

Implementasi Sistem *Data Mining* Penerimaan Siswa Baru Menggunakan Metode *Naive Bayes*

Veri Arinal¹, Nurul Khoiriyah^{2*}

^{1,2*} Sekolah Tinggi Ilmu Komputer Cipta Karya Informatika, Kota Jakarta Timur, Daerah Khusus Ibukota Jakarta, Indonesia.

Corresponding Email: kotaksuratida@gmail.com^{2*}

Histori Artikel:

Dikirim 13 Agustus 2026; *Diterima dalam bentuk revisi* 16 September 2026; *Diterima* 20 September 2026; *Diterbitkan* 30 September 2026. Semua hak dilindungi oleh Lembaga Penelitian dan Pengabdian Masyarakat (LPPM) STMIK Indonesia Banda Aceh.

Abstrak

Penerimaan siswa baru merupakan tahapan awal yang sangat krusial dalam keberlangsungan proses pendidikan di MTs NU Islamiyah, Tangerang, Banten. Meskipun sekolah mampu menerima peserta didik setiap tahun, fluktuasi jumlah calon siswa antara yang masuk dan tidak masuk menimbulkan kendala operasional, khususnya dalam perencanaan kebijakan kelembagaan serta penyediaan sarana dan prasarana pembelajaran. Penelitian ini bertujuan membangun model klasifikasi untuk memprediksi status penerimaan siswa baru menggunakan metode Naive Bayes berbasis bahasa pemrograman Python. Prediksi dilakukan dengan menganalisis atribut historis siswa, meliputi jenis kelamin, status ekonomi keluarga, dan jarak tempat tinggal ke sekolah. Dataset yang dianalisis mencakup 623 data latih (*training*) dan 82 data uji (*testing*). Algoritma mengestimasi probabilitas posterior setiap kategori berdasarkan data masa lalu. Hasil pengujian menunjukkan bahwa model memiliki performa klasifikasi yang sangat baik pada kelas masuk dengan nilai *precision* sebesar 87%, *recall* 100%, dan *f1-score* 93% (*support* 71). Namun, model belum mampu memprediksi kelas tidak masuk secara optimal dengan perolehan *precision*, *recall*, dan *f1-score* sebesar 0% (*support* 11).

Kata Kunci: *Data Mining*; Naive Bayes; Prediksi; Penerimaan Siswa Baru.

Abstract

New student admission is a pivotal initial phase in sustaining educational operations at MTs NU Islamiyah, Tangerang, Banten. Although the institution enrolls students annually, fluctuations between registered and unregistered applicants present significant operational challenges, particularly in institutional policy planning and the optimal allocation of learning infrastructure. This study aims to build a classification model to predict new student admission status using the Naive Bayes method implemented in Python. Predictions were conducted by analyzing historical student attributes, including gender, family economic status, and residential distance to the school. The analyzed dataset comprised 623 training instances and 82 testing instances. The algorithm calculated posterior probabilities for each category based on historical records. The testing results demonstrated that the model achieved strong classification performance for the admitted class, yielding a precision of 87%, a recall of 100%, and an f1-score of 93% (support: 71). Conversely, the model showed poor performance in predicting the non-admitted class, resulting in 0% precision, recall, and f1-score (support: 11).

Keyword: Classification Prediction; Data Mining; Naive Bayes; Student Admission.

1. Pendahuluan

Perkembangan teknologi informasi yang berlangsung pesat telah membawa transformasi mendasar pada tata kelola operasional di berbagai sektor industri, lembaga pemerintahan, dan institusi pendidikan formal. Pemanfaatan sistem komputasi modern memberikan dampak signifikan dalam mempermudah pengelolaan administrasi akademik, salah satunya pada kegiatan penerimaan peserta didik baru yang diselenggarakan secara rutin setiap tahun ajaran (Alamsyah, 2020; Zulfauzi *et al.*, 2020). Dalam ranah pendidikan, proses penerimaan siswa baru merupakan pintu gerbang utama yang dilalui calon peserta didik untuk memasuki jenjang pembelajaran formal (Irawan & Damanik, 2022). Kelancaran dan keberhasilan pelaksanaan agenda tahunan ini memegang peranan krusial bagi keberlangsungan institusi, khususnya pada lembaga pendidikan swasta seperti Madrasah Tsanawiyah (MTs) NU Islamiyah yang berlokasi di Tangerang, Banten. Sebagai institusi yang terus berupaya meningkatkan mutu pendidikan, keberhasilan penyelenggaraan pembelajaran sangat ditunjang oleh kesiapan tenaga pengajar, ketersediaan ruang kelas, kapasitas laboratorium, serta sarana dan prasarana pendukung lainnya guna mencetak generasi siswa yang terampil dan berwawasan luas.

Meskipun MTs NU Islamiyah secara konsisten berhasil menjaring siswa baru setiap tahunnya, pihak manajemen sekolah menghadapi kendala operasional yang berulang terkait ketidakpastian jumlah pendaftar. Angka calon siswa yang mendaftar dan melanjutkan proses registrasi ulang menunjukkan pola yang sangat fluktuatif antar-periode. Kondisi ketidakpastian ini menimbulkan permasalahan pelik bagi jajaran manajemen madrasah, terutama dalam merancang kebijakan strategis jangka pendek maupun jangka menengah. Apabila lonjakan siswa yang masuk melebihi perkiraan awal, sekolah rentan mengalami kendala keterbatasan daya tampung ruang kelas, kekurangan tenaga pendidik, dan ketidaksiapan fasilitas belajar. Sebaliknya, penurunan jumlah siswa yang masuk secara drastis dapat menyebabkan inefisiensi alokasi anggaran operasional dan pemborosan inventaris yang telah disediakan. Hingga saat ini, pihak sekolah menyimpan tumpukan data pendaftaran siswa dari tahun-tahun sebelumnya dalam arsip berkas cetak maupun lembar kerja digital sederhana, namun data tersebut belum dimanfaatkan secara optimal untuk mendukung pengambilan keputusan prediktif berbasis komputasi.

Guna mengatasi persoalan ketidakpastian tersebut, diperlukan penerapan teknik *data mining* atau penambangan data yang mampu mengekstraksi basis data pendaftaran historis menjadi informasi manajerial yang bernilai guna tinggi (Fahreza, 2022). Secara umum, *data mining* memiliki lima fungsi utama, yaitu estimasi, prediksi, klasifikasi, pengelompokan (*clustering*), dan asosiasi (Maruli Sitompul *et al.*, 2021). Dalam konteks permasalahan di MTs NU Islamiyah, pendekatan klasifikasi memegang peranan paling relevan untuk memilah kecenderungan calon pendaftar, apakah mereka akan berstatus melanjutkan pendaftaran (masuk) atau membatalkan pendaftaran (tidak masuk). Pola klasifikasi individual ini selanjutnya dapat diagregasikan menjadi dasar estimasi probabilitas penerimaan siswa baru pada tahun ajaran yang akan datang. Salah satu metode klasifikasi probabilitas yang terbukti andal dalam menangani data bertipe kategorikal dan memiliki keunggulan komputasi yang cepat adalah algoritma Naive Bayes.

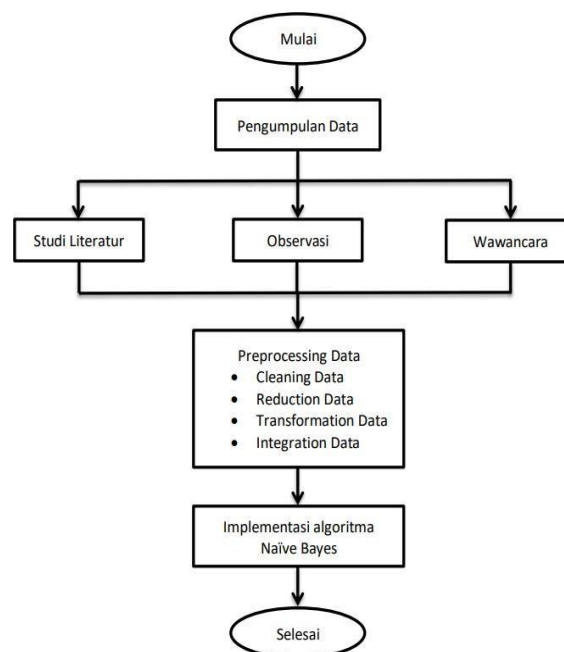
Algoritma Naive Bayes, yang berakar pada Teorema Bayes karya ilmuwan Thomas Bayes, merupakan metode pengklasifikasian statistik yang memprediksi peluang keterjadian suatu kelas di masa depan berdasarkan akumulasi bukti dan probabilitas kondisional dari pengalaman masa lalu (Imandasari *et al.*, 2019). Keunggulan utama metode ini terletak pada kemampuannya menghasilkan estimasi probabilitas yang akurat meskipun hanya menggunakan himpunan data latih yang terbatas, dengan asumsi independensi bersyarat antar-atribut prediktor. Dalam penelitian ini, atribut pendaftaran historis calon siswa seperti jenis kelamin, latar belakang ekonomi keluarga, dan estimasi jarak domisili diekstraksi dan dimodelkan ke dalam sistem klasifikasi berbasis bahasa pemrograman Python. Penerapan Naive Bayes bertujuan untuk mengukur performa pengklasifikasian status penerimaan siswa baru, mengidentifikasi faktor-faktor pendorong utama keputusan siswa untuk masuk ke MTs NU Islamiyah, serta memberikan gambaran proyeksi yang lebih terukur dan objektif. Dengan adanya sistem prediksi yang dibangun, pihak manajemen sekolah diharapkan memiliki

landasan kuantitatif yang kuat dalam menyusun kebijakan operasional, mengoptimalkan perencanaan fasilitas fisik, serta menata alokasi sumber daya sekolah secara terencana menjelang dibukanya tahun ajaran baru.

2. Metode Penelitian

2.1 Kerangka Penelitian

Pelaksanaan penelitian ini dilakukan melalui serangkaian tahapan sistematis yang mencakup pengumpulan data, pengolahan data (*preprocessing*), serta pengujian performa model klasifikasi (Informasi *et al.*, 2022). Penelitian ini menerapkan pendekatan deskriptif yang dirancang untuk menganalisis dan mengolah data pendaftaran siswa baru guna memproyeksikan kecenderungan calon peserta didik di masa mendatang (Jaedun, 2011). Alur kerja penelitian secara menyeluruh disajikan pada kerangka penelitian dalam Gambar 1.



Gambar 1. Kerangka Penelitian

2.2 Pengumpulan Data

Teknik pengumpulan data dilakukan untuk menghimpun seluruh informasi faktual dan data historis yang relevan dengan kebutuhan pemodelan prediksi (Irawan & Damanik, 2022). Teknik pengumpulan data yang diterapkan dalam penelitian ini mencakup: 1) observasi, yaitu peninjauan secara langsung pada objek penelitian di MTs NU Islamiyah (Senika & Iskandar, 2022) guna memperoleh data riwayat pendaftaran siswa baru pada tahun-tahun sebelumnya yang telah terverifikasi validitasnya oleh pihak sekolah; dan 2) wawancara, yaitu komunikasi tatap muka langsung dengan kepala sekolah dan bagian administrasi untuk menggali informasi mendalam mengenai dinamika, faktor kendala, serta pola fluktuasi penerimaan siswa baru setiap tahun ajaran.

2.3 Analisis Data

Tahap analisis data difokuskan pada proses seleksi, pembersihan, dan penyiapan data pendaftaran agar berada dalam format yang siap digunakan dalam pemodelan. Seleksi data dilakukan secara cermat dengan mengidentifikasi serta menangani permasalahan konsistensi format data, nilai

yang hilang (*missing value*), dan data ganda (*redundant data*) (Sinaga *et al.*, 2022). Secara konseptual, analisis data merupakan upaya menata dan menyusun secara sistematis catatan hasil observasi serta wawancara untuk meningkatkan pemahaman peneliti terhadap pola masalah yang dikaji (Muhadjir, 1998, sebagaimana dikutip dalam Rijali, 2019). Dalam konteks komputasi data mining, data yang telah bersih dianalisis menggunakan algoritma Naive Bayes melalui tahapan pelatihan (*training*) dan pengujian (*testing*) model.

2.4 Implementasi Sistem

Tahap implementasi merupakan proses penerjemahan rancangan alur klasifikasi dan formula matematis algoritma Naive Bayes ke dalam lingkungan komputasi menggunakan bahasa pemrograman Python. Pada tahap ini, model dilatih menggunakan data historis yang telah melalui tahap prapemrosesan, dievaluasi menggunakan data pengujian, serta disimpan untuk menguji kemampuan prediksi status pendaftaran calon siswa baru.

3. Hasil dan Pembahasan

3.1 Hasil

Pengumpulan data awal dilaksanakan melalui observasi langsung dan wawancara mendalam di lingkungan MTs NU Islamiyah. Hasil observasi mengidentifikasi ketersediaan catatan riwayat pendaftaran peserta didik dari kurun waktu operasional sebelumnya yang memuat berbagai atribut demografis dan administratif, antara lain usia, jenis kelamin, asal sekolah, penghasilan orang tua, dan alamat tempat tinggal. Sementara itu, wawancara dengan bagian tata usaha mengonfirmasi dua permasalahan pokok: 1) fluktuasi tajam jumlah siswa yang mendaftar ulang setiap tahun ajaran baru sehingga menyulitkan perencanaan sarana dan prasarana; serta 2) penumpukan data arsip pendaftaran dalam format fisik (*hard file*) maupun lembar kerja digital (*soft file*) yang belum pernah dimanfaatkan secara optimal melalui pemodelan komputasional berbasis teknologi informasi. Data mentah pendaftaran yang dihimpun mencakup kurun waktu sepuluh tahun (2014–2023) dengan total 710 baris rekaman. Struktur data awal memuat informasi identitas siswa, variabel sosial-ekonomi, dan status pendaftaran akhir sebagaimana ditunjukkan pada Tabel 1.

Tabel 1. Dataset Riwayat Pendaftaran Siswa Baru

No	Nama	Jenis Kelamin	Usia	Tempat Lahir	Status Keluarga	Tahun	Status
1	Nando Darmawan	L	12	Jakarta	Mampu	2014	Masuk
2	Oktavia Eka	P	11	Tangerang	Mampu	2014	Masuk
3	Putra Firmansyah	L	13	Serang	Tidak Mampu	2014	Tidak Masuk
4	Qoriyah Gita	P	13	Pandeglang	Tidak Mampu	2014	Tidak Masuk
5	Rafi Haris	L	13	Jakarta	Mampu	2014	Masuk
6	Sari Indriani	P	13	Jakarta	Mampu	2014	Masuk
7	Adi Jaya	L	13	Situbondo	Mampu	2014	Tidak Masuk
8	Candra Laksana	L	13	Jakarta	Tidak Mampu	2014	Masuk
9	Dito Mandiri	L	13	Tangerang	Tidak Mampu	2014	Masuk
10	Ella Nanda	P	13	Jakarta	Mampu	2014	Masuk

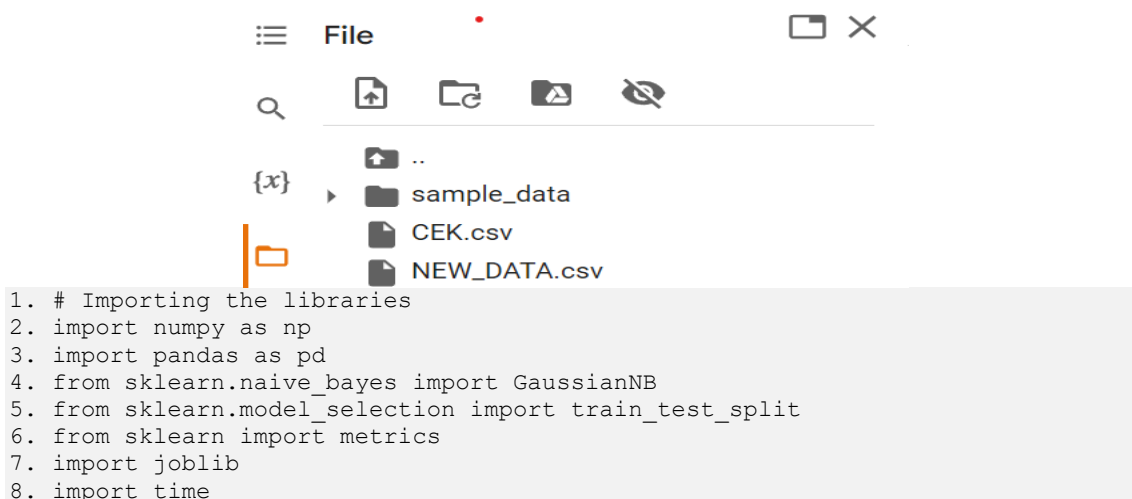
...	...	...	...	...	...	...	...
701	Ananda I.K	L	14	Banten	Mampu	2023	?
702	Aufa Aulia N.A	P	14	Banten	Mampu	2023	?
703	Bagus Dwiyanto	L	14	Serang	Tidak Mampu	2023	?

Tahap prapemrosesan data (*preprocessing*) dilakukan guna menyeleksi atribut yang relevan, membersihkan data yang tidak konsisten atau mengandung derau (*noise*), serta menyelaraskan format data agar dapat diproses oleh algoritma klasifikasi (Alghifari & Juardi, 2021). Atribut alamat tempat tinggal dikonversi menjadi fitur jarak tempuh ke madrasah melalui diskretisasi berbasis wilayah domisili, sebagaimana dirangkum pada Tabel 2.

Tabel 2. Aturan Konversi Atribut Alamat ke Jarak

Alamat / Wilayah	Kategori Jarak	Representasi Biner
Asembagus	< 1 km	1
Jangkar	> 1 km	0

Implementasi pemodelan dilakukan menggunakan pustaka Scikit-Learn pada lingkungan Google Colaboratory. Langkah awal komputasi diawali dengan pengunggahan berkas dataset ke direktori kerja sebagaimana diilustrasikan pada Gambar 2.



Gambar 2. Direktori Berkas Dataset pada Lingkungan Google Colaboratory

Setelah pustaka dasar seperti NumPy, Pandas, Scikit-Learn, dan Joblib diimpor, berkas data pendaftaran dibaca ke dalam struktur *DataFrame* Pandas dan sepuluh baris pertama ditampilkan untuk memastikan integritas pemuatan data sebagaimana diperlihatkan pada Gambar 3.

```

1. # Membaca data dari file excel ke dalam DataFrame
2. nama_file_xlsx = 'Cek.xlsx'
3. data = pd.read_excel(nama_file_xlsx)
4.
5. data.head(10)

```

Gambar 3. Pembacaan Sepuluh Baris Pertama Dataset

Tahapan berikutnya adalah seleksi atribut. Atribut-atribut yang tidak memiliki korelasi langsung terhadap kecenderungan penerimaan calon siswa—yaitu kolom *No*, *Nama*, *Usia*, *Tempat Lahir*, *Tgl*

Lahir, dan *Tahun*—dieliminasi dari *DataFrame*. Eksekusi fungsi penghapusan kolom ini disajikan pada Gambar 4.

```
1. def drop_data():
2.     print("Proses Drop Kolom yang Tidak Diperlukan")
3.     global data
4.     data = data.drop(columns=['No', 'Nama', 'Usia', 'Tempat Lahir', 'Tgl
    Lahir', 'Tahun'])
5.     time.sleep(10)
6.     print("Drop Berhasil")
7.     time.sleep(5)
8.     print("Proses Selesai")
9. drop_data()
```

```
➤ Proses Drop Kolom yang Tidak Diperlukan
Drop Berhasil
Proses Selesai
```

Gambar 4. Proses Eliminasi Kolom yang Tidak Diperlukan

Data yang telah direduksi kemudian dibersihkan melalui transformasi kategorikal ke bentuk numerik biner (Eka *et al.*, 2020). Atribut *Jenis Kelamin* dikodekan menjadi 1 (Laki-laki) dan 0 (Perempuan); atribut *Jarak* dikodekan menjadi 1 (< 1 km) dan 0 (> 1 km); atribut *Status Keluarga* dikodekan menjadi 1 (Mampu) dan 0 (Tidak Mampu); serta label target *Status* dikodekan menjadi 1 (Masuk) dan 0 (Tidak Masuk). Luaran sepuluh baris data hasil pembersihan dan transformasi biner ini ditampilkan pada Gambar 5.

```
# Mengubah tipe data menjadi category
data_c = data.astype('category')

# Pemetaan nilai kategorikal ke representasi biner numerik
data_c["Jenis Kelamin"] = data_c["Jenis Kelamin"].map({'Laki - Laki': 1,
    'Perempuan': 0})
data_c["Jarak"] = data_c["Jarak"].map({'< 1 Km': 1, '> 1 Km': 0})
data_c["STATUS KELUARGA"] = data_c["STATUS KELUARGA"].map({'Mampu': 1, 'Tidak
    Mampu': 0})
data_c["STATUS"] = data_c["STATUS"].map({'Masuk': 1, 'Tidak Masuk': 0})

# Menampilkan subset data hasil transformasi
data_subset = data_c.iloc[1:10]
print(data_subset)
```

```
➤  Jenis Kelamin STATUS KELUARGA Jarak STATUS
1      0      1      1      1
2      0      1      1      0
3      0      0      1      1
4      1      1      1      0
5      0      0      1      0
6      1      1      1      0
7      0      1      1      1
8      1      1      0      1
9      1      0      0      0
```

Gambar 5. Hasil Pembersihan dan Transformasi Data Kategorikal ke Numerik

Dataset yang telah bersih selanjutnya dipisahkan menjadi matriks fitur prediktor *X* (tiga kolom pertama: Jenis Kelamin, Status Keluarga, dan Jarak) serta vektor target *Y* (kolom Status). Struktur pemisahan fitur dan label target ini diperlihatkan pada Gambar 6.

```

1. # Memisahkan data label dan fitur
2. X = data_c.iloc[:, :3].values
3. Y = data_c.iloc[:, 3].values
4. Y

[1, 1, 0, 1, 0, ..., 1, 0, 1, 1, 1]
Length: 710
Categories (2, int64): [1, 0]

```

Gambar 6. Pemisahan Matriks Fitur dan Vektor Label Target

Secara teoretis, klasifikasi dilakukan berdasarkan Teorema Bayes yang menghitung nilai probabilitas kondisional kemunculan suatu kelas target berdasarkan data atribut historis (Senika & Iskandar, 2022; Suyanto, 2019), dengan persamaan:

$$P(H | X) = \frac{P(X|H) \cdot P(H)}{P(X)} \quad (1)$$

Keterangan:

- X : Data sampel dengan kelas yang belum diketahui
- H : Hipotesis bahwa sampel X merupakan bagian dari kelas spesifik tertentu
- $P(H | X)$: Probabilitas posterior dari hipotesis H berdasarkan kondisi data X
- $P(H)$: Probabilitas prior dari hipotesis H
- $P(X | H)$: Probabilitas kondisional dari X berdasarkan kondisi hipotesis H
- $P(X)$: Probabilitas dari kemunculan data atribut X

Dataset selanjutnya dibagi menjadi data latih (*training set*) dan data uji (*testing set*) menggunakan fungsi `train_test_split` dengan rasio pengujian sebesar 20% dan parameter acak `random_state = 710`. Model klasifikasi Gaussian Naive Bayes dilatih menggunakan data latih melalui fungsi `GaussianNB().fit(X_train, y_train)` dan disimpan secara permanen menggunakan format `joblib`. Pengujian model dilakukan dengan memuat kembali model terlatih untuk memprediksi label status pada himpunan data uji baru, yang menghasilkan deretan nilai prediksi biner sebagaimana disajikan pada Gambar 7.

```

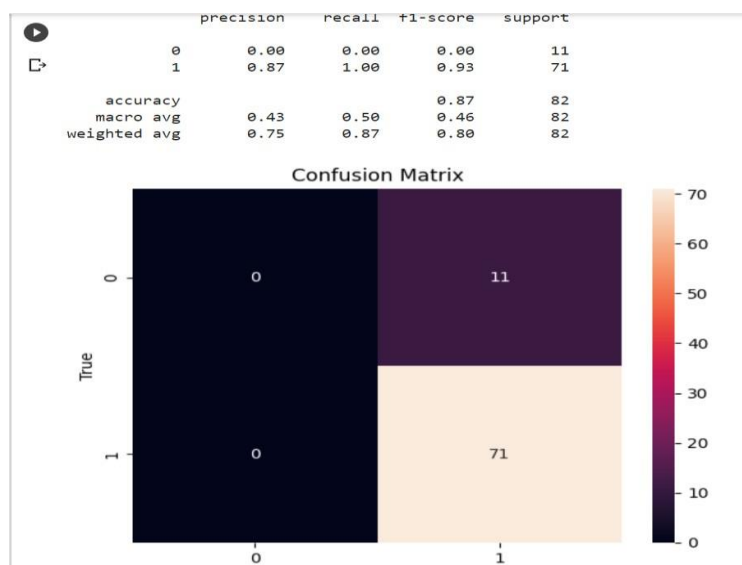
1. print("\nPredicted Status:")
2. print(y_pred)

Predicted Status:
[1 1 1 1 1 1 0 1 1 0 1 1 1 1 1 1 1 0 1 0 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1
 1 1 0 0 0 1 1 1 1 1 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 0]

```

Gambar 7. Luaran Vektor Prediksi Status Pendaftaran

Evaluasi kinerja model klasifikasi diukur menggunakan visualisasi matriks konfusi (*confusion matrix*) berbantuan pustaka Seaborn serta laporan klasifikasi (*classification report*) sebagaimana ditunjukkan pada Gambar 8.



Gambar 8. Visualisasi Matriks Konfusi dan Laporan Klasifikasi

Berdasarkan hasil evaluasi pada 82 sampel data uji (*test set*), model mengklasifikasikan seluruh sampel pengujian ke dalam kelas 1 (Masuk). Pada kelas Masuk (label 1), model mencatatkan nilai *precision* sebesar 87% (0,87), *recall* sebesar 100% (1,00), dan *f1-score* sebesar 93% (0,93) dengan jumlah sampel aktual (*support*) sebanyak 71 data. Sebaliknya, pada kelas Tidak Masuk (label 0), model menghasilkan performa sebesar 0% (0,00) untuk seluruh metrik (*precision*, *recall*, dan *f1-score*) dengan jumlah sampel aktual (*support*) sebanyak 11 data.

3.2 Pembahasan

Hasil evaluasi eksperimen menunjukkan adanya disparitas performa yang sangat ekstrem antara kelas Masuk dan kelas Tidak Masuk. Akurasi agregat yang tampak tinggi (mencapai sekitar 87%) sesungguhnya merupakan bias statistik yang dipicu oleh ketimpangan distribusi data (*class imbalance*) yang mencolok pada himpunan data historis pendaftaran di MTs NU Islamiyah. Dari 82 sampel pengujian, terdapat 71 siswa (86,59%) yang berstatus Masuk dan hanya 11 siswa (13,41%) yang berstatus Tidak Masuk. Karena populasi didominasi secara mutlak oleh calon siswa yang masuk, model Naive Bayes mengalami kecenderungan untuk memprediksi seluruh sampel ke dalam kelas mayoritas guna memaksimalkan fungsi probabilitas prior $P(H)$. Akibatnya, model gagal mengenali pola karakteristik dari calon siswa yang memutuskan untuk tidak melanjutkan pendaftaran (menghasilkan 0 nilai *true negative*).

Keterbatasan performa ini juga dipengaruhi oleh dua faktor teknis utama pada tahapan pemodelan. Pertama, pemilihan varian algoritma GaussianNB pada pustaka Scikit-Learn secara teoretis mengasumsikan bahwa fitur prediktor terdistribusi kontinu normal (Gaussian). Pada penelitian ini, seluruh fitur yang digunakan (*Jenis Kelamin*, *Status Keluarga*, dan *Jarak*) telah diubah menjadi data biner diskrit (0 dan 1). Penggunaan varian Naive Bayes yang dirancang khusus untuk data diskrit atau biner, seperti CategoricalNB atau BernoulliNB, lebih tepat secara konseptual dan berpotensi memberikan estimasi probabilitas kondisional yang lebih presisi pada fitur-fitur biner tersebut.

Reduksi fitur yang dilakukan pada tahap prapemrosesan menyisakan hanya tiga atribut prediktor (*Jenis Kelamin*, *Status Keluarga*, dan *Jarak*). Meskipun ketiga faktor demografis ini berpengaruh terhadap aksesibilitas dan minat calon siswa, keterbatasan jumlah variabel menyebabkan ruang fitur menjadi sangat sempit, sehingga kombinasi nilai fitur antara pendaftar yang masuk dan yang tidak masuk menjadi identik atau saling tumpang tindih (*overlap*). Ketiadaan variabel pembeda tambahan—seperti rata-rata nilai rapor, jalur pendaftaran, riwayat prestasi nonakademik, atau keberadaan saudara

kandung di sekolah—menyulitkan algoritma dalam menemukan batas keputusan (*decision boundary*) yang tegas bagi kelas minoritas.

Secara manajerial, meskipun sistem yang dibangun telah mampu mengotomatisasi alur pemrosesan data pendaftaran berbasis bahasa pemrograman Python, pihak madrasah belum dapat mengandalkan model ini sepenuhnya untuk mendeteksi potensi pembatalan pendaftaran (kelas Tidak Masuk). Model saat ini sangat efektif sebagai alat bantu rekapitulasi probabilitas bagi calon siswa yang berpotensi besar untuk melanjutkan studi, namun perbaikan model pada masa mendatang memerlukan penerapan teknik penyeimbangan data (seperti SMOTE atau pembobotan kelas *class_weight*), pemilihan estimator yang selaras dengan tipe data kategorikal, serta penambahan atribut sosio-akademik yang lebih komprehensif.

4. Kesimpulan dan Saran

Penelitian ini berhasil mengimplementasikan algoritma Naive Bayes untuk memprediksi status penerimaan siswa baru di MTs NU Islamiyah dengan memanfaatkan total 710 data pendaftaran yang terbagi menjadi 628 data latih dan 82 data uji menggunakan empat atribut. Hasil pengujian menunjukkan bahwa keterkaitan antarvariabel serta pemilihan atribut sangat memengaruhi performa model yang dibangun. Evaluasi kinerja model klasifikasi menghasilkan tingkat akurasi keseluruhan sebesar 87%. Kinerja prediksi pada kelas siswa yang masuk menunjukkan performa yang sangat optimal dengan perolehan nilai *precision* sebesar 87%, *recall* sebesar 100%, *f1-score* sebesar 93%, dan nilai *support* sebanyak 71 data. Sebaliknya, prediksi pada kelas siswa yang tidak masuk memperoleh nilai *precision*, *recall*, dan *f1-score* masing-masing sebesar 0% dengan nilai *support* sebanyak 11 data akibat adanya ketimpangan jumlah data antar-kelas.

Pengembangan penelitian selanjutnya disarankan untuk menambah variabel atau atribut baru yang relevan dengan faktor pendaftaran siswa agar dapat meningkatkan ketepatan klasifikasi model. Selain itu, penelitian berikutnya dapat menerapkan teknik penyeimbangan data untuk mengatasi ketimpangan distribusi kelas minoritas, serta menguji metode klasifikasi alternatif seperti *Decision Tree* atau *Random Forest* sebagai studi perbandingan guna menentukan algoritma terbaik yang paling andal bagi pihak sekolah.

5. Daftar Pustaka

- Alghifari, F., & Juardi, D. (2021). Penerapan data mining pada penjualan makanan dan minuman menggunakan metode algoritma Naïve Bayes. *Jurnal Ilmiah Informatika*, 9(2), 75–81. <https://doi.org/10.33884/jif.v9i02.3755>
- Fahreza, A. (2022). Penerapan data mining dengan metode single moving average dalam pengolahan data penerimaan siswa baru. *Seminar Nasional Ilmu Komputer (SNASIKOM)*, 2(1), 25–34.
- Imandasari, T., Irawan, E., Windarto, A. P., & Wanto, A. (2019). Algoritma Naive Bayes dalam klasifikasi lokasi pembangunan sumber air. *Prosiding Seminar Nasional Riset Information Science (SENARIS)*, 1, 750–761.
- Irawan, E., & Damanik, B. E. (2022). Teknik klasifikasi dalam memprediksi penerimaan siswa baru menggunakan metode Naïve Bayes. *Journal of Machine Learning and Artificial Intelligence (JOMLAI)*, 1(3), 209–218. <https://doi.org/10.55123/jomlai.v1i3.963>
- Jaedun, A. (2011). *Metodologi penelitian eksperimen*. Fakultas Teknik Universitas Negeri Yogyakarta.

- Jasri, M., Wijaya, A., & Sunggara, R. (2022). Penerapan data mining untuk klasifikasi penyakit Demam Berdarah Dengue (DBD) dengan metode Naïve Bayes (studi kasus Puskesmas Taman Krocok). *SMARTICS Journal*, 8(1), 35–42. <https://doi.org/10.21067/smartics.v8i1.7375>
- Putri, S. E. Y. (2021). Penerapan model Naive Bayes untuk memprediksi potensi pendaftaran siswa di SMK Taman Siswa Teluk Betung berbasis web. *Journal of Engineering, Computer Science and Information Technology (JECSIT)*, 1(2), 113–120.
- Rizal, S., & Luthfi, M. (2018). Penerapan algoritma Naïve Bayes untuk prediksi penerimaan siswa baru di SMK Al-Amien Wonorejo. *Explore IT: Jurnal Keilmuan dan Aplikasi Teknik Informatika*, 10(1), 14–21. <https://doi.org/10.35891/explorit.v10i1.1671>
- Senika, A., & Iskandar, D. (2022). Implementasi metode Naïve Bayes dalam penilaian kinerja sales marketing pada PT. Pachira Distrinusa. *Jurnal Media Informatika Budidarma*, 6(1), 701–709. <https://doi.org/10.30865/mib.v6i1.3331>
- Sinaga, S., Sembiring, R. W., & Sumarno, S. (2022). Penerapan algoritma Naive Bayes untuk klasifikasi prediksi penerimaan siswa baru. *Journal of Machine Learning and Data Analytics*, 1(1), 55–64.
- Sitompul, A. M., Suhada, S., & Saifulah, S. (2021). Teknik data mining dalam prediksi jumlah siswa baru dengan algoritma Naive Bayes. *Kesatria: Jurnal Penerapan Sistem Informasi (Komputer dan Manajemen)*, 2(2), 108–117.
- Suntoro, J. (2019). *Data mining: Algoritma dan implementasi dengan pemrograman PHP*. Elex Media Komputindo.
- Suyanto, D. (2019). *Data mining untuk klasifikasi dan klusterisasi data* (Edisi ke-1). Informatika Bandung.
- Umar, R., & Supardi, S. (2022). Penerapan metode Naive Bayes dalam mengklasifikasi penerima BLT pada Desa Pelangsian. *Jurnal JUPITER*, 14(2), 64–70.
- Zulfauzi, Z., & Alamsyah, M. N. (2020). Penerapan algoritma Naive Bayes untuk prediksi penerimaan mahasiswa baru studi kasus Universitas Bina Insan Fakultas Komputer. *Jurnal Teknologi Informasi MURA*, 12(2), 156–165. <https://doi.org/10.32767/jti.v12i02.1096>