Journal Sensi:Strategic Of Education in Information System
Accreditation rank 5
SK Dirjen DIKTI, RisTek No. 10/C/C3/DT.05.00/2025; E-ISSN: 2655-5298 P-ISSN: 2461-1409

PENERAPAN ALGORITMA *GAUSSIAN NAIVE BAYES* UNTUK KLASIFIKASI RISIKO DIABETES MELITUS BERDASARKAN DATA SIMULASI REKAM MEDIS PASIEN

Ilamsyah[1], Aditya Dwi Nurcahyo[2], Mutiara Anisa[3]
[1,2,3]Program Studi Ilmu Komputer Universitas Yatsi Madani Tangerang
email: [1]ilamsyah@uym.ac.id, [2]adityadwinurcahyo07@gmail.com,
[3]cmutiaraanss477@gmail.com

Abstrak

Diabetes melitus merupakan penyakit kronis dengan jumlah kasus yang cenderung meningkat dari waktu ke waktu dan dapat menyebabkan berbagai komplikasi apabila tidak dikenali serta ditangani secara lebih awal. Dalam hal ini, penerapan teknik *data mining* dapat digunakan untuk mengidentifikasi dan mengklasifikasikan tingkat risiko diabetes berdasarkan karakteristik yang dimiliki pasien, sehingga dapat membantu proses pengambilan keputusan secara lebih efektif. Penelitian ini menerapkan algoritma *Gaussian Naive Bayes* dalam melakukan klasifikasi risiko diabetes melitus dengan memanfaatkan 200 data simulasi yang menggambarkan karakteristik rekam medis pasien. Data yang digunakan mencakup beberapa atribut, yaitu umur, jenis kelamin, berat badan, tinggi badan, *Body Mass Index (BMI)*, tekanan darah, kadar glukosa, kadar kolesterol, aktivitas fisik, riwayat diabetes dalam keluarga, kebiasaan merokok, serta status diabetes yang digunakan sebagai variabel target. Proses penelitian dilakukan melalui beberapa tahapan, dimulai dari *Exploratory Data Analysis (EDA)* untuk memahami karakteristik data, dilanjutkan dengan *preprocessing*, pengubahan data kategorikal melalui metode *Label Encoding*, serta pembagian dataset menjadi data pelatihan dan data pengujian dengan perbandingan 80:20. Selanjutnya, model klasifikasi dibangun menggunakan algoritma *Gaussian Naive Bayes* dan kinerjanya dianalisis berdasarkan beberapa metrik evaluasi, meliputi *Accuracy, Precision, Recall, F1-Score, Confusion Matrix*, dan *Classification Report*. Berdasarkan hasil pengujian, model menghasilkan nilai *Accuracy* sebesar 97,50%, *Precision* 94,74%, *Recall* 100%, dan *F1-Score* 97,30%. Nilai tersebut memperlihatkan bahwa *Gaussian Naive Bayes* mampu memberikan hasil klasifikasi yang sangat baik pada dataset simulasi yang digunakan. Dengan demikian, algoritma tersebut memiliki potensi untuk digunakan sebagai salah satu pendekatan pendukung dalam proses identifikasi dan klasifikasi risiko diabetes melitus.

Kata Kunci: Data Mining, Diabetes Melitus, Gaussian Naïve Bayes, Klasifikasi, Rekam Medis Simulasi.

Abstract

Diabetes mellitus is a chronic disease with a number of cases that tends to increase over time and can lead to various complications if it is not recognized and treated at an early stage. In this context, the application of data mining techniques can be used to identify and classify the risk level of diabetes based on patients' characteristics, thereby supporting a more effective decision-making process. This study applies the Gaussian Naive Bayes algorithm to classify the risk of diabetes mellitus using 200 simulated data records representing the characteristics of patients' medical records. The data consist of several attributes, including age, gender, body weight, height, Body Mass Index (BMI), blood pressure, glucose level, cholesterol level, physical activity, family history of diabetes, smoking habits, and diabetes status as the target variable. The research process consists of several stages, beginning with Exploratory Data Analysis (EDA) to understand the characteristics of the dataset, followed by data preprocessing, conversion of categorical data using the Label Encoding method, and division of the dataset into training and testing sets using an 80:20 ratio. Subsequently, a classification model was developed using the Gaussian Naive Bayes algorithm, and its performance was evaluated using several evaluation metrics, including Accuracy, Precision, Recall, F1-Score, Confusion Matrix, and Classification Report. Based on the testing results, the model achieved an Accuracy of 97.50%, Precision of 94.74%, Recall of 100%, and F1-Score of 97.30%. These results indicate that the Gaussian Naive Bayes algorithm is capable of providing highly accurate classification results on the simulated dataset used in this study. Therefore, the algorithm has the potential to serve as a supporting approach for the identification and classification of diabetes mellitus risk.

Keywords: *Data Mining, Diabetes Mellitus, Gaussian Naive Bayes, Classification, Medical Records, Simulation.*

I. PENDAHULUAN

Diabetes melitus merupakan penyakit kronis yang hingga saat ini masih menjadi salah satu permasalahan penting dalam bidang kesehatan, seiring dengan terus bertambahnya jumlah kasus dari tahun ke tahun. Penyakit ini terjadi ketika tubuh mengalami gangguan dalam menghasilkan insulin atau tidak mampu menggunakan insulin secara optimal, sehingga menyebabkan peningkatan kadar glukosa di dalam darah. Apabila kondisi tersebut tidak dikenali dan memperoleh penanganan yang tepat sejak awal, diabetes dapat berkembang dan menimbulkan berbagai dampak kesehatan yang serius, termasuk penyakit kardiovaskular, gangguan fungsi ginjal, kerusakan sistem saraf, hingga masalah pada penglihatan. Kondisi tersebut menunjukkan pentingnya penerapan deteksi dini sebagai langkah untuk mengenali kemungkinan risiko diabetes secara lebih cepat, sehingga tindakan penanganan dapat dilakukan sedini mungkin dan risiko munculnya komplikasi dapat dikurangi. [1].

Kemajuan teknologi informasi telah membuka peluang yang semakin luas dalam penerapan *data mining* pada sektor kesehatan, khususnya untuk menganalisis data dalam jumlah besar dan menghasilkan informasi yang dapat digunakan sebagai bahan pertimbangan dalam pengambilan keputusan. Salah satu pendekatan yang dapat digunakan adalah metode klasifikasi, yang bertujuan menentukan suatu data ke dalam kelas tertentu dengan mempertimbangkan pola dan karakteristik yang terdapat pada data tersebut. Salah satu algoritma yang dapat diterapkan untuk tujuan tersebut adalah *Gaussian Naive Bayes*. Algoritma ini memiliki mekanisme perhitungan yang relatif sederhana dan proses pembelajaran yang cepat, serta dapat digunakan untuk mengolah atribut berbentuk numerik dengan mengasumsikan bahwa data mengikuti distribusi Gaussian. Kemampuan tersebut membuat *Gaussian Naive Bayes* menjadi salah satu metode yang relevan untuk digunakan dalam membangun model klasifikasi pada bidang kesehatan, termasuk dalam mengidentifikasi status atau risiko diabetes melitus. [2]

Hasil penelitian sebelumnya memperlihatkan bahwa metode *Naive Bayes* memiliki kemampuan yang cukup baik untuk digunakan dalam melakukan klasifikasi penyakit diabetes. Salah satu studi menemukan bahwa penerapan algoritma tersebut pada data pasien rumah sakit mampu menghasilkan tingkat akurasi sebesar 92,31% dalam menentukan klasifikasi diabetes [1]. Studi lainnya melakukan analisis perbandingan antara *Naive Bayes* dan *K-Nearest Neighbor (KNN)* dengan memanfaatkan dataset yang bersumber dari Kaggle untuk mengevaluasi metode yang memberikan kinerja klasifikasi lebih optimal [3]. Temuan dari penelitian lain juga mengindikasikan bahwa tingkat kinerja *Naive Bayes* dapat mengalami perubahan berdasarkan komposisi data yang digunakan untuk proses pelatihan dan pengujian. Hal ini menunjukkan bahwa penentuan proporsi data latih dan data uji merupakan aspek yang perlu diperhatikan karena dapat memengaruhi hasil pembentukan serta evaluasi model klasifikasi. [4].

Kajian terkait penerapan algoritma *Naive Bayes* dan *Gaussian Naive Bayes* untuk klasifikasi diabetes telah dilakukan pada beragam jenis dan karakteristik dataset. Sejumlah hasil penelitian memperlihatkan bahwa kedua pendekatan tersebut dapat memberikan kinerja klasifikasi yang kompetitif, sehingga berpotensi digunakan untuk membantu proses pengenalan risiko diabetes [5][6][7]. Temuan tersebut menunjukkan bahwa keberhasilan suatu model klasifikasi tidak semata-mata ditentukan oleh pemilihan algoritma, tetapi juga berkaitan dengan karakteristik data yang digunakan serta tahapan persiapan dan pengolahan data sebelum model dibangun. Pada penelitian-penelitian terdahulu, data yang digunakan umumnya berasal dari sumber yang telah tersedia, seperti dataset publik dari *Kaggle*, *National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK)*, maupun kumpulan data rekam medis. Berbeda dengan pendekatan tersebut, penelitian ini menggunakan sebanyak 200 data simulasi yang dirancang untuk menggambarkan karakteristik rekam medis pasien. Data tersebut mencakup sejumlah variabel prediktor, yaitu umur, jenis kelamin, berat badan, tinggi badan, *Body Mass Index (BMI)*, tekanan darah, kadar glukosa, kadar kolesterol, aktivitas fisik, riwayat diabetes dalam keluarga, dan status merokok, sedangkan status diabetes digunakan sebagai variabel target. Pemanfaatan data simulasi ini ditujukan untuk menguji kemampuan *Gaussian Naive Bayes* dalam melakukan klasifikasi berdasarkan karakteristik data yang dirancang sesuai dengan kebutuhan penelitian, sehingga hasil yang diperoleh diposisikan sebagai evaluasi terhadap model dan tidak dimaksudkan untuk menggantikan pengujian menggunakan data rekam medis pasien yang sebenarnya.

Mengacu pada permasalahan dan hasil kajian penelitian sebelumnya, penelitian ini diarahkan untuk menguji kemampuan algoritma *Gaussian Naive Bayes* dalam menentukan klasifikasi risiko diabetes melitus dengan memanfaatkan 200 data simulasi yang dirancang menyerupai karakteristik rekam medis pasien. Pelaksanaan penelitian dilakukan melalui beberapa tahapan, diawali dengan *Exploratory Data Analysis (EDA)* untuk memahami kondisi dan karakteristik dataset, kemudian dilanjutkan dengan proses *preprocessing* serta pengubahan atribut kategorikal ke dalam bentuk numerik menggunakan metode *Label Encoding*. Dataset yang telah dipersiapkan selanjutnya dibagi menjadi data pelatihan sebesar 80% dan data pengujian sebesar 20% untuk membangun serta menguji model *Gaussian Naive Bayes*. Kinerja model kemudian dianalisis menggunakan *Accuracy, Precision, Recall, F1-Score, Confusion Matrix*, dan *Classification Report* sebagai indikator evaluasi. Melalui rangkaian proses tersebut, penelitian ini diharapkan dapat memberikan informasi mengenai tingkat kemampuan *Gaussian Naive Bayes* dalam melakukan klasifikasi risiko diabetes berdasarkan dataset simulasi, sekaligus memberikan dasar atau bahan pertimbangan bagi penelitian berikutnya yang menggunakan dataset medis dengan jumlah, karakteristik, dan variasi data yang lebih luas.

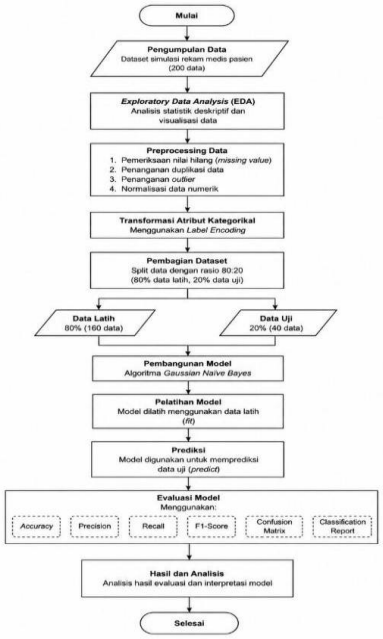
II. METODE PENELITIAN

2.1. Desain Penelitian

Penelitian ini menerapkan pendekatan kuantitatif dengan menggunakan metode klasifikasi dalam kerangka *data mining*. Fokus penelitian diarahkan pada penerapan algoritma *Gaussian Naive Bayes* untuk menentukan klasifikasi risiko diabetes melitus berdasarkan 200 data simulasi yang dirancang menyerupai karakteristik rekam medis pasien. Pelaksanaan penelitian dilakukan melalui serangkaian tahapan yang terstruktur, mulai dari penyusunan dan pengumpulan dataset, analisis karakteristik data menggunakan *Exploratory Data Analysis (EDA)*, tahap *preprocessing*, pembagian data menjadi kelompok pelatihan dan pengujian, pembentukan model klasifikasi, hingga pengukuran kinerja model yang dihasilkan.

Exploratory Data Analysis (EDA) merupakan tahapan analisis yang digunakan untuk memahami kondisi awal suatu dataset dengan mengeksplorasi pola, karakteristik, distribusi, serta kemungkinan adanya ketidaksesuaian pada data melalui statistik deskriptif dan visualisasi [8]. Dalam penelitian ini, EDA dilakukan dengan meninjau struktur dataset, tipe setiap atribut, jumlah data, karakteristik statistik, serta distribusi pada variabel target. Setelah karakteristik data diketahui, proses dilanjutkan dengan *preprocessing* yang mencakup pemeriksaan terhadap keberadaan *missing value* dan data duplikat, kemudian dilanjutkan dengan pengubahan atribut kategorikal ke bentuk numerik menggunakan metode *Label Encoding*. Tahapan tersebut dilakukan agar data memiliki format yang sesuai untuk diproses oleh algoritma *Gaussian Naive Bayes*. Dataset yang telah melalui proses persiapan selanjutnya dipisahkan menggunakan metode *train-test split* dengan komposisi 80% sebagai data latih dan 20% sebagai data uji. Data latih digunakan dalam proses pembentukan model, sedangkan data uji berfungsi untuk mengukur kemampuan model ketika menghadapi data yang tidak digunakan pada tahap pembelajaran. Selanjutnya, performa model dianalisis menggunakan beberapa metrik, yaitu *Accuracy, Precision, Recall, F1-Score, Confusion Matrix*, dan *Classification Report*. Penggunaan berbagai indikator tersebut bertujuan memperoleh gambaran yang lebih menyeluruh mengenai kemampuan *Gaussian Naive Bayes* dalam melakukan klasifikasi risiko diabetes melitus. Alur lengkap dari setiap tahapan penelitian tersebut disajikan pada Gambar 1.

.



Gambar 1. Alur Penelitian

2.2. Dataset Penelitian

Data yang digunakan pada penelitian ini berupa 200 record simulasi yang dirancang untuk menggambarkan kondisi dan karakteristik rekam medis pasien terkait risiko diabetes melitus. Penyusunan dataset dilakukan dengan mempertimbangkan sejumlah faktor yang relevan dan umum digunakan sebagai indikator dalam proses identifikasi risiko diabetes. Seluruh data disimpan dalam berkas berformat *Microsoft Excel (.xlsx)*, kemudian diproses dan dianalisis menggunakan bahasa pemrograman Python melalui lingkungan *Google Colaboratory (Google Colab)*. Dataset memuat kombinasi atribut numerik dan kategorikal yang berfungsi sebagai variabel masukan (*feature*), sementara atribut Diabetes ditetapkan sebagai variabel keluaran (*label*) yang menjadi dasar proses klasifikasi. Sebelum model dibangun, data terlebih dahulu dianalisis melalui tahap *Exploratory Data Analysis (EDA)* untuk memperoleh pemahaman mengenai struktur, distribusi, dan karakteristik setiap atribut. Setelah itu, dilakukan tahapan *preprocessing* untuk mempersiapkan dataset agar berada dalam format yang sesuai sehingga dapat digunakan pada proses pelatihan maupun pengujian algoritma *Gaussian Naive Bayes*.

2.2.1. Sumber Dataset

Data penelitian ini tidak berasal dari rekam medis pasien secara langsung, melainkan berupa dataset simulasi yang dirancang secara khusus oleh peneliti untuk menggambarkan karakteristik pasien yang berkaitan dengan risiko diabetes melitus. Dataset tersebut mencakup 200 data individu dengan variasi pada sejumlah karakteristik yang relevan terhadap kondisi diabetes. Dalam proses penyusunannya, peneliti menggunakan atribut yang lazim digunakan pada penelitian klasifikasi diabetes sebagai dasar pembentukan data, sehingga dataset dapat digunakan untuk merepresentasikan kondisi pasien secara terstruktur dalam tahap pengujian model. Pendekatan ini sekaligus membedakan penelitian dari studi sebelumnya yang umumnya memanfaatkan dataset publik atau data rekam medis dari fasilitas kesehatan. Dengan menggunakan data simulasi yang disusun secara mandiri, penelitian ini secara khusus diarahkan untuk menguji kemampuan dan performa algoritma *Gaussian Naive Bayes* dalam melakukan klasifikasi risiko diabetes melitus.

2.2.2. Atribut Dataset

Dataset penelitian terdiri atas 13 atribut, yang meliputi satu atribut identitas, sebelas atribut prediktor (*feature*), dan satu atribut target (*label*). Deskripsi masing-masing atribut disajikan pada Tabel 1.

Tabel 1. Atribut Dataset Penelitian

No	Atribut	Tipe Data	Keterangan
1	ID Pasien	String	Identitas unik setiap pasien
2	Umur	Integer	Usia pasien (tahun)
3	Jenis Kelamin	Kategorikal	Laki-laki atau Perempuan
4	Berat Badan	Numerik	Berat badan (kg)
5	Tinggi Badan	Numerik	Tinggi badan (cm)
6	BMI	Numerik	*Body Mass Index*
7	Tekanan Darah	Numerik	Tekanan darah pasien (mmHg)
8	Glukosa	Numerik	Kadar glukosa darah (mg/dL)
9	Kolesterol	Numerik	Kadar kolesterol darah (mg/dL)
10	Aktivitas Fisik	Kategorikal	Rendah, Sedang, atau Tinggi

11	Riwayat Diabetes Keluarga	Kategorikal	Ya atau Tidak
12	Status Merokok	Kategorikal	Ya atau Tidak
13	**Diabetes**	**Kategorikal**	Positif atau Negatif (variabel target)

Atribut ID Pasien hanya digunakan sebagai penanda unik untuk membedakan setiap record dalam dataset sehingga tidak digunakan sebagai masukan pada proses pembentukan model. Sementara itu, proses klasifikasi memanfaatkan 11 atribut lainnya sebagai variabel prediktor yang menjadi dasar dalam menghasilkan keputusan klasifikasi. Adapun atribut Diabetes ditetapkan sebagai variabel target atau *label* yang menunjukkan kelas akhir setiap pasien, yaitu termasuk kategori diabetes atau tidak diabetes.

2.3. Tahapan Penelitian

Pelaksanaan penelitian disusun melalui rangkaian proses yang terstruktur, dimulai dari tahap penyiapan dataset, pengolahan dan analisis data, hingga pengujian serta evaluasi terhadap model klasifikasi yang dihasilkan. Seluruh tahapan pemrosesan data dilakukan dengan memanfaatkan bahasa pemrograman Python dalam lingkungan Google Colaboratory (Google Colab). Uraian mengenai setiap tahapan penelitian selanjutnya disajikan secara lebih rinci pada masing-masing subbab berikut.

2.3.1. Pengumpulan Dataset

Pada tahap awal, penelitian diawali dengan menyiapkan dataset yang akan digunakan sebagai sumber data dalam proses klasifikasi. Data yang digunakan berjumlah 200 record simulasi yang dirancang untuk menggambarkan informasi rekam medis pasien dan tersedia dalam bentuk berkas Microsoft Excel (.xlsx). Berkas tersebut kemudian dimasukkan ke dalam lingkungan Google Colaboratory (Google Colab) dan dibaca menggunakan pustaka Pandas pada Python. Proses ini dilakukan agar data dapat diakses, dikelola, serta dipersiapkan secara lebih mudah untuk tahapan analisis dan pengolahan berikutnya.

2.3.2. Exploratory Data Analysis (EDA)

Sebelum model klasifikasi dibangun, dilakukan tahap Exploratory Data Analysis (EDA) dengan tujuan memperoleh gambaran awal mengenai kondisi dan karakteristik dataset yang digunakan. Pemeriksaan terhadap struktur data dilakukan menggunakan fungsi info(), sedangkan fungsi shape digunakan untuk mengetahui dimensi dataset berdasarkan jumlah baris dan atribut yang tersedia. Selanjutnya, karakteristik statistik dari data dianalisis melalui fungsi describe() untuk memperoleh informasi deskriptif pada atribut numerik. Selain analisis tersebut, distribusi pada variabel target juga divisualisasikan untuk melihat jumlah dan proporsi data yang termasuk ke dalam masing-masing kategori status diabetes. Hasil EDA ini digunakan sebagai dasar untuk memahami kondisi dataset sebelum memasuki tahap preprocessing dan pembangunan model klasifikasi.

2.3.3. Preprocessing Data

Tahap preprocessing dilakukan untuk memastikan seluruh data berada dalam kondisi dan format yang sesuai sebelum digunakan oleh algoritma Gaussian Naive Bayes. Pada proses ini, dataset terlebih dahulu diperiksa untuk mengidentifikasi kemungkinan adanya missing value maupun record yang memiliki data duplikat. Setelah kondisi data dipastikan, atribut yang memiliki tipe kategorikal dikonversi ke bentuk numerik melalui teknik Label Encoding sehingga dapat diterima dan diproses oleh algoritma klasifikasi. Hasil dari rangkaian proses tersebut berupa dataset yang telah dipersiapkan dan selanjutnya dapat digunakan sebagai input pada tahap pembagian data, pelatihan, serta pengujian model.

2.3.4. Pembagian Dataset

Setelah seluruh tahapan preprocessing selesai, dataset dipisahkan menjadi dua kelompok untuk kebutuhan pembentukan dan pengujian model. Proses pemisahan dilakukan dengan memanfaatkan fungsi train_test_split() yang tersedia pada pustaka Scikit-learn, dengan komposisi 80% data digunakan sebagai training data dan 20% sisanya sebagai testing data. Berdasarkan total 200 record, pembagian tersebut menghasilkan 160 data untuk proses pelatihan dan 40 data untuk proses pengujian. Data pelatihan selanjutnya digunakan sebagai dasar pembelajaran algoritma Gaussian Naive Bayes dalam membentuk model klasifikasi, sedangkan data pengujian digunakan untuk mengetahui sejauh mana model mampu memberikan klasifikasi yang tepat terhadap data yang tidak terlibat dalam proses pelatihan.

2.3.5. Pembangunan Model Gaussian Naïve Bayes

Setelah data latih tersedia, tahap berikutnya adalah membentuk model klasifikasi dengan menerapkan algoritma *Gaussian Naive Bayes* yang disediakan oleh pustaka *Scikit-learn*. Proses pembelajaran model dilakukan dengan menggunakan seluruh data yang telah ditetapkan sebagai data latih pada tahap pembagian dataset. Melalui proses tersebut, model mempelajari pola hubungan antara atribut prediktor dengan kelas target yang telah ditentukan. Model yang telah selesai dilatih kemudian diterapkan pada data uji untuk menghasilkan prediksi kelas bagi data yang sebelumnya tidak digunakan dalam proses pembelajaran. Hasil prediksi tersebut selanjutnya menjadi dasar dalam melakukan pengukuran dan analisis kinerja model pada tahap evaluasi.

2.3.6. Evaluasi Model

Tahap evaluasi merupakan proses yang dilakukan setelah model klasifikasi selesai dibentuk untuk mengetahui sejauh mana model mampu bekerja pada data yang tidak digunakan selama proses pembelajaran. Pengukuran kinerja dapat dilakukan dengan menguji hasil prediksi terhadap data uji serta menerapkan sejumlah metrik evaluasi yang relevan [9]. Dalam penelitian ini, evaluasi ditujukan untuk mengetahui kemampuan *Gaussian Naive Bayes* dalam menentukan kelas risiko diabetes berdasarkan data pengujian. Tingkat kinerja model dianalisis melalui nilai *Accuracy, Precision, Recall*, dan *F1-Score*. Untuk memperoleh gambaran yang lebih rinci mengenai kesesuaian antara kelas aktual dan hasil prediksi, evaluasi juga dilengkapi dengan *Confusion Matrix* dan *Classification Report*. Kombinasi indikator tersebut digunakan untuk memberikan penilaian yang lebih menyeluruh terhadap kemampuan model dalam melakukan klasifikasi pada dataset yang digunakan.

2.4. Algoritma Gaussian Naïve Bayes

Gaussian Naive Bayes merupakan varian dari algoritma Naive Bayes yang ditujukan untuk data dengan atribut numerik kontinu dan termasuk metode supervised learning yang dikenal memiliki proses pembelajaran sederhana serta waktu komputasi relatif cepat [10]. Algoritma ini menggunakan asumsi bahwa setiap atribut numerik mengikuti distribusi normal (Gaussian distribution), sehingga pada penelitian ini diterapkan untuk mengklasifikasikan risiko diabetes melitus berdasarkan karakteristik pada dataset simulasi rekam medis pasien; secara matematis, proses penentuan probabilitas kelas tersebut didasarkan pada Teorema Bayes yang dinyatakan sebagai berikut:

$$P(y \mid X) = \frac{P(X \mid y) \times P(y)}{P(X)}$$

dengan keterangan:

- $(y \mid X)$= probabilitas data X termasuk ke dalam kelas y (*posterior probability*);
- $(X \mid y)$= probabilitas kemunculan atribut X apabila diketahui berada pada kelas y (*likelihood*);
- P(y)= probabilitas awal (*prior probability*) dari kelas y;
- P(X)= probabilitas kemunculan data X (*evidence*).

Hasil klasifikasi ditentukan dengan memilih kelas yang memiliki nilai probabilitas posterior paling tinggi dibandingkan kelas lainnya. Dalam penelitian ini, penerapan algoritma tersebut dilakukan menggunakan fungsi GaussianNB() dari pustaka *Scikit-learn* dengan bahasa pemrograman Python untuk membangun model klasifikasi

III. HASIL DAN PEMBAHASAN

3.1. Hasil Penelitian

3.1.1. Exploratory Data Analysis (EDA)

Analisis Exploratory Data Analysis (EDA) dilakukan sebagai tahap awal untuk memperoleh pemahaman mengenai kondisi dan struktur dataset sebelum model klasifikasi dibangun. Proses ini mencakup pemeriksaan informasi dataset menggunakan fungsi info(), identifikasi dimensi data melalui shape, analisis statistik deskriptif dengan describe(), serta visualisasi distribusi variabel target untuk melihat komposisi masing-masing kelas.

a. Informasi Dataset (*df.info()*)

Pemeriksaan struktur dataset dilakukan menggunakan fungsi *df.info()*. Hasilnya menunjukkan bahwa dataset terdiri atas 200 data dengan 13 atribut, yang meliputi atribut identitas pasien, atribut prediktor, dan atribut target. Seluruh atribut berhasil dibaca oleh sistem dengan tipe data yang sesuai, baik berupa numerik maupun kategorikal. Selain itu, tidak ditemukan atribut yang memiliki nilai kosong sehingga seluruh data dapat digunakan pada tahap pengolahan berikutnya

```
# ============================================================
# INFORMASI DATASET
# ============================================================

print("="*60)
print("INFORMASI DATASET")
print("="*60)

df.info()
```

```
============================================================
INFORMASI DATASET
============================================================
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200 entries, 0 to 199
Data columns (total 13 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   ID Pasien                  200 non-null    object
 1   Umur                       199 non-null    float64
 2   Jenis Kelamin              199 non-null    object
 3   Berat Badan (kg)           200 non-null    int64
 4   Tinggi Badan (cm)          199 non-null    float64
 5   BMI                        199 non-null    float64
 6   Tekanan Darah              199 non-null    object
 7   Glukosa (mg/dL)            198 non-null    float64
 8   Kolesterol (mg/dL)         199 non-null    float64
 9   Aktivitas Fisik            199 non-null    object
 10  Riwayat Diabetes Keluarga  199 non-null    object
 11  Status Merokok             200 non-null    object
 12  Diabetes                   200 non-null    object
dtypes: float64(5), int64(1), object(7)
memory usage: 20.4+ KB
```

Gambar 2. Hasil Pemeriksaan Informasi Dataset Menggunakan df.info()

b. Ukuran Dataset (*df.shape*)

Ukuran dataset diperiksa menggunakan fungsi *df.shape* untuk mengetahui jumlah baris dan kolom pada dataset. Berdasarkan hasil yang diperoleh, dataset memiliki ukuran (200, 13) yang menunjukkan bahwa terdapat 200 data dengan 13 atribut.

```
# ============================================================
# UKURAN DATASET
# ============================================================

print("="*60)
print("UKURAN DATASET")
print("="*60)

print("Jumlah Baris :", df.shape[0])
print("Jumlah Kolom :", df.shape[1])
```

```
============================================================
UKURAN DATASET
============================================================
Jumlah Baris : 200
Jumlah Kolom : 13
```

Gambar 3. Hasil Pemeriksaan Ukuran Dataset (*df.shape*)

c. Statistik Deskriptif (*df.describe()*)

Statistik deskriptif diperoleh menggunakan fungsi *df.describe()* untuk memberikan informasi mengenai karakteristik atribut numerik pada dataset. Informasi yang ditampilkan meliputi jumlah data (*count*), nilai rata-rata (*mean*), simpangan baku (*standard deviation*), nilai minimum (*minimum*), kuartil (25%, 50%, dan 75%), serta nilai maksimum (*maximum*). Hasil statistik deskriptif digunakan untuk memberikan gambaran awal mengenai penyebaran data sebelum dilakukan proses klasifikasi.

```
# ============================================================
# STATISTIK DESKRIPTIF
# ============================================================

print("="*60)
print("STATISTIK DESKRIPTIF")
print("="*60)

df.describe()
```

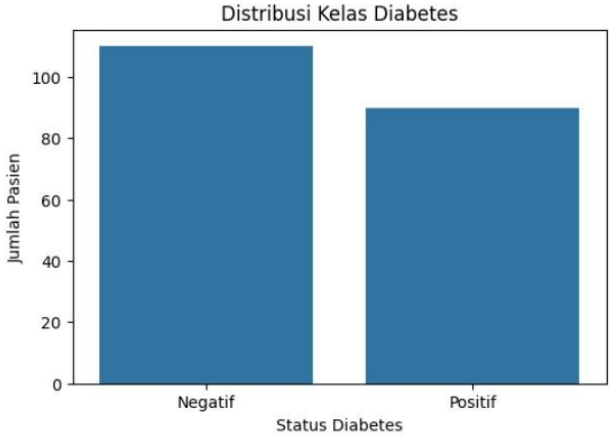
```
============================================================
STATISTIK DESKRIPTIF
============================================================
```

	Umur	Berat Badan (kg)	Tinggi Badan (cm)	BMI	Glukosa (mg/dL)	Kolesterol (mg/dL)
count	199.000000	200.000000	199.000000	199.000000	198.000000	199.000000
mean	46.472362	74.920000	164.045226	28.188945	151.121212	228.934673
std	16.825545	14.729166	9.745393	6.678813	56.513737	49.335081
min	18.000000	45.000000	148.000000	14.200000	58.000000	150.000000
25%	34.000000	66.000000	155.000000	23.550000	103.000000	195.000000
50%	45.000000	76.000000	165.000000	28.400000	130.000000	225.000000
75%	56.000000	85.000000	172.000000	32.100000	199.500000	264.500000
max	80.000000	105.000000	182.000000	47.500000	305.000000	405.000000

Gambar 4. Hasil Statistik Deskriptif Dataset (*df.describe()*)

d. Distribusi Kelas Target

Distribusi kelas target dilakukan untuk mengetahui jumlah data pada masing-masing kelas, yaitu Diabetes dan Tidak Diabetes. Visualisasi distribusi kelas memberikan gambaran mengenai proporsi data yang akan digunakan dalam proses pelatihan dan pengujian model. Informasi ini penting untuk mengetahui apakah distribusi data relatif seimbang sehingga model dapat mempelajari karakteristik setiap kelas dengan baik.



Gambar 5. Distribusi Kelas Target

3.1.2. Hasil *Preprocessing*

Sebelum memasuki tahap pelatihan, dataset terlebih dahulu dipersiapkan melalui proses *preprocessing* agar sesuai dengan kebutuhan algoritma *Gaussian Naive Bayes*. Tahapan ini mencakup identifikasi nilai yang hilang (*missing value*), pengecekan terhadap kemungkinan adanya data ganda, serta pengubahan variabel kategorikal menjadi bentuk numerik menggunakan metode *Label Encoding*. Hasil pengolahan tersebut digunakan sebagai dataset akhir yang siap diproses pada tahap pemodelan..

a. Missing Value

Keberadaan *missing value* pada dataset diperiksa menggunakan fungsi isnull().sum() untuk mengidentifikasi jumlah nilai kosong pada masing-masing atribut. Hasil pemeriksaan menunjukkan bahwa nilai yang hilang terdapat pada atribut Umur, Jenis Kelamin, Tinggi Badan, BMI, Tekanan Darah, Glukosa, Kolesterol, Aktivitas Fisik, dan Riwayat Diabetes Keluarga, sedangkan ID Pasien, Berat Badan, Status Merokok, serta Diabetes tidak memiliki nilai kosong. Kondisi tersebut ditampilkan pada Gambar 6, kemudian nilai yang hilang ditangani melalui proses imputasi dengan mempertimbangkan tipe data setiap atribut agar seluruh record dapat digunakan dalam pembentukan model. Kode yang digunakan untuk proses tersebut disajikan pada Lampiran A. Setelah imputasi selesai dilakukan, dataset dilanjutkan ke tahap pemeriksaan data duplikat sebagai bagian dari proses *preprocessing*.

```
# ============================================================
# PEMERIKSAAN MISSING VALUE
# ============================================================

print("=" * 60)
print("PEMERIKSAAN MISSING VALUE")
print("=" * 60)

missing = df.isnull().sum()

print(missing)
```

```
============================================================
PEMERIKSAAN MISSING VALUE
============================================================
ID Pasien                   0
Umur                        1
Jenis Kelamin               1
Berat Badan (kg)            0
Tinggi Badan (cm)           1
BMI                         1
Tekanan Darah               1
Glukosa (mg/dL)             2
Kolesterol (mg/dL)          1
Aktivitas Fisik             1
Riwayat Diabetes Keluarga   1
Status Merokok              0
Diabetes                    0
dtype: int64
```

Gambar 6. Hasil Pemeriksaan *Missing Value*

b. Data Duplikat

Selanjutnya dilakukan pemeriksaan data duplikat menggunakan fungsi *duplicated().sum()*. Hasil pemeriksaan menunjukkan bahwa tidak terdapat data yang terduplikasi pada dataset sehingga seluruh 200 data dapat digunakan pada proses klasifikasi.

```
# ============================================================
# PEMERIKSAAN DUPLIKAT
# ============================================================

print("=" * 60)
print("PEMERIKSAAN DATA DUPLIKAT")
print("=" * 60)

print("Jumlah Data Duplikat :", df.duplicated().sum())
```

```
============================================================
PEMERIKSAAN DATA DUPLIKAT
============================================================
Jumlah Data Duplikat : 0
```

Gambar 7. Hasil Pemeriksaan Data Duplikat

c. *Label Encoding*

Setelah pemeriksaan dan penanganan data selesai, atribut yang masih berbentuk kategorikal dikonversi ke dalam nilai numerik menggunakan metode Label Encoding. Transformasi ini diperlukan agar informasi kategorikal dapat direpresentasikan dalam bentuk yang dapat diterima oleh algoritma Gaussian Naive Bayes. Dengan selesainya proses tersebut, seluruh atribut yang digunakan dalam pemodelan telah memiliki format numerik sehingga dataset dapat dilanjutkan ke tahap pembagian data dan pembentukan model klasifikasi.

```
# ============================================================
# ENCODING DATA KATEGORIK
# ============================================================

label_encoder = LabelEncoder()

kolom_kategori = [
    "Jenis Kelamin",
    "Tekanan Darah",
    "Aktivitas Fisik",
    "Riwayat Diabetes Keluarga",
    "Status Merokok",
    "Diabetes"
]

for kolom in kolom_kategori:
    df[kolom] = label_encoder.fit_transform(df[kolom])

print("="*60)
print("ENCODING BERHASIL")
print("="*60)

df.head()
```

```
============================================================
ENCODING BERHASIL
============================================================
```

	ID Pasien	Umur	Jenis Kelamin	Berat Badan (kg)	Tinggi Badan (cm)	BMI	Tekanan Darah	Glukosa (mg/dL)	Kolesterol (mg/dL)	Aktivitas Fisik	Riwayat Diabetes Keluarga	Status Merokok	Diabetes
0	P001	25.0	1	77	182.0	23.2	0	84.0	178.0	1	0	0	0
1	P002	49.0	1	45	165.0	16.5	0	85.0	186.0	1	0	1	0
2	P003	49.0	1	70	163.0	26.3	0	100.0	198.0	0	0	0	0
3	P004	39.0	0	63	168.0	22.3	0	134.0	232.0	1	0	0	0
4	P005	41.0	1	72	179.0	22.5	0	135.0	224.0	1	1	0	0

Gambar 8. Hasil *Label Encoding*

3.1.3. Pembagian Dataset

Setelah preprocessing selesai, dataset dipisahkan menjadi variabel prediktor (feature) yang terdiri atas 11 atribut masukan dan variabel target (label) berupa atribut Diabetes yang menjadi kelas yang akan diprediksi oleh algoritma Gaussian Naive Bayes. Selanjutnya, fungsi train_test_split() dari pustaka Scikit-learn digunakan untuk membagi dataset dengan komposisi 80% data latih dan 20% data uji serta menetapkan random_state = 42 agar hasil pembagian tetap konsisten pada setiap eksekusi. Dari total 200 data, diperoleh 160 data latih yang mencakup 11 atribut prediktor dan 160 nilai target serta 40 data uji dengan 11 atribut prediktor dan 40 nilai target; hasil pembagian ini kemudian digunakan sebagai dasar dalam proses pelatihan dan evaluasi model Gaussian Naive Bayes, sebagaimana ditampilkan pada Gambar 9.

```
# ============================================================
# TRAIN TEST SPLIT
# ============================================================

X_train, X_test, y_train, y_test = train_test_split(
    X,
    y,
    test_size=0.2,
    random_state=42
)

print("="*60)
print("DATA TRAINING DAN TESTING")
print("="*60)

print("X Train :", X_train.shape)
print("X Test  :", X_test.shape)
print("Y Train :", y_train.shape)
print("Y Test  :", y_test.shape)
```

```
============================================================
DATA TRAINING DAN TESTING
============================================================
X Train : (160, 11)
X Test  : (40, 11)
Y Train : (160,)
Y Test  : (40,)
```

Gambar 9. Hasil Pembagian *Data Training* dan *Data Testing*

Sebagai ringkasan, hasil pembagian dataset disajikan pada **Tabel 2**.

Tabel 2. Hasil Pembagian Dataset

Data	Jumlah Data	Jumlah Atribut Prediktor
Data Training	160	11
Data Testing	40	11

Berdasarkan Tabel 2, komposisi dataset telah mengikuti perbandingan 80:20 sebagaimana ditetapkan dalam metodologi penelitian. Porsi data yang lebih besar dialokasikan sebagai training data untuk proses pembelajaran model Gaussian Naive Bayes, sedangkan bagian lainnya digunakan sebagai testing data untuk mengukur kinerja model pada data yang tidak digunakan selama pelatihan. Pembagian tersebut dimaksudkan untuk menyediakan jumlah data pembelajaran yang memadai sekaligus mempertahankan data yang cukup untuk proses evaluasi model secara objektif.

3.1.4. Hasil Pembangunan Model Gaussian Naïve Bayes

Tahap pemodelan dilakukan setelah dataset selesai dipisahkan menjadi kelompok pelatihan dan pengujian, dengan menerapkan algoritma Gaussian Naive Bayes melalui kelas GaussianNB() dari pustaka Scikit-learn pada Python. Sebanyak 160 data pelatihan yang memuat 11 atribut prediktor digunakan untuk membentuk model, sementara atribut Diabetes ditetapkan sebagai variabel target; melalui proses training tersebut, algoritma mempelajari pola keterkaitan antara karakteristik masukan dan kelas target sebagai dasar dalam menghasilkan prediksi pada data uji.

```
# ============================================================
# MEMBANGUN MODEL GAUSSIAN NAIVE BAYES
# ============================================================

model = GaussianNB()

model.fit(X_train, y_train)

print("="*60)
print("MODEL BERHASIL DILATIH")
print("="*60)
```

```
============================================================
MODEL BERHASIL DILATIH
============================================================
```

Gambar 10. Proses Pembangunan Model *Gaussian Naive Bayes*

Model yang telah selesai dilatih kemudian diterapkan pada 40 data uji yang tidak terlibat dalam proses pembelajaran untuk menghasilkan klasifikasi. Proses prediksi dilakukan menggunakan fungsi predict(), sehingga setiap record pada data pengujian memperoleh hasil kelas berdasarkan pola yang telah dipelajari oleh model..

```
# ============================================================
# PREDIKSI
# ============================================================

y_pred = model.predict(X_test)

hasil = pd.DataFrame({
    "Aktual": y_test.values,
    "Prediksi": y_pred
})

hasil.head(10)
```

	Aktual	Prediksi
0	0	0
1	0	0
2	0	1
3	1	1
4	1	1
5	1	1
6	0	0
7	1	1
8	1	1
9	0	0

Gambar 11. Proses Prediksi Menggunakan Model *Gaussian Naïve Bayes*

Hasil prediksi yang diperoleh selanjutnya digunakan sebagai dasar untuk melakukan evaluasi performa model menggunakan beberapa metrik pengukuran, yaitu *Accuracy*, *Precision*, *Recall*, *F1-Score*, *Confusion Matrix,* dan *Classification Report*, yang akan dibahas pada subbab berikutnya.

3.1.5. Evaluasi Model

Evaluasi dilakukan untuk mengetahui tingkat kemampuan model *Gaussian Naive Bayes* dalam mengklasifikasikan risiko diabetes melitus berdasarkan dataset simulasi yang digunakan. Pengujian kinerja model dilakukan terhadap 40 data yang telah dipisahkan sebagai *testing data*, dengan menggunakan beberapa indikator, yaitu *Accuracy, Precision, Recall, F1-Score, Confusion Matrix*, dan *Classification Report*, sehingga performa model dapat dianalisis secara lebih menyeluruh.

a. *Accuracy, Precision, Recall, dan F1-Score*

Kinerja model klasifikasi dapat dinilai melalui Accuracy, Precision, Recall, dan F1-Score, yang masing-masing memberikan sudut pandang berbeda terhadap hasil prediksi. Accuracy merepresentasikan proporsi keseluruhan prediksi yang sesuai dengan kelas sebenarnya, sedangkan Precision menunjukkan tingkat ketepatan model ketika menetapkan suatu data sebagai kelas positif dan Recall menggambarkan kemampuan model menemukan seluruh data yang memang termasuk kelas positif; sementara itu, F1-Score digunakan untuk melihat keseimbangan antara Precision dan Recall. Hasil pengujian menunjukkan bahwa Gaussian Naive Bayes menghasilkan Accuracy 97,50%, Precision 94,74%, Recall 100%, dan F1-Score 97,30%, yang mengindikasikan kemampuan klasifikasi yang sangat baik dalam membedakan risiko diabetes pada dataset simulasi yang digunakan.

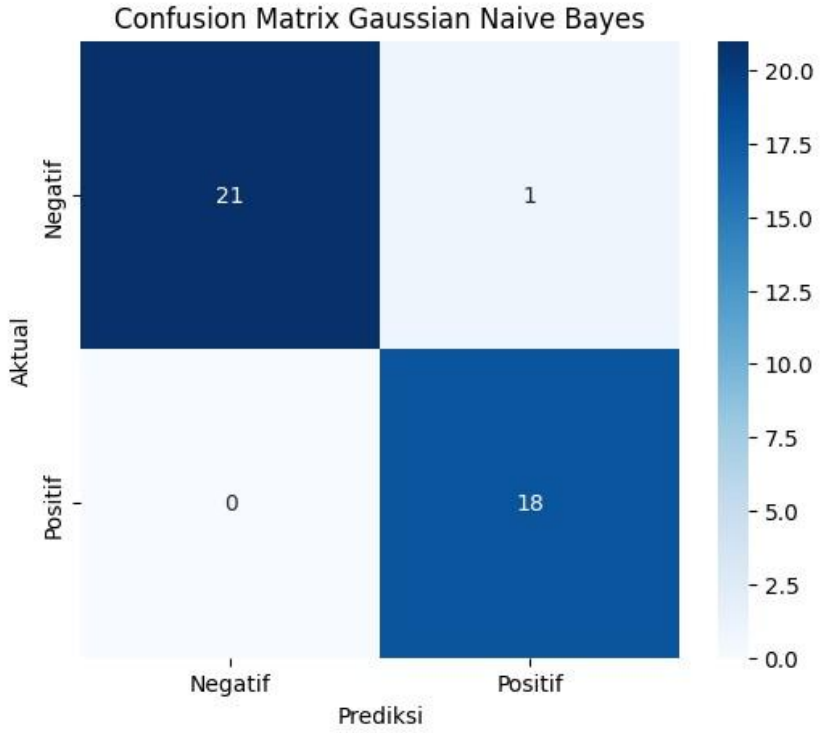
```
# ============================================================
# EVALUASI MODEL
# ============================================================

accuracy = accuracy_score(y_test, y_pred)
precision = precision_score(y_test, y_pred)
recall = recall_score(y_test, y_pred)
f1 = f1_score(y_test, y_pred)

print("="*60)
print("HASIL EVALUASI MODEL")
print("="*60)

print(f"Accuracy  : {accuracy:.2%}")
print(f"Precision : {precision:.2%}")
print(f"Recall    : {recall:.2%}")
print(f"F1-Score  : {f1:.2%}")
```

```
============================================================
HASIL EVALUASI MODEL
============================================================
Accuracy  : 97.50%
Precision : 94.74%
Recall    : 100.00%
F1-Score  : 97.30%
```

Gambar 12. Hasil *Accuracy, Precision, Recall,* dan *F1-Score*

b. *Confusion Matrix*

Untuk melihat kesesuaian antara kelas aktual dan hasil prediksi model, digunakan Confusion Matrix yang menyajikan jumlah klasifikasi benar maupun salah pada setiap kategori. Hasil pengujian model dalam bentuk Confusion Matrix ditampilkan pada Gambar 13.



Gambar 13. *Confusion Matrix Gaussian Naive Bayes*

Berdasarkan *Confusion Matrix* tersebut, diperoleh rincian hasil klasifikasi sebagaimana disajikan pada Tabel 3.

Tabel 3. Hasil *Confusion Matrix*

Aktual	**Prediksi Tidak Diabetes**	**Prediksi Diabetes**
Tidak Diabetes	21	1
Diabetes	0	18

Berdasarkan *Confusion Matrix*, sebanyak 21 data dari kategori Tidak Diabetes berhasil diprediksi secara tepat, sementara 18 data pada kategori Diabetes juga diklasifikasikan dengan benar. Kesalahan prediksi hanya ditemukan pada satu kasus, yaitu data yang sebenarnya termasuk Tidak Diabetes tetapi dikategorikan oleh model sebagai Diabetes, sedangkan seluruh data pada kelas Diabetes berhasil dikenali tanpa kesalahan sebagai Tidak Diabetes.

c. *Classification Report*

Hasil Classification Report memberikan gambaran kinerja model pada masing-masing kelas melalui nilai Precision, Recall, F1-Score, dan Support. Pada kelas Negatif, model menghasilkan Precision 1,00, Recall 0,95, serta F1-Score 0,98, sedangkan kelas Positif memperoleh Precision 0,95, Recall 1,00, dan F1-Score 0,97. Secara keseluruhan, nilai weighted average mencapai 0,98 untuk Precision, 0,97 untuk Recall, dan 0,98 untuk F1-Score, yang menunjukkan bahwa model mampu memberikan hasil klasifikasi dengan kinerja tinggi dan relatif konsisten pada seluruh data pengujian.

```
# ============================================================
# CLASSIFICATION REPORT
# ============================================================

print("="*60)
print("CLASSIFICATION REPORT")
print("="*60)

print(classification_report(
    y_test,
    y_pred,
    target_names=["Negatif","Positif"]
))
```

```
============================================================
CLASSIFICATION REPORT
============================================================
              precision    recall  f1-score   support

     Negatif       1.00      0.95      0.98        22
     Positif       0.95      1.00      0.97        18

    accuracy                           0.97        40
   macro avg       0.97      0.98      0.97        40
weighted avg       0.98      0.97      0.98        40
```

Gambar 14. *Classification Report Gaussian Naive Bayes*

Tingginya nilai *Precision, Recall*, dan *F1-Score* pada kedua kelas memperlihatkan bahwa *Gaussian Naive Bayes* mampu menghasilkan klasifikasi yang konsisten terhadap data simulasi rekam medis pasien. Secara keseluruhan, hasil pengujian dari berbagai metrik menunjukkan kinerja model yang sangat baik pada data uji, dengan rangkuman hasil evaluasi disajikan pada Tabel 4.

Tabel 4. Ringkasan Hasil Evaluasi Model *Gaussian Naive Bayes*

Metrik Evaluasi	Nilai
Accuracy	**97,50%**
Precision	**94,74%**
Recall	**100%**
F1-Score	**97,30%**

Hasil pengujian memperlihatkan bahwa penerapan *Gaussian Naive Bayes* mampu memberikan kinerja klasifikasi yang tinggi pada dataset simulasi rekam medis pasien. Model menghasilkan *Accuracy* sebesar 97,50%, dengan *Precision* mencapai 94,74%, *Recall* 100%, serta *F1-Score* sebesar 97,30%. Capaian tersebut menunjukkan bahwa model memiliki kemampuan yang sangat baik dalam membedakan data berdasarkan kategori risiko diabetes melitus pada data pengujian yang digunakan.

3.2. Pembahasan

Pengujian yang dilakukan memperlihatkan bahwa algoritma *Gaussian Naive Bayes* mampu memberikan hasil klasifikasi yang baik dalam mengidentifikasi risiko diabetes melitus pada dataset simulasi yang dirancang menyerupai rekam medis pasien. Pembahasan hasil penelitian selanjutnya difokuskan pada interpretasi kinerja model berdasarkan metrik evaluasi yang diperoleh, keterkaitannya dengan temuan dari penelitian sebelumnya, serta pembahasan mengenai keunggulan dan keterbatasan pendekatan yang diterapkan.

3.2.1. Analisis Hasil Evaluasi

Hasil pengujian terhadap data uji memperlihatkan bahwa *Gaussian Naive Bayes* menghasilkan *Accuracy* sebesar 97,50%, *Precision* 94,74%, *Recall* 100%, dan *F1-Score* 97,30%. Capaian tersebut menunjukkan bahwa model mampu membedakan pasien yang termasuk kategori diabetes dan tidak diabetes dengan tingkat ketepatan yang tinggi pada dataset simulasi. Nilai *Recall* yang mencapai 100% menjadi salah satu hasil penting karena seluruh data yang sebenarnya berada pada kelas diabetes berhasil dikenali oleh model dan tidak ditemukan *false negative*. Dalam konteks klasifikasi risiko diabetes, kemampuan tersebut memiliki arti penting karena kegagalan mengenali pasien yang berisiko dapat menyebabkan kondisi tersebut tidak teridentifikasi pada tahap awal. Di sisi lain, *Precision* sebesar 94,74% menunjukkan bahwa sebagian besar data yang diprediksi sebagai diabetes memang berasal dari kelas tersebut, sedangkan *F1-Score* sebesar 97,30% memperlihatkan adanya keseimbangan yang baik antara kemampuan menemukan kasus positif dan ketepatan prediksi positif. Konsisten dengan hasil tersebut, *Confusion Matrix* memperlihatkan bahwa 39 dari 40 data pengujian berhasil diklasifikasikan secara tepat dan hanya satu data yang mengalami kesalahan prediksi. Secara keseluruhan, hasil ini menunjukkan bahwa model mampu menangkap pola yang terdapat pada atribut dataset simulasi dan menghasilkan performa klasifikasi yang tinggi, meskipun hasil tersebut tetap perlu ditafsirkan sesuai dengan karakteristik data simulasi yang digunakan

3.2.2. Perbandingan dengan Penelitian Sebelumnya

Pengujian yang dilakukan dalam penelitian ini memperlihatkan bahwa *Gaussian Naive Bayes* mampu menghasilkan kinerja klasifikasi yang tinggi pada dataset simulasi yang merepresentasikan rekam medis pasien. Model memperoleh *Accuracy* sebesar 97,50%, *Precision* 94,74%, *Recall* 100%, dan *F1-Score* 97,30%, yang menunjukkan kemampuan model dalam mengenali pola pada atribut masukan dan menggunakannya untuk menentukan kelas diabetes secara tepat. Hasil tersebut memiliki kesesuaian dengan temuan sejumlah penelitian sebelumnya mengenai penerapan *Naive Bayes* untuk klasifikasi diabetes. Desiani et al. (2024), misalnya, melaporkan bahwa *Naive Bayes* memberikan hasil yang baik dalam proses deteksi diabetes dan menunjukkan kinerja yang lebih optimal ketika menggunakan skema *training split* dibandingkan pendekatan *k-fold cross validation*. Temuan lain dilaporkan oleh Khasanah et al. (2022), yang memperoleh akurasi terbaik sebesar 92,31% melalui algoritma *Naive Bayes* pada konfigurasi pembagian dataset tertentu untuk klasifikasi penyakit diabetes. Kesamaan hasil tersebut memperkuat indikasi bahwa pendekatan berbasis *Naive Bayes* dapat menjadi metode yang efektif untuk menangani permasalahan klasifikasi pada data kesehatan, meskipun tingkat performanya tetap dapat dipengaruhi oleh karakteristik dataset dan strategi pengujian yang diterapkan.

Perbedaan utama penelitian ini dengan studi sebelumnya terletak pada karakteristik data dan rangkaian pengolahan yang diterapkan. Jika penelitian terdahulu umumnya menggunakan sumber data yang telah tersedia, seperti dataset dari *Kaggle*, *National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK)*, atau rekam medis rumah sakit, penelitian ini menggunakan 200 data simulasi yang dirancang untuk menggambarkan karakteristik rekam medis pasien. Pengolahan data dalam penelitian ini dilakukan melalui beberapa tahapan, yaitu *Exploratory Data Analysis (EDA)*, *preprocessing*, pemisahan data dengan komposisi 80% data latih dan 20% data uji, pembentukan model *Gaussian Naive Bayes*, serta pengukuran kinerja menggunakan *Accuracy, Precision, Recall, F1-Score, Confusion Matrix*, dan *Classification Report*.

Nilai evaluasi yang diperoleh pada dataset penelitian ini menunjukkan hasil yang lebih tinggi dibandingkan beberapa hasil yang dilaporkan dalam penelitian terdahulu, tetapi perbedaan tersebut tidak dapat digunakan untuk menyimpulkan bahwa model yang dikembangkan lebih unggul secara universal. Hal ini disebabkan oleh adanya perbedaan jumlah dan karakteristik data, atribut yang digunakan, sumber dataset, serta tahapan pemrosesan dan skema evaluasi pada

masing-masing penelitian. Dengan demikian, temuan penelitian ini lebih tepat diinterpretasikan sebagai bukti bahwa *Gaussian Naive Bayes* memberikan kinerja yang sangat baik pada dataset simulasi yang digunakan, bukan sebagai representasi performa algoritma pada seluruh jenis dataset diabetes.

3.2.3. Kelebihan dan Keterbatasan Penelitian

Dari sisi metodologi, penelitian ini memiliki beberapa aspek yang menjadi keunggulan, salah satunya adalah penggunaan *Gaussian Naive Bayes* yang relatif mudah diimplementasikan karena proses pembelajarannya sederhana dan membutuhkan waktu komputasi yang tidak terlalu besar. Model yang dihasilkan juga menunjukkan kinerja tinggi berdasarkan nilai *Accuracy, Precision, Recall*, dan *F1-Score* pada dataset yang digunakan. Selain itu, rangkaian penelitian disusun secara terstruktur, mulai dari analisis awal melalui *EDA*, persiapan data, pemisahan dataset, pembentukan model, hingga evaluasi, sehingga tahapan yang dilakukan dapat dijadikan acuan untuk penelitian serupa. Namun demikian, terdapat beberapa batasan yang perlu diperhatikan dalam menginterpretasikan hasil penelitian. Dataset yang digunakan hanya berupa 200 data simulasi sehingga belum sepenuhnya mencerminkan keragaman kondisi pasien pada situasi klinis sebenarnya. Penelitian ini juga belum melakukan pengujian komparatif dengan algoritma lain karena hanya berfokus pada *Gaussian Naive Bayes*. Pengembangan selanjutnya dapat diarahkan pada penggunaan dataset yang lebih besar dan berasal dari rekam medis nyata, sekaligus menguji beberapa metode klasifikasi, seperti *Decision Tree, Random Forest*, atau *Support Vector Machine*, sehingga performa berbagai pendekatan dapat dibandingkan secara lebih objektif dan berpotensi menghasilkan model yang lebih optimal.

IV. KESIMPULAN

Penelitian ini berhasil menerapkan algoritma *Gaussian Naive Bayes* dalam proses klasifikasi risiko diabetes melitus menggunakan 200 data simulasi yang dirancang untuk merepresentasikan karakteristik rekam medis pasien. Pelaksanaan penelitian mencakup analisis awal melalui *Exploratory Data Analysis (EDA)*, *preprocessing* data, pemisahan dataset menjadi 80% data pelatihan dan 20% data pengujian, pembentukan model klasifikasi, serta pengukuran kinerja menggunakan *Accuracy, Precision, Recall, F1-Score, Confusion Matrix*, dan *Classification Report*. Hasil pengujian memperlihatkan bahwa model mencapai *Accuracy* 97,50%, *Precision* 94,74%, *Recall* 100%, dan *F1-Score* 97,30%, yang menunjukkan kemampuan klasifikasi yang tinggi pada dataset simulasi yang digunakan. Meskipun demikian, hasil tersebut masih terbatas pada data simulasi dengan jumlah record yang relatif kecil sehingga belum dapat digunakan untuk menggambarkan performa model pada populasi pasien secara umum. Penelitian berikutnya disarankan untuk menggunakan dataset rekam medis nyata dengan jumlah dan keragaman data yang lebih besar serta melakukan perbandingan dengan beberapa algoritma klasifikasi lain agar diperoleh evaluasi yang lebih luas mengenai efektivitas masing-masing metode.

DAFTAR PUSTAKA

[1]. Khasanah, L. U., Nasution, Y. N., Deny, F., & Amijaya, T. (2022). Klasifikasi Penyakit Diabetes Melitus Menggunakan Algoritma Naïve Bayes Classifier. 1(1), 41–50. http://jurnal.fmipa.unmul.ac.id/index.php/basis

[2]. Rozi, F., Dwi Pamungkas, B., Panggayuh, V., & Satria Putra, F. (2024). JoEICT (Journal of Education and ICT) Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/joeict Gaussian Naive Bayes For Early Diabetes Prediction: A Comprehensive Evaluation Of Classification Performance Across Varying Training Test Proportions. 8(2), 63–69. https://doi.org/10.29100/joeict.v8i2.9282

[3]. Irawan, F., Suprapti, T., & Bahtiar, A. (2023). Klasifikasi Algoritma Naive Bayes dan K-Nearest Neighbor pada Penderita Diabetes. 18, 80.

[4]. Arrayyan, A. Z., Setiawan, H., & Putra, K. T. (2024). Naive Bayes for Diabetes Prediction: Developing a Classification Model for Risk Identification in Specific Populations. Semesta Teknika, 27(1), 28–36. https://doi.org/10.18196/st.v27i1.21008

[5]. Najla Salsabila, L., Riyo Dwi Pangga, M., Mauhub Yasser, S., Arin Riyani, N., & Aminah, S. (2024). Application of Naïve Bayes Algorithm for Diabetes Prediction. Jurnal UJMC, 10(1), 56–68.

[6]. Reza Pahlevi, I. (2025). Penerapan Naive Bayes Untuk Prediksi Penyakit Diabetes dengan Menggunakan Rapid Miner. Jurnal Cendekia Ilmiah, 4(4).

[7]. Handayani, F., Firdaus, R., Wahyudi, A., Fu'adah Amran, H., Taufiq, R. M., Informatika, T., Komputer, I., & Riau, U. M. (2025). Kombinasi Algoritma Gaussian Naïve Bayes Dan Adaboost Untuk Meningkatkan Akurasi Dalam Klasifikasi Penyakit Diabetes.

[8]. Nugroho, A, Y. Penerapan Teknik analisis data untuk prediksi penjualan Exploratory Data Analysis (EDA)

[9]. Wororomi, Jonathan, Felix Reba, Samuel Alexander Mandowen, Alvian M.Sroyer, Holomoan Edy manurung. 2024. Data Mining (Memahami Pola Di Balik Angka). CV.EUREKA MEDIA AKSARA

[10]. Permana, Alfin Noval, Juwita Adinda, Roberto Kaban. 2026. Implementasi Naive Bayesuntuk Memprediksi Prestasi Belajar Siswa MTs Fathurrahman Padang Tualan. Merkurius: Jurnal Riset Sistem Informasi dan Teknik InformatikaVolume. 4 Nomor. 4 Juli 2026