Jurnal Ilmiah Ilmu Komputer & Elektronika (MURAKOM)
Volume. 2 Nomor. 2 April 2026
e-ISSN : 3123-1985; p-ISSN : 3110-4754, Hal. 25-38
DOI: https://doi.org/10.64365/murakom.v2i2.282
Available online at: *https://jurnal.muaraedukasi.id/index.php/murakom*

Optimasi Naive Bayes dengan Diskritisasi dan Penanganan Outlier untuk Deteksi Diabetes pada Dataset Pima Indians

Gus Rosauli Pandiangan[1], Angelica Barus[2], Mohd. Rafiif Albani[3], Nuriana Sipahutar [4]
[1-4]Universitas Negeri Medan, Indonesia
rosapandiangan71@gmail.com[1], angelicabrs.4243250029@mhs.unimed.ac.id[2], rafiifalbani21@gmail.com[3]
Alamat: Jalan Willem Iskandar, Pasar V, Medan Estate, Kecamatan Percut Sei Tuan, Kabupaten Deli Serdang, Sumatera Utara
Korespondensi penulis: *rosapandiangan71@gmail.com*

Abstract. *Diabetes often remains asymptomatic in its early stages, leading to delayed medical intervention for many individuals. Computational technology, specifically the Gaussian Naive Bayes (GNB) method, offers a solution for rapid risk detection. However, health datasets frequently present challenges such as inconsistent entries, implausible values, and extreme outliers. This study aims to improve data quality by implementing rigorous preprocessing, including the imputation of invalid values, outlier removal using the Interquartile Range (IQR) method, and the discretization (binning) of complex medical variables into more interpretable categories. The results demonstrate a significant performance enhancement, with the model achieving an accuracy of 80.60% in predicting diabetes risk. Blood glucose levels and Body Mass Index (BMI) were identified as the most critical predictors. In conclusion, by prioritizing data cleaning and transformation, even fundamental algorithms can serve as highly effective early screening tools for medical personnel in community health centers or small clinics.*

Keywords: *Diabetes Prediction, Gaussian Naive Bayes, Data Preprocessing, Outlier Removal, Discretization.*

Abstrak. Diabetes sering kali tidak menunjukkan gejala awal sehingga banyak orang terlambat menanganinya. Teknologi komputer sebenarnya bisa membantu mendeteksi risiko ini dengan cepat melalui metode Gaussian Naive Bayes. Namun, tantangannya adalah data kesehatan yang tersedia sering kali berantakan, memiliki angka yang tidak masuk akal, atau data yang terlalu ekstrem. Penelitian ini bertujuan memperbaiki kualitas data tersebut dengan cara membersihkan angka yang salah, membuang data yang tidak wajar, dan menyederhanakan angka medis yang rumit menjadi kelompok yang lebih mudah dibaca sistem. Hasilnya sangat baik karena komputer menjadi jauh lebih akurat dalam menebak risiko diabetes dengan tingkat keberhasilan mencapai 80,60%. Faktor yang paling menentukan dalam prediksi ini adalah kadar gula darah dan berat badan pasien. Kesimpulannya dengan membereskan data yang berantakan terlebih dahulu maka teknologi sederhana pun bisa menjadi alat deteksi dini yang sangat membantu tenaga medis di puskesmas atau klinik kecil.

Kata kunci: Diabetes, Klasifikasi, Gaussian Naive Bayes, Preprocessing, Diskritisasi

1. LATAR BELAKANG

Diabetes Melitus (DM) telah menjadi salah satu tantangan kesehatan global yang paling signifikan pada abad ke-21, dengan angka prevalensi yang terus meningkat secara drastis di berbagai belahan dunia. Penyakit metabolik ini sering kali bersifat asimtomatik pada fase awal, sehingga banyak penderita yang tidak menyadari kondisi mereka hingga terjadi komplikasi serius seperti kerusakan saraf, gagal ginjal, hingga penyakit kardiovaskular. Di tengah meningkatnya beban sistem kesehatan, kebutuhan akan alat skrining dini yang cepat, murah, dan akurat menjadi sangat mendesak. Teknologi pembelajaran mesin (*Machine Learning*) menawarkan solusi potensial melalui pengembangan model prediktif yang mampu mengidentifikasi risiko diabetes berdasarkan data klinis rutin pasien.

Salah satu algoritma klasik yang sering digunakan dalam klasifikasi medis adalah Gaussian Naive Bayes (GNB). Algoritma ini memiliki keunggulan pada kecepatan komputasi

Received: Februari 15, 2026; Revised: Maret 20, 2026; Accepted: April 01, 2026; Online Available: April 02, 2026; Published: April 20, 2026;

yang tinggi, kebutuhan data pelatihan yang efisien, serta hasil prediksi berbasis probabilitas yang mudah diinterpretasikan oleh tenaga medis. Namun, performa GNB sering kali terbatas ketika dihadapkan pada data kesehatan dunia nyata yang "kotor". Dataset medis, seperti *Pima Indians Diabetes Database* (PIDD), sering kali mengandung nilai tidak masuk akal (seperti nilai nol pada tekanan darah atau BMI), distribusi data yang tidak normal, serta keberadaan *outlier* yang ekstrem. Masalah-masalah teknis ini jika tidak ditangani dengan benar akan mengaburkan batas keputusan algoritma dan menurunkan akurasi prediksi secara signifikan.

Berdasarkan fenomena tersebut, penelitian ini merumuskan masalah utama pada bagaimana meningkatkan daya prediksi algoritma Naive Bayes yang sederhana agar mampu menghasilkan performa yang setara dengan algoritma kompleks. Rumusan masalah dalam penelitian ini difokuskan pada dua hal: pertama, sejauh mana teknik *preprocessing* seperti imputasi median dan eliminasi *outlier* dapat menstabilkan distribusi data medis; dan kedua, bagaimana penerapan teknik diskritisasi fitur (*binning*) dapat membantu algoritma GNB dalam menangani variabel kontinu yang memiliki variansi tinggi. Dengan menjawab tantangan ini, diharapkan teknologi sederhana dapat menjadi instrumen deteksi dini yang handal.

Tujuan utama dari penelitian ini adalah untuk merancang dan mengevaluasi alur kerja (*pipeline*) optimasi data yang mampu meningkatkan performa klasifikasi diabetes pada algoritma Naive Bayes. Secara spesifik, penelitian ini bertujuan untuk membuktikan bahwa melalui penanganan *outlier* menggunakan metode *Interquartile Range* (IQR) dan transformasi fitur numerik ke dalam kategori klinis, model dapat mencapai nilai akurasi dan AUC-ROC yang lebih tinggi. Selain itu, penelitian ini bertujuan untuk memberikan analisis mendalam mengenai fitur medis apa saja yang paling berkontribusi terhadap risiko diabetes, sehingga dapat memberikan wawasan tambahan bagi praktisi kesehatan dalam melakukan diagnosis awal.

Manfaat dari penelitian ini diharapkan dapat memberikan kontribusi ganda, baik secara teoretis maupun praktis. Secara teoretis, penelitian ini memperkaya literatur mengenai strategi optimasi data pada algoritma klasifikasi probabilistik. Secara praktis, model yang dihasilkan dapat menjadi referensi bagi pengembangan sistem pendukung keputusan klinis di fasilitas kesehatan primer dengan sumber daya komputasi terbatas, seperti puskesmas atau klinik kecil. Dengan membereskan integritas data di awal, penelitian ini menegaskan bahwa algoritma sederhana sekalipun dapat menjadi alat deteksi dini yang sangat efektif dalam upaya menekan angka komplikasi diabetes di masyarakat.

2. KAJIAN TEORITIS

Jumlah penderita diabetes di seluruh dunia terus bertambah dari tahun ke tahun. Berdasarkan laporan edisi ke-11 International Diabetes Federation (IDF) Diabetes Atlas, lebih dari 500 juta orang dewasa hidup dengan diabetes pada tahun 2024, dan angka ini diperkirakan akan mendekati 900 juta jiwa pada tahun 2050 (Genitsaridi et al., 2026). Tanpa penanganan yang tepat, penyakit ini dapat memicu komplikasi serius seperti gagal ginjal, kerusakan saraf, dan penyakit jantung. Kondisi inilah yang mendorong kebutuhan terhadap sistem deteksi dini yang mampu bekerja cepat, akurat, dan dapat diterapkan pada populasi besar.

Dalam beberapa tahun terakhir, pendekatan berbasis *machine learning* banyak digunakan untuk membangun model prediksi diabetes. Algoritma seperti *Random Forest, Support Vector Machine, Decision Tree*, dan *Naive Bayes* telah diuji dan dibandingkan pada berbagai dataset klinis (Edeh et al., 2022). Di antara algoritma tersebut, *Naive Bayes* memiliki keunggulan dari sisi kecepatan komputasi dan kemudahan interpretasi hasil, dua hal yang sangat dibutuhkan dalam konteks pengambilan keputusan medis (Tasin et al., 2023). Dataset *Pima Indians Diabetes* yang dikumpulkan oleh National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) menjadi salah satu data rujukan yang paling sering dipakai dalam penelitian sejenis (Chang et al., 2023). Dataset ini memuat 768 rekam medis pasien wanita keturunan Pima Indian berusia minimal 21 tahun, dengan delapan atribut klinis yang mencakup kadar glukosa, tekanan darah, indeks massa tubuh (BMI), kadar insulin, dan beberapa variabel lain yang berkaitan dengan risiko diabetes (Ramesh et al., 2021).

Sayangnya, dataset ini tidak bebas dari permasalahan kualitas data. Sejumlah atribut medis seperti kadar glukosa, tekanan darah, dan BMI memiliki nilai nol yang secara klinis tidak mungkin terjadi—nilai-nilai tersebut sesungguhnya adalah data yang tidak tercatat (*missing values*) dan bukan nol yang sesungguhnya (Salih, 2024). Dari total 768 data, sebanyak 374 pasien tercatat memiliki nilai insulin nol dan 227 pasien dengan nilai ketebalan kulit nol. Di luar itu, sebaran data pada beberapa fitur juga mengandung nilai-nilai ekstrem yang cukup jauh dari pola umum, yang bila dibiarkan dapat memengaruhi stabilitas dan akurasi model (Ashisha et al., 2024). Ketidakberesan-ketidakberesan inilah yang menjadikan tahap persiapan data menjadi bagian yang tidak bisa diabaikan sebelum membangun model klasifikasi.

Untuk menangani permasalahan tersebut, penelitian ini menggunakan tiga pendekatan secara berurutan. Pertama, nilai-nilai yang tidak valid diganti menggunakan imputasi median, sebuah teknik yang tidak mudah terpengaruh oleh data ekstrem (Salih, 2024). Kedua, data-data yang terlalu jauh dari rentang wajar diidentifikasi dan dihapus menggunakan metode *Interquartile Range* (IQR), yakni dengan menentukan batas bawah dan atas berdasarkan Q1$dan Q3 dari sebaran data (Rahman et al., 2025). Ketiga, seluruh atribut numerik dikonversi menjadi kategori menggunakan teknik diskritisasi *binning*. Transformasi ini penting karena *Naive Bayes* secara alami lebih cocok bekerja dengan data kategoris; ketika data masih berbentuk kontinu dengan distribusi tidak normal, perkiraan probabilitasnya menjadi kurang andal (Feng et al., 2023).

Sejumlah penelitian sebelumnya telah menerapkan *Naive Bayes* pada dataset *Pima Indians*, namun umumnya hanya menangani satu atau dua permasalahan data sekaligus—misalnya hanya melakukan imputasi tanpa diskritisasi, atau menghapus *outlier* tanpa

mengikutinya dengan transformasi kategoris (Al-Hameli et al., 2023; Edeh et al., 2022). Penelitian ini mencoba mengisi celah tersebut dengan menggabungkan ketiga tahap tersebut secara berurutan, kemudian mengukur dampaknya terhadap akurasi, *recall*, dan nilai AUC-ROC model. Dengan cara ini, diharapkan dapat ditunjukkan bahwa persiapan data yang matang mampu mendongkrak performa *Naive Bayes* secara nyata, sehingga model yang dihasilkan layak untuk dipublikasikan dan berpotensi diterapkan sebagai alat bantu skrining diabetes berbasis data.

3. METODE PENELITIAN

Penelitian ini menerapkan pendekatan eksperimental kuantitatif yang berfokus pada optimasi alur kerja *data mining* untuk meningkatkan performa klasifikasi algoritma Gaussian Naive Bayes (GNB). Mengingat algoritma Naive Bayes sangat bergantung pada asumsi distribusi data yang ideal, penelitian ini menempatkan teknik *preprocessing* data sebagai instrumen utama dalam memitigasi kelemahan data mentah yang sering kali mengandung derau (*noise*) dan ketidakkonsistenan.Tahapan penelitian ini dilakukan secara sistematis, termasuk pengumpulan data, preprocessing, pemodelan, dan evaluasi model.

3.1 Dataset

Data yang digunakan dalam penelitian ini adalah dataset *Pima Indians Diabetes Database* (PIDD), yang diperoleh melalui platform repositori data Kaggle. Dataset ini secara historis dikumpulkan oleh *National Institute of Diabetes and Digestive and Kidney Diseases* dan telah menjadi standar baku (*benchmark*) dalam berbagai penelitian klasifikasi medis di seluruh dunia. Pemilihan dataset ini didasarkan pada relevansi klinis dari atribut-atribut yang terkandung di dalamnya, yang mencakup parameter fisiologis krusial seperti frekuensi kehamilan, konsentrasi glukosa plasma, tekanan darah diastolik, ketebalan lipatan kulit trisep, kadar insulin serum, Indeks Massa Tubuh (BMI), fungsi silsilah diabetes (*Diabetes Pedigree Function*), serta usia pasien. Penggunaan dataset ini sangat penting karena menyediakan gambaran riil mengenai faktor risiko diabetes tipe 2 pada kelompok etnis tertentu, sehingga memungkinkan algoritma pembelajaran mesin untuk mengekstraksi pola-pola probabilitas yang kompleks dari variabel-variabel medis yang saling berkaitan.

3.2 Preprocessing Data

a. Data Cleaning

Pada tahap ini, dilakukan identifikasi dan penanganan terhadap data yang tidak valid atau kotor yang dapat menurunkan performa prediktif model. Fokus utama adalah pada deteksi nilai nol (0) pada metrik medis kritis seperti kadar glukosa, tekanan darah, ketebalan kulit, kadar insulin, dan BMI, yang secara biologis tidak mungkin terjadi pada pasien yang masih hidup. Nilai-nilai nol tersebut dikategorikan sebagai *missing values* terselubung yang jika dibiarkan akan memberikan bias negatif pada perhitungan rata-rata dan varians dalam algoritma Naive Bayes.

Untuk mengatasi masalah ini, nilai-nilai tersebut diganti menggunakan teknik imputasi median. Pemilihan median sebagai nilai pengganti dilakukan karena sifatnya yang jauh lebih stabil (*robust*) terhadap keberadaan *outlier* dibandingkan dengan nilai rata-rata (*mean*). Dalam distribusi data medis yang sering kali menceng (*skewed*), penggunaan median memastikan bahwa nilai imputasi tetap berada pada pusat distribusi massa data tanpa terpengaruh oleh nilai ekstrem. Dalam literatur data medis, teknik

imputasi yang tepat terbukti secara signifikan mampu meningkatkan kualitas dataset, mengurangi *noise*, dan memberikan fondasi yang lebih kokoh bagi model klasifikasi untuk mencapai tingkat akurasi dan generalisasi yang lebih tinggi.

b. Penanganan Outlier

Metode Interquartile Range (IQR), yang dihitung sebagai selisih antara kuartil ketiga dan kuartil pertama, seperti yang ditunjukkan pada Persamaan (1), digunakan untuk mengidentifikasi outlier.

$$IQR = Q3 - Q1$$

Persamaan (2) dan (3) digunakan untuk menentukan batas bawah dan atas untuk mendeteksi outlier, yaitu:

$$Q1 - 1.5 \times IQR$$
$$Q3 + 1.5 \times IQR$$

Untuk meningkatkan kualitas data dan kestabilan model, data yang di luar rentang tersebut dihapus dari dataset. Ini dianggap sebagai outlier.

c. Diskritisasi Data

Diskritisasi dilakukan dengan menggunakan metode binning untuk mengubah data numerik menjadi format kategorikal. Tujuan dari proses ini adalah untuk meningkatkan kinerja algoritma berbasis probabilistik seperti Naive Bayes dan juga menyederhanakan distribusi data.

Pola distribusi data menjadi lebih terstruktur karena nilai kontinu dikelompokkan ke dalam interval yang lebih kecil. Ini mempermudah proses pembelajaran model

3.3 Pemodelan Naive Bayes

Dalam penelitian ini, algoritma Naive Bayes digunakan karena sangat efisien dalam komputasi dan dapat memberikan hasil klasifikasi yang baik pada dataset medis. Metode ini berasal dari Teorema Bayes, yang dinyatakan dalam Persamaan (4).

$$P(H \mid X) = \frac{P(X \mid H)P(H)}{P(X)}$$

Pada persamaan tersebut, Hmenyatakan hipotesis kelas dan Xmerupakan data fitur yang diamati. Nilai $P(H \mid X)$merupakan probabilitas posterior, yaitu probabilitas suatu kelas berdasarkan data yang diberikan. Sementara itu, $P(X \mid H)$adalah likelihood, $P(H)$merupakan probabilitas prior, dan $P(X)$adalah evidence.

Model dibangun dengan data yang telah melalui tahap preprocessing. Selanjutnya, model dilatih dengan data pelatihan untuk mengetahui pola probabilitas dari masing-masing fitur terhadap kelas target.

3.4 Pengujian Model

Dataset dibagi menjadi data pelatihan dan data pengujian dengan rasio 80 : 20. Untuk memastikan bahwa model memiliki cukup data untuk proses pelatihan, pembagian ini bertujuan untuk memastikan bahwa model dapat diuji secara objektif pada data yang belum pernah digunakan sebelumnya.

3.5 Evaluasi Model

Sebagaimana ditunjukkan pada Persamaan (5), evaluasi performa model dilakukan dengan menggunakan metrik seperti akurasi, recall, dan AUC-ROC. Akurasi digunakan untuk mengukur tingkat keseluruhan ketepatan prediksi model.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Pada persamaan tersebut, *True Positive* (TP) menunjukkan jumlah data positif yang terklasifikasi dengan benar, *True Negative* (TN) menunjukkan jumlah data negatif yang terklasifikasi dengan benar, sedangkan *False Positive* (FP) dan *False Negative* (FN) masing-masing menunjukkan kesalahan klasifikasi.

Kemampuan model untuk menemukan kasus positif juga diukur dengan metrik recall. Di sisi lain, AUC-ROC digunakan untuk mengukur kemampuan model untuk membedakan secara keseluruhan antara kelas negatif dan positif.

4. HASIL DAN PEMBAHASAN

Bagian ini memaparkan hasil implementasi eksperimen serta analisis mendalam terhadap performa model Gaussian Naive Bayes (GNB) yang telah dioptimalkan. Evaluasi dimulai dengan menyajikan proses transformasi data mentah melalui tahapan *preprocessing* yang ketat, dilanjutkan dengan pengujian model menggunakan berbagai metrik standar evaluasi medis. Fokus utama dari pembahasan ini adalah mengungkap bagaimana integrasi teknik penanganan *outlier* dan diskritisasi fitur (binning) memberikan kontribusi nyata dalam memperkuat kemampuan prediksi model. Melalui penyajian data yang komprehensif mulai dari matriks kontingensi, kurva karakteristik operasi (ROC), hingga analisis kepentingan fitur bab ini akan membuktikan efektivitas skema yang diusulkan. Hasil ini tidak hanya menunjukkan peningkatan akurasi pada prediksi diabetes, tetapi juga memberikan perspektif baru dalam literatur penelitian terkait pengolahan data medis.

4.1 Hasil Eksplorasi Data dan Preprocessing

Tahap eksplorasi data mengungkap sejumlah temuan penting mengenai kualitas PIDD. Dari 768 sampel awal, ditemukan total 763 nilai nol yang tersebar di lima fitur medis kritis: Glucose (5 nilai, 0,65%), BloodPressure (35 nilai, 4,56%), SkinThickness (227 nilai, 29,56%), Insulin (374 nilai, 48,70%), dan BMI (11 nilai, 1,43%). Fitur Insulin menunjukkan proporsi missing value tertinggi — hampir separuh dari seluruh nilainya tidak valid. Kondisi ini mencerminkan tantangan nyata dalam pengumpulan data klinis di lapangan, di mana beberapa pemeriksaan mungkin tidak dilakukan karena keterbatasan fasilitas atau biaya.

Setelah imputasi median diterapkan, dataset tetap berjumlah 768 sampel namun kini bebas dari nilai yang secara medis tidak valid. Nilai median yang digunakan untuk imputasi adalah: Glucose = 117,0 mg/dL, BloodPressure = 72,0 mmHg, SkinThickness = 29,0 mm, Insulin = 125,0 μU/mL, dan BMI = 32,3 kg/m². Nilai-nilai ini konsisten dengan statistik populasi yang dilaporkan pada studi epidemiologi suku Pima Indian [11]. Penerapan IQR kemudian menghapus 436 sampel outlier (56,77%), menyisakan 332 sampel yang bersih dan terdistribusi lebih simetris. Meskipun pengurangan sampel ini cukup substansial, hasilnya adalah dataset yang lebih representatif dan bebas dari distorsi ekstrem yang dapat menyesatkan model.

Hasil diskritisasi menunjukkan distribusi yang secara klinis masuk akal untuk populasi ini. Untuk BMI, sebagian besar sampel berada pada kategori Obese (194 sampel, 58,43%), diikuti Overweight (85 sampel, 25,60%), Normal (50 sampel, 15,06%), dan Underweight (3 sampel, 0,90%). Distribusi ini konsisten dengan dokumentasi tingginya prevalensi obesitas pada suku Pima Indian, yang merupakan salah satu faktor risiko utama diabetes tipe 2 di komunitas tersebut [11]. Untuk usia, kelompok Young Adult mendominasi (142 sampel, 42,77%), diikuti Adult (118 sampel, 35,54%), Middle Aged (51 sampel, 15,36%), dan Senior (21 sampel, 6,33%).

4.2 Analisis Confusion Matrix

Confusion matrix model GNB yang telah dioptimalkan pada 67 sampel data uji ditampilkan pada Tabel 3 berikut. Matriks ini merupakan fondasi interpretasi klinis dari performa model, karena setiap sel memiliki konsekuensi yang berbeda dalam konteks skrining medis.

Tabel 1. Confusion Matrix Model GNB yang Telah Dioptimalkan (n = 67)

	Prediksi: Tidak Diabetes (0)	**Prediksi: Diabetes (1)**
Aktual: Tidak Diabetes (0)	TN = 39	FP = 5
Aktual: Diabetes (1)	FN = 8	TP = 15

Dari Tabel 1 dapat dibaca bahwa model menghasilkan 39 True Negative (TN) — pasien tidak diabetes yang diprediksi dengan benar — dan 15 True Positive (TP) — pasien diabetes yang berhasil teridentifikasi. Terdapat 5 False Positive (FP), yaitu pasien sehat yang diprediksi sebagai diabetes. Dalam konteks skrining, FP hanya akan membuat pasien menjalani pemeriksaan lanjutan yang lebih mendalam, sehingga risikonya relatif rendah secara klinis. Yang paling perlu mendapat perhatian adalah 8 kasus False Negative (FN) — pasien diabetes yang lolos dari deteksi. FN memiliki implikasi klinis paling serius karena pasien yang tidak terdeteksi tidak akan mendapatkan intervensi dini, sehingga penyakitnya berisiko berkembang ke tahap yang lebih parah.

4.3 Analisis Metrik Performa Klasifikasi

Tabel 2 menyajikan seluruh metrik evaluasi model secara komprehensif beserta interpretasinya dalam konteks aplikasi skrining diabetes.

Tabel 2. Ringkasan Metrik Performa Model GNB yang Dioptimalkan

Metrik	**Nilai**	**Interpretasi**	**Kategori**
Akurasi (Accuracy)	80,60%	Persentase total prediksi yang benar dari seluruh data uji	Sangat Baik
Presisi (Precision)	75,00%	Dari seluruh prediksi positif, 75% benar-benar pasien diabetes	Dapat Diterima
Recall (Sensitivity)	65,22%	Model mendeteksi 65,22% dari seluruh kasus diabetes yang ada	Cukup Baik

F1-Score	69,77%	Rata-rata harmonik Presisi dan Recall, ukuran performa seimbang	Cukup Baik
Spesifisitas	88,64%	Model benar mengidentifikasi 88,64% pasien yang tidak menderita diabetes	Sangat Baik
AUC-ROC	87,55%	Kemampuan diskriminasi model di semua ambang batas klasifikasi	Excellent ✓
Average Precision (AP)	80,63%	Luas area di bawah kurva Precision-Recall, ringkasan performa di kelas imbalance	Sangat Baik

Akurasi keseluruhan sebesar 80,60% menunjukkan bahwa dari setiap 10 pasien yang diuji, rata-rata 8 di antaranya diklasifikasikan dengan benar. Angka ini merupakan peningkatan yang bermakna dibandingkan dengan penelitian sejenis yang menggunakan GNB tanpa pipeline preprocessing serupa, yang umumnya melaporkan akurasi di kisaran 75–77% [5]. Presisi 75,00% berarti bahwa tiga dari empat pasien yang diprediksi positif memang benar-benar menderita diabetes — sebuah tingkat keandalan yang cukup baik untuk menghindari kecemasan berlebihan pada pasien yang sesungguhnya sehat.

Recall (Sensitivitas) sebesar 65,22% menunjukkan bahwa model berhasil mendeteksi sekitar dua dari tiga kasus diabetes yang sesungguhnya ada. Nilai ini perlu dicermati lebih lanjut, karena dalam konteks skrining penyakit kronis, nilai Recall yang tinggi umumnya lebih diprioritaskan daripada Presisi — lebih baik mendeteksi lebih banyak kasus meski ada beberapa yang ternyata negatif (FP), daripada melewatkan kasus positif (FN). Meski demikian, Spesifisitas yang tinggi (88,64%) membuktikan bahwa model sangat andal dalam mengidentifikasi pasien yang tidak diabetes, sehingga tidak membebani sistem layanan kesehatan dengan rujukan yang tidak perlu. F1-Score sebesar 69,77% sebagai rata-rata harmonik Presisi dan Recall merupakan ukuran performa yang seimbang, dan nilainya yang mendekati 70% menunjukkan bahwa model tidak terlalu condong ke salah satu dimensi tersebut.

4.4 Analisis Kurva ROC dan Nilai AUC

Nilai AUC-ROC sebesar 0,8755 (87,55%) merupakan pencapaian penting yang membuktikan kemampuan diskriminasi model secara global. Kurva ROC memvisualisasikan trade-off antara True Positive Rate (Sensitivitas) dan False Positive Rate (1 − Spesifisitas) di seluruh kemungkinan ambang batas keputusan, sehingga memberikan gambaran performa model yang lebih komprehensif dibandingkan akurasi tunggal. Berdasarkan skema klasifikasi Hosmer dan Lemeshow, nilai AUC dalam rentang [0,80–0,90] dikategorikan sebagai diskriminasi "excellent" [12]. Artinya, jika model diberikan satu pasien diabetes dan satu pasien sehat secara acak, terdapat probabilitas 87,55% bahwa model akan memberikan skor risiko yang lebih tinggi pada pasien diabetes.

Implikasi klinis dari AUC 0,8755 sangat signifikan: model ini memiliki kemampuan diskriminasi yang cukup kuat untuk dijadikan alat skrining lini pertama di fasilitas kesehatan primer, seperti puskesmas atau klinik layanan dasar, di mana tenaga medis spesialis dan peralatan diagnostik canggih mungkin tidak tersedia. Titik optimal pada kurva ROC (titik Youden's J) yang menunjukkan

keseimbangan terbaik antara Sensitivitas dan Spesifisitas dapat digunakan sebagai ambang batas keputusan yang direkomendasikan dalam implementasi praktis.

Average Precision (AP) sebesar 80,63% pada kurva Precision-Recall juga memperkuat validitas model. Kurva Precision-Recall sangat informatif pada kondisi ketidakseimbangan kelas, karena berfokus pada performa kelas positif (diabetes) tanpa terpengaruh oleh dominansi kelas negatif. Nilai AP 80,63% yang jauh di atas baseline no-skill (≈34,3% berdasarkan prevalensi kelas positif pada data uji) mengonfirmasi bahwa model memberikan nilai prediktif yang substansial melampaui prediksi acak.

4.5 Analisis Permutation Feature Importance

Tabel 3 menyajikan hasil analisis Permutation Feature Importance yang mengukur seberapa besar penurunan akurasi model ketika nilai setiap fitur diacak secara acak. Fitur yang penurunan akurasinya paling besar dianggap paling penting bagi model.

Tabel 3. Permutation Feature Importance Model GNB

No.	Fitur	Importance	Interpretasi Klinis
1	Glucose	+0,1179	Paling dominan. Kadar glukosa plasma adalah kriteria diagnosis utama diabetes berdasarkan ADA.
2	BMI_Cat_Enc	+0,0269	Hasil diskritisasi BMI. Membuktikan binning meningkatkan kontribusi fitur terhadap model.
3	SkinThickness	+0,0194	Ketebalan lipatan kulit, indikator cadangan lemak tubuh yang berkorelasi dengan resistansi insulin.
4	BloodPressure	+0,0164	Tekanan darah tinggi merupakan komorbiditas umum pada pasien diabetes tipe 2.
5	Pregnancies	+0,0070	Riwayat kehamilan berulang meningkatkan risiko diabetes gestasional dan tipe 2.
6	Insulin	0,0000	Kontribusi nol kemungkinan akibat tingginya tingkat imputasi (48,70% nilai awal adalah 0).
7	Age_Cat_Enc	−0,0204	Kontribusi negatif; usia kategorikal saja tidak cukup diskriminatif setelah variabel lain dikontrol.

8	DiabetesPedigreeFunction	−0,0224	Skor genetik menunjukkan korelasi lemah setelah preprocessing; mungkin perlu transformasi lanjutan.

Glucose terbukti menjadi fitur paling dominan dengan nilai importance sebesar 0,1179, jauh melampaui fitur-fitur lainnya. Temuan ini sepenuhnya konsisten dengan pengetahuan medis yang mapan: kadar glukosa plasma merupakan parameter diagnostik utama diabetes berdasarkan standar American Diabetes Association (ADA) — ambang batas glukosa puasa ≥126 mg/dL atau glukosa 2 jam ≥200 mg/dL adalah kriteria diagnosis diabetes [13]. Dominansi fitur ini mengonfirmasi bahwa model telah berhasil menangkap sinyal klinis yang paling relevan dari data.

Temuan menarik kedua adalah posisi BMI_Cat_Enc sebagai fitur terpenting kedua (importance = 0,0269). Ini merupakan bukti langsung bahwa teknik diskritisasi yang diterapkan pada fitur BMI berhasil meningkatkan kualitas informasi yang dapat ditangkap oleh model. Secara mekanistik, pembagian BMI ke dalam empat kategori WHO mengurangi variansi dalam masing-masing kategori, sehingga batas antar kelas menjadi lebih tajam dan mudah dipisahkan oleh model. Hal ini juga menjelaskan mengapa dalam persamaan likelihood Gaussian (Persamaan 2), nilai σ^2_{ki} yang lebih kecil akan menghasilkan distribusi likelihood yang lebih runcing dan diskriminatif, yang pada akhirnya meningkatkan kemampuan model membedakan pasien diabetes dari yang tidak.

Sebaliknya, fitur Insulin menunjukkan nilai importance nol, yang mungkin disebabkan oleh tingginya proporsi nilai yang diimputasi (48,70% dari nilai awal adalah nol). Imputasi dengan satu nilai median yang seragam dapat mengurangi variabilitas alami fitur tersebut, sehingga kontribusinya terhadap diskriminasi kelas melemah. Fitur Age_Cat_Enc dan DiabetesPedigreeFunction menunjukkan kontribusi negatif, yang mengindikasikan bahwa pengacakan kedua fitur ini justru sedikit meningkatkan akurasi — fenomena yang dapat terjadi ketika fitur mengandung noise yang mengganggu prediksi model.

4.6 Perbandingan dengan Penelitian Terdahulu

Untuk memposisikan kontribusi penelitian ini dalam konteks literatur yang lebih luas, Tabel 6 membandingkan performa model yang diusulkan dengan beberapa penelitian terdahulu yang menggunakan PIDD dengan pendekatan berbasis Naive Bayes.

Tabel 4. Perbandingan Performa dengan Penelitian Terdahulu pada PIDD

Penelitian	Metode	Akurasi	AUC-ROC
Aini et al. (2021)	Naive Bayes (tanpa preprocessing lanjutan)	75,46%	Tidak dilaporkan
Sisodia & Sisodia (2018)	Naive Bayes + seleksi fitur sederhana	76,30%	~0,78
Maniruzzaman et al. (2017)	LDA + Naive Bayes	77,60%	0,82

Penelitian Ini (2025)	GNB + Median + IQR + Diskritisasi	80,60%	0,8755 ✓

Dari Tabel 4 terlihat bahwa pendekatan yang diusulkan dalam penelitian ini menghasilkan performa AUC-ROC tertinggi (0,8755) di antara penelitian sejenis yang menggunakan Naive Bayes pada PIDD. Peningkatan ini tidak dicapai melalui penggantian algoritma dengan model yang lebih kompleks, melainkan semata-mata melalui optimasi kualitas data pada tahap preprocessing. Hal ini membuktikan tesis utama penelitian: bahwa investasi pada tahap pembersihan dan transformasi data memberikan return yang signifikan terhadap performa bahkan algoritma yang sederhana sekalipun.

5. KESIMPULAN DAN SARAN

Penelitian ini berhasil membuktikan bahwa kualitas data (*data quality*) memegang peranan yang lebih krusial dibandingkan kompleksitas algoritma dalam meningkatkan akurasi prediksi diabetes. Melalui penerapan *pipeline preprocessing* yang sistematis meliputi imputasi median untuk menangani data kosong, penghapusan *outlier* dengan metode IQR, serta teknik diskritisasi fitur algoritma Gaussian Naive Bayes yang sederhana mampu mencapai tingkat akurasi sebesar 80,60% dengan nilai AUC-ROC mencapai 0,8755. Hasil ini mengonfirmasi bahwa penanganan fitur medis yang tidak masuk akal dan transformasi variabel kontinu menjadi kategori (binning) secara efektif mempertajam batas keputusan model. Dengan demikian, model yang diusulkan memiliki kemampuan diskriminasi yang "Excellent" dan sangat layak untuk diimplementasikan sebagai instrumen skrining dini di fasilitas kesehatan dengan sumber daya komputasi terbatas.

Meskipun model ini menunjukkan performa yang sangat baik, masih terdapat ruang pengembangan untuk penelitian selanjutnya guna memperkuat validitas hasil. Peneliti menyarankan penggunaan teknik penyeimbangan data (*data balancing*) seperti SMOTE untuk mengatasi masalah ketidakseimbangan kelas antara pasien diabetes dan non-diabetes guna meningkatkan nilai *Recall*. Selain itu, eksplorasi terhadap metode diskritisasi yang lebih dinamis atau penggunaan algoritma *ensemble* dapat dipertimbangkan untuk melihat sejauh mana performa model dapat ditingkatkan tanpa mengorbankan aspek interpretabilitas medis. Terakhir, pengujian model pada dataset klinis lokal yang lebih bervariasi sangat direkomendasikan agar sistem prediksi ini memiliki generalisasi yang lebih luas dalam mendukung pengambilan keputusan klinis di lapangan.

DAFTAR REFERENSI

Al-Hameli, B. A., Alsewari, A. A., Basurra, S. S., Bhogal, J., & Ali, M. A. H. (2023). Diabetes disease prediction system using HNB classifier based on discretization method. *Journal of Integrative Bioinformatics*, 20(1). https://doi.org/10.1515/jib-2021-0037

Ashisha, G. R., Mary, X. A., Kanaga, E. G. M., Andrew, J., & Eunice, R. J. (2024). Random Oversampling-Based Diabetes Classification via Machine Learning Algorithms. *International Journal of Computational Intelligence Systems*, 17(1), 270. https://doi.org/10.1007/s44196-024-00678-3

Chang, V., Bailey, J., Xu, Q. A., & Sun, Z. (2023). Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms. *Neural Computing and Applications*, 35(22), 16157–16173. https://doi.org/10.1007/s00521-022-07049-z

Damanik, A. R., Sumijan, S., & Nurcahyo, G. W. (2021). Prediksi Tingkat Kepuasan dalam Pembelajaran Daring Menggunakan Algoritma Naïve Bayes. *Jurnal Sistim Informasi dan Teknologi*, 3(3), 88–94. https://doi.org/10.37034/jsisfotek.v3i3.49

Depari, A. D. S., Kirana, C. C., Oktariana, C. N., & Akbar, F. (2025). Prediksi Risiko Diabetes Dengan Metode Naive Bayes: Identifikasi Faktor Risiko Utama dan Evaluasi Akurasi Model. *JATI (Jurnal Mahasiswa Teknik Informatika)*, 9(4).

Edeh, M. O., et al. (2022). A Classification Algorithm-Based Hybrid Diabetes Prediction Model. *Frontiers in Public Health*, 10. https://doi.org/10.3389/fpubh.2022.829519

Fajriati, N., & Prasetiyo, B. (2023). Optimasi Algoritma Naive Bayes dengan Diskritisasi K-Means pada Diagnosis Penyakit Jantung. *Jurnal Teknologi Informasi dan Ilmu Komputer*, 10(3), 503–512. https://doi.org/10.25126/jtiik.2023106510

Feng, X., Cai, Y., & Xin, R. (2023). Optimizing diabetes classification with a machine learning-based framework. *BMC Bioinformatics*, 24(1), 428. https://doi.org/10.1186/s12859-023-05467-x

Genitsaridi, I., et al. (2024). 11th edition of the IDF Diabetes Atlas: global, regional, and national diabetes prevalence estimates for 2024. *The Lancet Diabetes & Endocrinology*, 12(2), 149–156. https://doi.org/10.1016/S2213-8587(23)00299-2

Han, J., Kamber, M., & Pei, J. (2022). *Data Mining: Concepts and Techniques* (4th ed.). Morgan Kaufmann.

Hidayat, A., & Surarso, B. (2022). Penerapan Algoritma Naive Bayes untuk Prediksi Diagnosa Penyakit Diabetes Mellitus. *Jurnal Masyarakat Informatika*, 13(1), 1-10. https://doi.org/10.14710/jmi.v13i1.44281

Kurniawati, R., & Kristiyanti, D. A. (2021). Analisis Pengaruh Outlier pada Algoritma Klasifikasi untuk Prediksi Penyakit Diabetes. *Jurnal Informatika*, 8(2), 175-182. https://doi.org/10.31294/ji.v8i2.10543

Larose, D. T., & Larose, C. D. (2020). *Data Science Methodology: A Step-by-Step Guide to Project Success*. Wiley.

Nasari, F., & Darma, S. (2021). *Penerapan Algoritma Klasifikasi Data Mining*. Yayasan Kita Menulis.

Pratama, A. R., & Widiastiwi, Y. (2020). Implementasi Algoritma Naive Bayes Dalam Mengklasifikasi Penyakit Diabetes Mellitus. *Jurnal Informatika dan Multimedia*, 12(2), 55-63.

Rahman, F., Hossain, S., Tiang, J.-J., & Nahid, A.-A. (2025). Diabetes Prediction Using Feature Selection Algorithms and Boosting-Based Machine Learning Classifiers. *Diagnostics*, 15(20), 2622. https://doi.org/10.3390/diagnostics15202622

Ramesh, J., Aburukba, R., & Sagahyroon, A. (2021). A remote healthcare monitoring framework for diabetes prediction using machine learning. *Healthcare Technology Letters*, 8(3), 45–57. https://doi.org/10.1049/htl2.12010

Ridwan, A. (2020). Penerapan Algoritma Naïve Bayes Untuk Klasifikasi Penyakit Diabetes Mellitus. *Jurnal Sistem Komputer dan Kecerdasan Buatan*, 4(1), 15–21. https://doi.org/10.47970/siskom-kb.v4i1.156

Salih, M. S. (2024). Diabetic Prediction based on Machine Learning Using PIMA Indian Dataset. *Communications on Applied Nonlinear Analysis*, 31(5s), 138–156. https://doi.org/10.52783/cana.v31.1008

Sasmita, S., & Wati, M. (2022). Perbandingan Teknik Diskritisasi Data pada Algoritma Naive Bayes untuk Klasifikasi Penyakit. *Jurnal Ilmu Komputer dan Informatika*, 6(1), 44-52.

Suyanto. (2023). *Machine Learning Tingkat Lanjut: Teori dan Implementasi*. Informatika.

Tasin, I., Nabil, T. U., Islam, S., & Khan, R. (2023). Diabetes prediction using machine learning and explainable AI techniques. *Healthcare Technology Letters*, 10(1–2), 1–10. https://doi.org/10.1049/htl2.12039

Wahyuni, S., & Fadlil, A. (2023). Optimasi Naive Bayes Menggunakan Normalisasi Data untuk Prediksi Diabetes Pima Indian. *Jurnal Ilmiah Teknik Informatika*, 17(2), 120-130.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2021). *Data Mining: Practical Machine Learning Tools and Techniques* (5th ed.). Morgan Kaufmann.

Zahra, S. A., & Muslim, M. A. (2021). Analisis Performa Naive Bayes dengan Seleksi Fitur untuk Deteksi Diabetes. *Jurnal Edukasi dan Penelitian Informatika*, 7(1), 12-18. https://doi.org/10.26418/jp.v7i1.44231