedua kelas pada data penelitian ini.

Ditinjau dari penyebaran data, kelas Kurang Berprestasi menunjukkan standar deviasi yang lebih besar pada atribut keaktifan ($\sigma = 5,02$), UTS ($\sigma = 6,15$), dan UAS ($\sigma = 7,89$) dibandingkan kelas Berprestasi yang masing-masing bernilai 3,34, 3,09, dan 3,81. Temuan ini memperlihatkan bahwa variasi nilai pada kelompok Kurang Berprestasi cenderung lebih lebar untuk ketiga atribut tersebut. Sebaliknya, standar deviasi kehadiran pada kelas Berprestasi sebesar 3,57, sedikit lebih tinggi daripada kelas Kurang Berprestasi sebesar 3,00. Dengan demikian, perbedaan antarkelas tidak hanya terlihat pada nilai rata-rata, tetapi juga pada tingkat penyebaran setiap atribut. Parameter μ dan σ tersebut selanjutnya digunakan untuk menghitung *likelihood* masing-masing data siswa dalam proses *Gaussian Naive Bayes*.

Perhitungan Likelihood dan Posterior

Tahap selanjutnya adalah menghitung nilai *likelihood* setiap atribut siswa terhadap kelas Berprestasi dan Kurang Berprestasi menggunakan fungsi distribusi Gaussian. Perhitungan dilakukan berdasarkan parameter *mean* (μ) dan standar deviasi (σ) yang telah diperoleh pada tahap sebelumnya. Nilai *likelihood* dari atribut kehadiran, keaktifan, UTS, dan UAS kemudian dikombinasikan dengan probabilitas *prior* masing-masing kelas untuk menghasilkan probabilitas *posterior*. Siswa ditetapkan ke dalam kelas yang memiliki nilai *posterior* lebih tinggi. Hasil implementasi perhitungan terhadap data siswa disajikan pada Tabel 4 dengan tetap mengikuti nilai pada file sumber.

Tabel 4. Hasil Perhitungan Likelihood dan Posterior Siswa

N o	Kl s	Nama	P(Had B)	P(Akt B)	P(UTS B)	P(UAS B)	Post B	P(Had KB)	P(Akt KB)	P(UTS KB)	P(UAS KB)	Post KB	Predik si	Bena r?
1	VI I	VITHA ZAHFIRA	0,0701 1	0,101 62	0,1237 6	0,1017 2	4,43 E- 05	0,12241	0,04716	0,02195	0,01392	8,94 E- 07	Berpres tasi	✓
2	VI I	ARGA DINATA	0,0883 6	0,060 54	0,1069 5	0,0902 5	2,55 E- 05	0,10093	0,06547	0,02752	0,00703	6,48 E- 07	Berpres tasi	✓
3	VI II	ABIZAR PRANAN DA	0,0701 1	0,025 22	0,0832 5	0,1017 2	7,39 E- 06	0,12241	0,07754	0,03361	0,01392	2,25 E- 06	Berpres tasi	✗
4	VI II	ARDIANS YAH PRATAM A	0,0701 1	0,060 54	0,1069 5	0,0312 9	7,01 E- 06	0,12241	0,06547	0,02752	0,03154	3,52 E- 06	Berpres tasi	✗
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
7 6	IX	ZHAIRA SYAFAN A	0,0000 9	0	0,0000 1	0,0000 2	5,63 E- 23	0,00183	0,00484	0,05280	0,04843	1,15 E- 08	Kurang B.	✓
7 7	IX	IKHSAN DWIAND RA TARIGAN	0,0883 6	0,119 27	0,1237 6	0,0922 3	5,94 E- 05	0,10093	0,02898	0,02195	0,01693	5,51 E- 07	Berpres tasi	✓

Sumber: Hasil Implementasi Python Gaussian Naive Bayes, 2026

Berdasarkan Tabel 4, hasil klasifikasi ditentukan melalui perbandingan nilai *posterior* pada kedua kelas. Sebagai contoh, Vitha Zahfira memperoleh nilai *posterior* kelas Berprestasi sebesar $4,43 \times 10^{-5}$, lebih tinggi dibandingkan *posterior* kelas Kurang Berprestasi sebesar $8,94 \times 10^{-7}$, sehingga diprediksi sebagai siswa Berprestasi. Pola serupa terlihat pada Arga Dinata dengan nilai *posterior* Berprestasi sebesar $2,55 \times 10^{-5}$, sedangkan nilai *posterior* Kurang Berprestasi hanya $6,48 \times 10^{-7}$. Pada data lain, Zhaira Syafana memiliki nilai *posterior* Kurang Berprestasi sebesar $1,15 \times 10^{-8}$, jauh lebih tinggi dibandingkan *posterior* Berprestasi sebesar $5,63 \times 10^{-23}$, sehingga model menetapkan ke kelas Kurang Berprestasi. Nilai-nilai tersebut mengikuti tabel perhitungan pada halaman 4 file sumber.

Dari keseluruhan 77 data siswa, model menghasilkan 73 prediksi yang sesuai dengan label aktual dan 4 prediksi yang tidak sesuai. Dengan demikian, tingkat kecocokan prediksi pada keseluruhan data mencapai 94,81%, sedangkan proporsi kesalahan sebesar 5,19%. Empat kasus ketidaksesuaian prediksi ditemukan pada Abizar Prananda, Ardiansyah Pratama, Muhammad Guntur Maulana, dan Mia Zahira, yang seluruhnya berasal dari kelas VIII. Berdasarkan file sumber, keempat data tersebut berada di sekitar batas penentuan kategori prestasi, yaitu rata-rata nilai 80, sehingga karakteristik akademiknya berada pada area yang relatif sulit dipisahkan secara tegas oleh model.

Evaluasi Model dengan Confusion Matrix

Evaluasi model dilakukan untuk mengukur kemampuan *Gaussian Naive Bayes* dalam mengklasifikasikan data yang tidak digunakan pada proses pelatihan. Sesuai dengan rancangan penelitian, dataset dibagi menggunakan skema *Hold-Out Split Data* dengan rasio 80:20, yaitu 61 siswa sebagai data *training* dan 16 siswa sebagai data *testing*. Hasil prediksi terhadap data *testing* kemudian dibandingkan dengan kelas aktual menggunakan *confusion matrix*. Berdasarkan pengujian, diperoleh 7 data *True Positive* (TP), 8 data *True Negative* (TN), tidak terdapat *False Positive* (FP), dan terdapat 1 data *False Negative* (FN). Hasil tersebut disajikan pada Tabel 5 dan tetap mengikuti data evaluasi dalam file sumber.

Tabel 5. Confusion Matrix Evaluasi Model Gaussian Naive Bayes

Aktual \ Prediksi	Prediksi: Berprestasi	Prediksi: Kurang Berprestasi
Berprestasi	7 (<i>True Positive</i> / TP)	1 (<i>False Negative</i> / FN)
Kurang Berprestasi	0 (<i>False Positive</i> / FP)	8 (<i>True Negative</i> / TN)

Sumber: Hasil Evaluasi Model, 2026

Berdasarkan nilai pada *confusion matrix*, performa model dihitung menggunakan metrik akurasi, presisi, dan *recall*. Perhitungan akurasi dilakukan dengan membandingkan jumlah prediksi benar terhadap keseluruhan data *testing*, sebagai berikut:

$$\text{Akurasi} = \frac{7+8}{16} \times 100\% = \frac{15}{16} \times 100\% = 93,75\% \quad (11)$$

Hasil tersebut menunjukkan bahwa model mampu mengklasifikasikan 15 dari 16 data *testing* secara tepat. Selanjutnya, nilai presisi dihitung berdasarkan perbandingan *True Positive* terhadap keseluruhan data yang diprediksi sebagai kelas positif:

$$\text{Presisi} = \frac{7}{7+0} \times 100\% = \frac{7}{7} \times 100\% = 100,00\% \quad (12)$$

Nilai presisi sebesar 100,00% menunjukkan bahwa seluruh siswa yang diprediksi model sebagai Berprestasi pada data pengujian memang termasuk dalam kelas tersebut berdasarkan label aktual. Kondisi ini didukung oleh nilai *False Positive* sebesar 0, yang berarti tidak terdapat siswa Kurang Berprestasi yang keliru diprediksi sebagai Berprestasi. Adapun nilai *recall* dihitung sebagai berikut:

$$\text{Recall} = \frac{7}{7+1} \times 100\% = \frac{7}{8} \times 100\% = 87,50\% \quad (13)$$

Nilai *recall* sebesar 87,50% memperlihatkan bahwa model berhasil mengenali 7 dari 8 siswa yang secara aktual berada pada kategori Berprestasi, sedangkan satu siswa lainnya diklasifikasikan sebagai Kurang Berprestasi. Secara keseluruhan, kombinasi akurasi 93,75%, presisi 100,00%, dan *recall* 87,50% menunjukkan bahwa model memiliki tingkat ketepatan klasifikasi yang tinggi pada 16 data pengujian. Seluruh angka evaluasi ini sama dengan hasil *confusion matrix* dan perhitungan metrik pada file sumber.

5. KESIMPULAN DAN SARAN

Penerapan *Gaussian Naive Bayes* pada 77 data siswa kelas VII, VIII, dan IX MTs Fathurrahman Padang Tualang menunjukkan bahwa pola akademik berdasarkan kehadiran, keaktifan, nilai UTS, dan nilai UAS dapat digunakan untuk memprediksi kategori prestasi belajar. Dari keseluruhan data, sebanyak 38 siswa atau 49,35% termasuk kategori Berprestasi dan 39 siswa atau 50,65% berada pada kategori Kurang Berprestasi. Hasil pengujian menggunakan skema *Hold-Out Split Data* dengan 61 data *training* dan 16 data *testing* menghasilkan akurasi 93,75%, presisi 100,00%, dan *recall* 87,50%. Pada parameter distribusi Gaussian, atribut UAS memperlihatkan selisih *mean* antarkelas paling besar, yaitu 13,59 poin, dengan rata-rata 80,92 pada kelas Berprestasi dan 67,33 pada kelas Kurang Berprestasi. Hasil tersebut menunjukkan bahwa *Gaussian Naive Bayes* mampu memberikan performa klasifikasi

yang tinggi pada data pengujian dan dapat dimanfaatkan sebagai pendekatan analitis untuk mendukung identifikasi prestasi belajar siswa di lingkungan madrasah.

Penelitian ini masih terbatas pada 77 siswa dari satu madrasah dan menggunakan satu kali pembagian data melalui skema *hold-out*, sehingga hasil yang diperoleh perlu dipahami sesuai ruang lingkup data penelitian. Penelitian selanjutnya disarankan melibatkan jumlah data yang lebih besar, mencakup lebih banyak periode akademik, serta menerapkan metode evaluasi seperti *k-fold cross validation* agar kestabilan performa model dapat diuji secara lebih menyeluruh. Pengembangan berikutnya juga dapat mempertimbangkan atribut akademik tambahan dan membandingkan *Gaussian Naive Bayes* dengan algoritma klasifikasi lain untuk memperoleh gambaran performa model yang lebih komprehensif.

UCAPAN TERIMA KASIH

Penulis menyampaikan terima kasih kepada pihak MTs Fathurrahman Padang Tualang yang telah memberikan izin dan dukungan selama pelaksanaan penelitian, khususnya dalam penyediaan data akademik siswa kelas VII, VIII, dan IX tahun ajaran 2025/2026. Apresiasi juga disampaikan kepada seluruh pihak yang telah memberikan dukungan dan kontribusi dalam proses pengumpulan data, pengolahan data, hingga penyusunan hasil penelitian ini.

DAFTAR REFERENSI

- Alsaaidah, B., Al-Diabat, M., Al-Oqaily, A., & Alzu'bi, A. (2022). Educational data mining system for predicting students' academic performance using Naive Bayes classifier. *Journal of Intelligent Systems*, 31(1), 419–433.
- Ardiansyah, R., & Suryana, T. (2023). Penerapan algoritma Naive Bayes untuk klasifikasi prestasi belajar siswa sekolah menengah pertama. *Jurnal Informatika dan Sistem Informasi*, 4(2), 89–97.
- Fitriani, N., Maulana, I., & Sari, R. (2024). Klasifikasi prestasi akademik mahasiswa menggunakan Gaussian Naive Bayes dengan evaluasi k-fold cross validation. *Jurnal Ilmiah Teknologi Informasi Terapan*, 10(2), 115–123.
- Hasanah, N., & Mulyani, S. (2024). Penerapan Gaussian Naive Bayes pada prediksi prestasi belajar siswa SMP menggunakan data nilai akademik. *Jurnal Sistem Informasi dan Teknologi*, 6(1), 45–53.
- Mardiyanto, A., Kurniawan, B., & Putri, D. (2025). Implementasi data mining untuk prediksi keberhasilan akademik siswa madrasah menggunakan metode klasifikasi probabilistik. *Jurnal Pendidikan Islam dan Teknologi*, 3(1), 22–34.
- Nugroho, A., & Prasetyo, E. (2022). Implementasi Naive Bayes untuk prediksi kelulusan mahasiswa berdasarkan data akademik. *Jurnal Informatika dan Rekayasa Perangkat Lunak*, 4(1), 22–31.

- Prasetya, D. A., & Wibowo, F. (2023). Evaluasi model machine learning untuk prediksi nilai akademik siswa sekolah dasar dan menengah. *Jurnal Teknologi dan Sistem Komputer*, 11(3), 178–186.
- Ramadhani, F., & Fitria, L. (2022). Data mining untuk prediksi prestasi akademik: Tinjauan sistematis. *Jurnal Ilmu Komputer dan Informasi*, 15(1), 1–15.
- Rish, I. (2021). An empirical study of the Naive Bayes classifier. *Journal of Universal Computer Science*, 27(1), 1–15.
- Santoso, H., Prabowo, A., & Hidayah, I. (2024). Comparative study of machine learning algorithms for student academic performance prediction in Indonesian secondary schools. *Indonesian Journal of Computing and Cybernetics Systems*, 18(1), 55–67.
- Septiani, R., & Kurniawan, H. (2021). Analisis perbandingan metode klasifikasi data mining untuk prediksi prestasi siswa. *Jurnal Pendidikan Teknologi dan Kejuruan*, 27(2), 175–183.
- Sukmawati, R., & Hermawan, D. (2023). Sistem pendukung keputusan prediksi prestasi belajar siswa menggunakan Naive Bayes berbasis web. *Jurnal Rekayasa Sistem dan Teknologi Informasi*, 7(1), 40–48.
- Wahyuningsih, S., Hartati, S., & Nugroho, E. (2025). Penerapan Naive Bayes Gaussian dalam klasifikasi potensi akademik siswa madrasah tsanawiyah berbasis data historis. *Jurnal Sistem Cerdas dan Komputasi*, 8(1), 1–12.
- Wijaya, B., Santoso, E., & Purnomo, A. (2023). Prediksi prestasi belajar siswa sekolah dasar menggunakan algoritma Naive Bayes: Studi kasus di Kota Semarang. *Jurnal Ilmiah Informatika*, 11(2), 45–54.
- Zhang, H. (2021). The optimality of Naive Bayes revisited: A statistical perspective. *Computational Statistics & Data Analysis*, 154, 107098.