

ANALISIS KINERJA METODE *NAÏVE BAYES* DALAM MENGKLASIFIKASIKAN PENYAKIT DIABETES

SKRIPSI

OLEH:

ILYAS NASUTION

218160039



**PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS TEKNIK
UNIVERSITAS MEDAN AREA
MEDAN
2025**

UNIVERSITAS MEDAN AREA

© Hak Cipta Di Lindungi Undang-Undang

Document Accepted 22/12/25

1. Dilarang Mengutip sebagian atau seluruh dokumen ini tanpa mencantumkan sumber
2. Pengutipan hanya untuk keperluan pendidikan, penelitian dan penulisan karya ilmiah
3. Dilarang memperbanyak sebagian atau seluruh karya ini dalam bentuk apapun tanpa izin Universitas Medan Area

Access From (repositori.uma.ac.id)22/12/25

Analisis Kinerja Metode *Naïve Bayes* dalam Mengklasifikasikan Penyakit Diabetes

SKRIPSI

**Diajukan sebagai Salah Satu Syarat untuk Memperoleh
Gelar Sarjana di Fakultas Teknik
Universitas Medan Area**

Oleh:

ILYAS NASUTION

218160039

**PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS TEKNIK
UNIVERSITAS MEDAN AREA
MEDAN
2025**

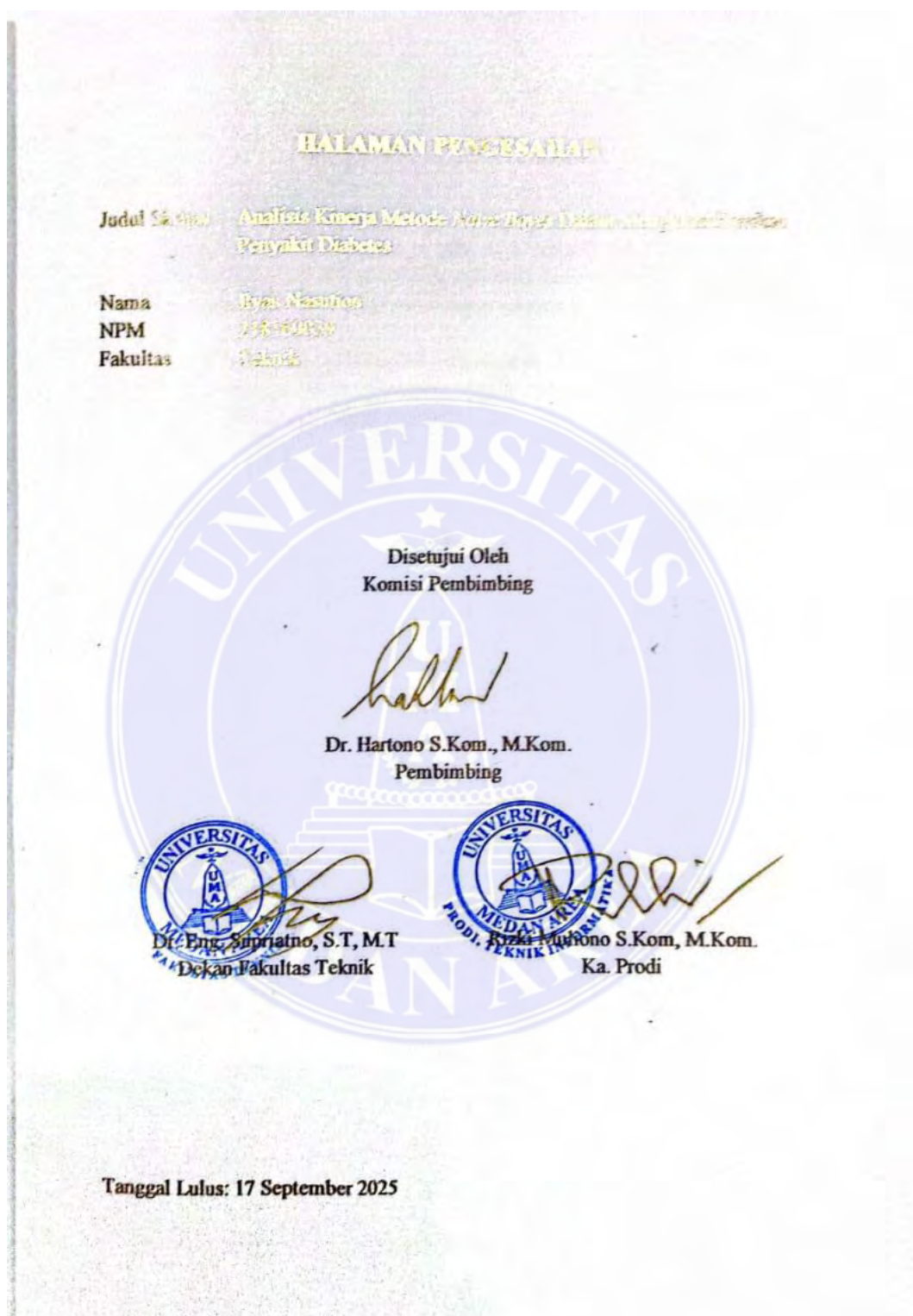
UNIVERSITAS MEDAN AREA

© Hak Cipta Di Lindungi Undang-Undang

1. Dilarang Mengutip sebagian atau seluruh dokumen ini tanpa mencantumkan sumber
2. Pengutipan hanya untuk keperluan pendidikan, penelitian dan penulisan karya ilmiah
3. Dilarang memperbanyak sebagian atau seluruh karya ini dalam bentuk apapun tanpa izin Universitas Medan Area

Document Accepted 22/12/25

Access From (repositori.uma.ac.id)22/12/25



HALAMAN PERNYATAAN

Saya menyatakan bahwa skripsi yang saya susun, sebagai syarat memperoleh gelar serjana merupakan hasil karya tulis saya sendiri. Adapun bagian-bagian tertentu dalam penulisan skripsi ini yang saya kutip dari hasil karya orang lain telah dituliskan sumbernya secara jelas sesuai dengan norma, kaidah, dan etika penulisan ilmiah.

Saya bersedia menerima sanksi pencabutan gelar akademik yang saya peroleh dan sanksi-sanksi lainnya dengan peraturan yang berlaku, apabila di kemudian hari ditemukan adanya plagiat dalam skripsi ini.

Medan, 17 September 2025



Ilyas Nasution
218160039

**HALAMAN PERNYATAAN PERSETUJUAN PUBLIKASI
TUGAS AKHIR/SKRIPSI/TESIS UNTUK KEPENTINGAN AKADEMIS**

Sebagai sivitas akademik Universitas Medan Area, saya yang bertanda tangan di bawah ini:

Nama : Ilyas Nasution
NPM : 218160039
Program Studi : Informatika
Fakultas : Teknik
Jenis Karya : Skripsi

Demi pengembangan ilmu pengetahuan, menyetujui untuk memberikan kepada Universitas Medan Area **Hak Bebas Royalti Noneksklusif (*Non-exclusive Royalty-Free Right*)** atas karya ilmiah saya yang berjudul :

Analisis Kinerja Metode *Naïve Bayes* Dalam Mengklasifikasikan Penyakit Diabetes

Beserta perangkat yang ada (jika diperlukan). Dengan Hak Bebas Royalti Noneksklusif ini Universitas Medan Area berhak menyimpan, mengalihmedia/format-kan, mengelola dalam bentuk pangkalan data (*database*), merawat, dan mempublikasikan Skripsi saya selama tetap mencantumkan nama saya sebagai penulis/pencipta dan sebagai pemilik Hak Cipta.

Demikian pernyataan ini saya buat dengan sebenarnya.

Dibuat di : Medan
Pada tanggal : 17 September 2025
Yang menyatakan



(Ilyas Nasution)
218160039

RIWAYAT HIDUP

Ilyas Nasution lahir di Kota Padang Sidempuan pada tanggal 10 Juni 2003. Penulis merupakan anak kedua dari empat bersaudara dari ayah Darwin Nasution dan ibunda Israwati Harahap. Penulis memulai pendidikan di TK Aisyah Duri, Riau. Kemudian melanjutkan ke jenjang SD Negeri 027 Mandau hingga kelas 5 yang kemudian pindah ke SD Negeri 022 Teladan, Jaya Mukti, Dumai, namun hanya sampai kelas 6. Selanjutnya, penulis kembali pindah ke SD Negeri 0507 Latong, Padang Lawas, Tapanuli Selatan, dan menyelesaikan pendidikan dasar di sekolah tersebut hingga lulus pada tahun 2015.

Pada tahun yang sama, penulis melanjutkan pendidikan di tingkat menengah pertama, yaitu di SMP Negeri 24 Medan, dan menamatkannya pada tahun 2018. Setelah itu, penulis melanjutkan ke jenjang menengah atas di SMKS Tritech Informatika Medan sejak tahun 2018 hingga lulus pada tahun 2021. Setelah itu, penulis melanjutkan pendidikan ke perguruan tinggi di Universitas Medan Area pada tahun 2021 hingga tamat di tahun 2025.

KATA PENGANTAR

Puji syukur kita panjatkan kehadiran Allah SWT, yang telah melimpahkan rahmat dan karunia-Nya sehingga penulis dapat menyelesaikan skripsi yang berjudul "**Analisis Kinerja Metode *Naïve Bayes* Dalam Mengklasifikasikan Penyakit Diabetes**". Skripsi ini disusun sebagai salah satu syarat untuk menyelesaikan pendidikan Program Sarjana pada Fakultas Teknik Program Studi Teknik Informatika di Universitas Medan Area. Dalam penyusunan skripsi ini, penulis telah menerima banyak bantuan, bimbingan, dan dukungan dari berbagai pihak. Oleh karena itu, penulis ingin menyampaikan rasa terima kasih yang sebesar-besarnya kepada :

1. Allah SWT. yang selalu memberikan kenikmatan dan kesehatan sehingga penulis masih bisa menjalankan hidup dengan baik dan selalu memberikan penulis perlindungan dan kekuatan.
2. Bapak Prof. Dr. Dadan Ramdan, M.Eng., M.Sc., selaku Rektor Universitas Medan Area.
3. Bapak Dr. Eng. Supriatno, S.T., M.T. selaku Dekan Fakultas Teknik Universitas Medan Area.
4. Bapak Rizki Muliono, S.Kom., M.Kom., selaku Kepala Program Studi Teknik Informatika.
5. Bapak Dr. Hartono S.Kom, M.Kom., selaku dosen pembimbing yang telah meluangkan waktu untuk membimbing dan memberikan arahan kepada penulis dalam proses menyelesaikan skripsi ini.
6. Bapak Dr. Ir. Rahmad Syah, S.Kom., M.Kom., Ibu Nanda Novita S.Kom., M.Kom., dan Bapak Rizki Muliono, S.Kom., M.Kom., selaku panitia dalam

pelaksanaan pengujian penelitian penulis yang telah memberikan banyak saran dan masukan kepada penulis.

7. Seluruh Dosen Teknik Informatika Universitas Medan Area yang telah memberikan penulis saran yang bermanfaat.
8. Ayah dan Ibunda penulis, Darwin Nasution dan Israwati yang dengan penuh kasih sayang, ketulusan, dan kerja keras telah berjuang untuk menyekolahkan penulis hingga ke jenjang perkuliahan. Doa dan dukungan mereka menjadi kekuatan terbesar bagi penulis untuk terus berjuang dan menyelesaikan skripsi ini.
9. IT Support Teknik Informatika bang Robby Kurniawan Sari Damanik, ST yang telah membantu menyiapkan pemberkasan yang diperlukan dari awal sampai selesai.
10. Teman-teman kelas penulis yang telah memberikan dukungan dan bantuan selama proses penyusunan skripsi ini.

Medan, 17 September 2025



Ilyas Nasution
218160039

ABSTRAK

Diabetes merupakan suatu penyakit kronis yang mempengaruhi jutaan orang di seluruh dunia dan menjadi tantangan besar bagi kesehatan masyarakat. Dalam beberapa tahun terakhir, metode *machine learning* telah banyak diterapkan di bidang medis untuk membantu klasifikasi dan diagnosis penyakit. Diantara berbagai algoritma klasifikasi, metode *Naïve Bayes* mendapat perhatian karena kesederhanaan, efisiensi, dan keefektifannya dalam menangani data probabilistik. Penelitian ini berfokus pada tiga varian *Naïve Bayes*, yaitu *Gaussian Naïve Bayes*, *Bernoulli Naïve Bayes*, dan *Multinomial Naïve Bayes*, yang masing-masing memiliki karakteristik unik yang menjadikannya cocok untuk berbagai jenis dataset dalam klasifikasi medis. Dataset yang digunakan dalam penelitian ini memiliki total 768 baris data dan jumlah 9 kolom. Pra-pemrosesan data meliputi pemeriksaan *missing value*, pembersihan data *missing value*, serta normalisasi data untuk menyamakan skala fitur numerik. Metode yang digunakan ialah *Min-Max Scalling*, yang mengubah nilai fitur menjadi berada dalam rentang 0 hingga 1. Dalam penelitian ini, dilakukan tiga skenario pembagian data dengan proporsi data yang berbeda yaitu 90:10, 80:20, dan 70:30. Evaluasi dilakukan menggunakan *Confusion Matrix* dengan metrik akurasi, presisi, *recall*, dan *f1-score*. Hasil dari pengujian menunjukkan bahwa metode *Gaussian Naïve Bayes* secara konsisten mendapatkan akurasi tertinggi dari tiga skenario pengujian. Pengujian pertama, kedua, dan ketiga metode *Gaussian Naïve Bayes* mendapatkan akurasi sebesar 80%, 82.28%, dan 76.27%. Dari ketiga pengujian tersebut, dapat disimpulkan bahwa *Gaussian Naïve Bayes* lebih optimal dalam mengklasifikasikan penyakit diabetes.

Kata Kunci: Analisis, *Naïve Bayes*, Klasifikasi, Diabetes, *Confusion Matrix*

ABSTRACT

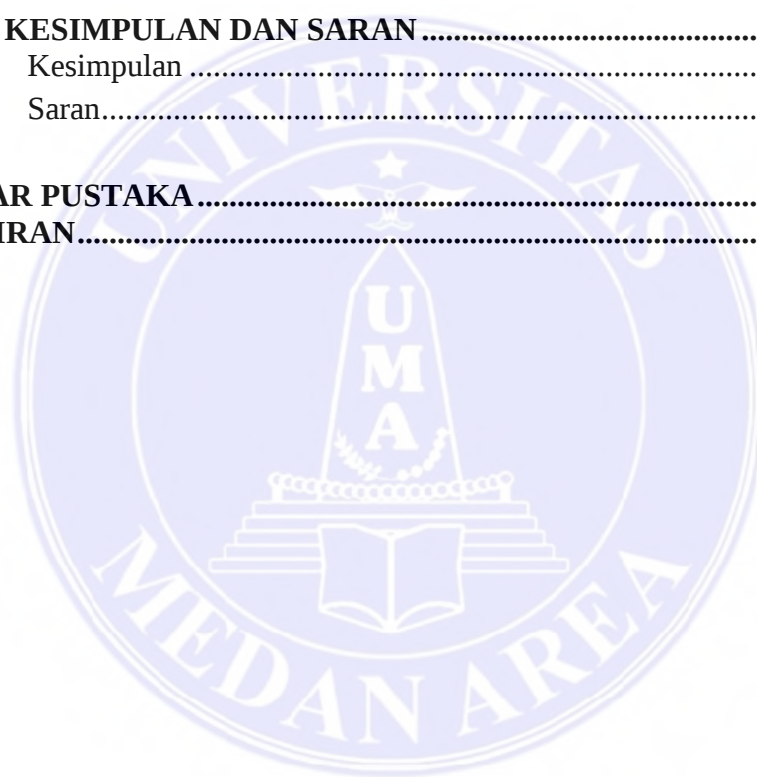
Diabetes is a chronic disease that affects millions of people worldwide and poses a major challenge to public health. In recent years, machine learning methods have been widely applied in the medical field to assist in disease classification and diagnosis. Among various classification algorithms, the Naïve Bayes method has gained attention due to its simplicity, efficiency, and effectiveness in handling probabilistic data. This study focuses on three variants of Naïve Bayes, Gaussian Naïve Bayes, Bernoulli Naïve Bayes, and Multinomial Naïve Bayes, each of which has unique characteristics that make them suitable for different types of datasets in medical classification. The dataset used in this research consists of 768 rows and 9 columns. Data preprocessing includes checking for missing values, cleaning missing data, and normalizing data to standardize the scale of numerical features. The method used is Min-Max Scaling, which transforms feature values into a range between 0 and 1. In this study, three data-splitting scenarios were conducted with different proportions: 90:10, 80:20, and 70:30. Evaluation was performed using the Confusion Matrix with accuracy, precision, recall, and f1-score metrics. The experimental results show that the Gaussian Naïve Bayes method consistently achieved the highest accuracy across the three testing scenarios. In the first, second, and third experiments, Gaussian Naïve Bayes achieved accuracies of 80%, 82.28%, and 76.27%, respectively. Based on these results, it can be concluded that Gaussian Naïve Bayes is more optimal in classifying diabetes disease.

Keywords: Analysis, Naïve Bayes, Classification, Diabetes, Confusion Matrix

DAFTAR ISI

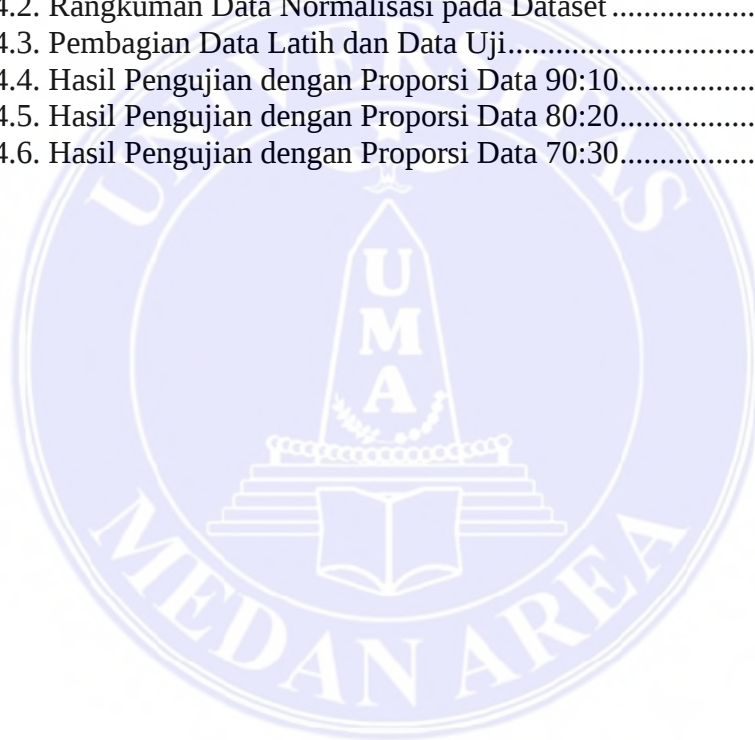
HALAMAN PENGESAHAN	i
HALAMAN PERNYATAAN	Error! Bookmark not defined.
RIWAYAT HIDUP	iv
KATA PENGANTAR	v
ABSTRAK	vii
ABSTRACT	viii
DAFTAR ISI	ix
DAFTAR TABEL	xi
DAFTAR GAMBAR	xii
DAFTAR LAMPIRAN	xiii
 BAB I PENDAHULUAN	 1
1.1. Latar Belakang	1
1.2. Rumusan Masalah	2
1.3. Batasan Masalah.....	3
1.4. Tujuan Penelitian	3
1.5. Manfaat Penelitian	3
 BAB II TINJAUAN PUSTAKA	 5
2.1. Algoritma <i>Naïve Bayes</i>	5
2.1.1. Algoritma <i>Gaussian Naïve Bayes</i>	6
2.1.2. Algoritma <i>Bernoulli Naïve Bayes</i>	7
2.1.3. Algoritma <i>Multinomial Naïve Bayes</i>	8
2.2. Klasifikasi	9
2.3. <i>Confusion Matrix</i>	10
2.4. Penyakit Diabetes.....	11
2.5. Penelitian Terdahulu	12
 BAB III METODOLOGI PENELITIAN	 14
3.1. Kebutuhan Sistem, Perangkat Keras dan Perangkat Lunak	14
3.2. Tahapan Penelitian	15
3.2.2. Pengumpulan Data	16
3.2.3. Pra-pemrosesan Data.....	18
3.2.4. Implementasi	20
3.2.5. Evaluasi	21
3.3. Perhitungan Manual	22
3.3.1. Perhitungan Manual untuk GNB.....	23
3.3.2. Perhitungan Manual untuk BNB	24
3.3.3. Perhitungan Manual untuk MNB.....	25

BAB IV HASIL DAN PEMBAHASAN	28
4.1. Data Penelitian	28
4.2. Pra-pemrosesan Data.....	29
4.2.1. Memeriksa Data <i>Missing Value</i>	29
4.2.2. Penanganan <i>Missing Value</i>	30
4.2.3. Normalisasi Data	31
4.2.4. Pembagian Data (Splitting)	33
4.3. Implementasi	33
4.3.1. Pengujian dengan Proporsi Data 90:10	34
4.3.2. Pengujian dengan Proporsi Data 80:20	36
4.3.3. Pengujian dengan Proporsi Data 70:30	38
BAB V KESIMPULAN DAN SARAN	42
5.1. Kesimpulan	42
5.2. Saran.....	43
DAFTAR PUSTAKA.....	45
LAMPIRAN.....	52



DAFTAR TABEL

	Halaman
Tabel 2.1. Rangkuman Penelitian Terdahulu	12
Tabel 3.1. Rangkuman Dataset	17
Tabel 3.2. Dataset Setelah Prapemrosesan	19
Tabel 3.3. Data Latih <i>Dummy</i>	22
Tabel 3.4. Data Uji <i>Dummy</i>	22
Tabel 3.5. Rata-rata dan Varians per-Kelas	23
Tabel 3.6. Data Setelah Binarisasi	24
Tabel 3.7. Data Total Nilai per-Kelas	25
Tabel 4.1. Deskripsi Atribut	28
Tabel 4.2. Rangkuman Data Normalisasi pada Dataset	32
Tabel 4.3. Pembagian Data Latih dan Data Uji	33
Tabel 4.4. Hasil Pengujian dengan Proporsi Data 90:10	35
Tabel 4.5. Hasil Pengujian dengan Proporsi Data 80:20	37
Tabel 4.6. Hasil Pengujian dengan Proporsi Data 70:30	39



DAFTAR GAMBAR

	Halaman
Gambar 3.1. Alur Tahapan Penelitian	15
Gambar 4.1. Grafik <i>Missing Value</i> per Kolom	29
Gambar 4.2. Grafik <i>Missing Value</i> Setelah Dibersihkan	30
Gambar 4.3. Hasil <i>Confusion Matrix</i> (a) <i>Gaussian Naïve Bayes</i> , (b) <i>Bernoulli Naïve Bayes</i> , (c) <i>Multinomial Naïve Bayes</i> , (d) <i>Naïve Bayes Klasik</i>	36
Gambar 4.4. Hasil <i>Confusion Matrix</i> (a) <i>Gaussian Naïve Bayes</i> , (b) <i>Bernoulli Naïve Bayes</i> , (c) <i>Multinomial Naïve Bayes</i> , (d) <i>Naïve Bayes Klasik</i>	38
Gambar 4.5. Hasil <i>Confusion Matrix</i> (a) <i>Gaussian Naïve Bayes</i> , (b) <i>Bernoulli Naïve Bayes</i> , (c) <i>Multinomial Naïve Bayes</i> , (d) <i>Naïve Bayes Klasik</i>	40



DAFTAR LAMPIRAN

	Halaman
1. Kode Program	52



BAB I

PENDAHULUAN

1.1. Latar Belakang

Diabetes merupakan suatu penyakit kronis yang mempengaruhi jutaan orang di seluruh dunia dan menjadi tantangan besar bagi kesehatan masyarakat. Menurut Organisasi Kesehatan Dunia (WHO), jumlah kasus diabetes terus meningkat, dengan faktor-faktor seperti gaya hidup yang buruk, kurangnya aktivitas fisik, dan dapat juga disebabkan oleh faktor genetik (Ong dkk, 2023).

Dalam beberapa tahun terakhir, metode *machine learning* telah banyak diterapkan di bidang medis untuk membantu klasifikasi dan diagnosis penyakit. Diantara berbagai algoritma klasifikasi, metode *Naïve Bayes* mendapat perhatian karena kesederhanaan, efisiensi, dan keefektifannya dalam menangani data probabilistik. *Naïve Bayes* adalah pengklasifikasi statistik berdasarkan teorema Bayes yang mengasumsikan independensi antara fitur-fitur dalam data (A'an Prasetyo dkk, 2024).

Salah satu pertimbangan utama dalam penerapan metode *Naïve Bayes* adalah pemilihan variannya, karena varian yang berbeda dioptimalkan untuk jenis distribusi data yang berbeda. Penelitian ini berfokus pada tiga varian utama *Naïve Bayes*, yaitu *Gaussian Naïve Bayes*, *Bernoulli Naïve Bayes*, dan *Multinomial Naïve Bayes*, yang masing-masing memiliki karakteristik unik yang menjadikannya cocok untuk berbagai jenis dataset dalam klasifikasi medis (Dhuhita & Zone, 2023). *Gaussian Naïve Bayes* (GNB) mengasumsikan bahwa fitur kontinu mengikuti distribusi normal (*Gaussian*), sehingga sangat berguna untuk dataset yang

mengandung fitur numerik seperti kadar glukosa darah, indeks massa tubuh (BMI), dan usia, yang merupakan faktor penting dalam diagnosis penyakit diabetes (As'ad, 2023). Di sisi lain, *Bernoulli Naïve Bayes* (BNB) dirancang untuk fitur biner atau *boolean*, di mana setiap fitur mewakili ada atau tidaknya suatu karakteristik tertentu. Hal ini menjadikannya efektif dalam skenario di mana dataset mencakup variabel kategorikal yang dapat dikonversi ke dalam bentuk biner, seperti apakah seorang pasien memiliki gejala tertentu atau tidak (Taslim dkk, 2023). Sementara itu, *Multinomial Naïve Bayes* (MNB) sangat efektif dalam klasifikasi teks dan dataset yang fitur-fiturnya berupa frekuensi diskrit atau kategori tertentu (Hassan dkk, 2022).

Ketiga varian ini dipilih untuk memberikan perbandingan komprehensif terhadap kinerja *Naïve Bayes* dalam mengklasifikasikan diabetes. Dengan menganalisis efektivitasnya dalam menangani data numerik, biner, dan kategorikal, penelitian ini bertujuan untuk menentukan kinerja varian yang paling kompatibel dalam mengklasifikasikan penyakit diabetes (Putra dkk, 2024). Evaluasi akan berfokus pada metrik kinerja utama seperti akurasi, presisi, *recall*, dan *f1-score* guna mengukur kinerjanya dalam mengklasifikasikan diabetes.

1.2. Rumusan Masalah

Penelitian ini merumuskan masalah sebagai berikut yaitu, bagaimana hasil dari perbandingan kinerja varian *Naïve Bayes* dalam mengklasifikasikan penyakit diabetes.

1.3. Batasan Masalah

Batasan masalah diperlukan untuk mempermudah penulisan skripsi ini dan agar pembahasan lebih terarah dan sesuai dengan latar belakang yang sudah diuraikan.

Penelitian ini hanya dibatasi pada hal-hal berikut :

1. Metode yang digunakan adalah *Naïve Bayes* varian *Gaussian Naïve Bayes*, *Bernoulli Naïve Bayes*, dan *Multinomial Naïve Bayes*.
2. Metrik evaluasi yang digunakan merupakan akurasi, presisi, *recall*, dan *f1-score*.
3. Bahasa pemrograman yang digunakan adalah *Python*.
4. Data yang digunakan merupakan data sekunder yang diperoleh dari *Mendeley Data*.

1.4. Tujuan Penelitian

Penelitian ini bertujuan untuk membandingkan kinerja varian algoritma *Naïve Bayes* dalam mengklasifikasikan penyakit diabetes guna mengetahui varian yang paling optimal berdasarkan metrik evaluasi.

1.5. Manfaat Penelitian

Adapun manfaat penelitian ini sebagai berikut :

1. Memberikan pemahaman mengenai perbedaan kinerja varian *Naïve Bayes* dalam mengklasifikasikan penyakit diabetes.

2. Memberikan wawasan serta petunjuk bagi penelitian selanjutnya dalam memilih metode klasifikasi yang lebih sesuai untuk permasalahan prediksi penyakit diabetes dari metode yang telah diuji.
3. Menjadi referensi bagi pengembangan kajian lebih lanjut terkait penerapan algoritma klasifikasi pada bidang kesehatan, khususnya dalam pemanfaatan *data mining* untuk mendukung analisis dan pengambilan keputusan berbasis data.
4. Memberikan kontribusi dalam memperkaya literatur terkait penerapan algoritma klasifikasi berbasis *Naïve Bayes* pada kasus kesehatan, sehingga dapat dijadikan acuan bagi penelitian serupa di masa mendatang.
5. Menyediakan gambaran nyata mengenai pengaruh proporsi data latih dan data uji terhadap performa model klasifikasi, yang dapat dijadikan pertimbangan dalam perancangan eksperimen pada penelitian lain.

BAB II

TINJAUAN PUSTAKA

2.1. Algoritma *Naïve Bayes*

Algoritma *Naïve Bayes* adalah salah satu algoritma pengklasifikasi yang paling populer, dengan tingkat akurasi rata-rata yang memuaskan. Metode klasifikasi *Naïve Bayes* didasarkan pada probabilitas sederhana dan dirancang untuk digunakan dengan asumsi independensi antara variabel penjelas (Elistiana dkk, 2023). Algoritma *Naïve Bayes* adalah pengklasifikasi probabilistik sederhana yang menghitung sekumpulan probabilitas dengan menjumlahkan frekuensi dan kombinasi nilai dari sekumpulan data yang diberikan. Berikut adalah bentuk umum Teorema Bayes yang ditunjukkan pada persamaan 2.1 :

$$P(H|X) = \frac{P(X|H) \times P(H)}{P(X)} \quad (2.1)$$

Di sini, X adalah data dengan kelas yang belum diketahui, H adalah hipotesis bahwa X termasuk dalam kelas tertentu, $P(H|X)$ adalah probabilitas hipotesis H berdasarkan data X , $P(X|H)$ adalah probabilitas data X muncul jika hipotesis H benar, $P(H)$ adalah probabilitas awal dari hipotesis H , dan $P(X)$ adalah probabilitas keseluruhan dari data X (Mulyani dkk, 2022). Algoritma *Naïve Bayes* memiliki berbagai varian, di antaranya *Gaussian Naïve Bayes*, *Bernoulli Naïve Bayes*, dan *Multinomial Naïve Bayes*, yang akan digunakan dalam penelitian ini.

2.1.1. Algoritma *Gaussian Naïve Bayes*

Algoritma *Gaussian Naïve Bayes* (GNB) merupakan sebuah metode klasifikasi probabilistik yang didasari oleh Teorema *Bayes* dengan mengasumsikan adanya independensi yang kuat (*Naïve*) antar fitur (Ige & Kiekintveld, 2023). Algoritma GNB sangat sederhana dan mudah diimplementasikan, dan tidak memerlukan banyak data pelatihan. Salah satu kelemahan utama algoritma GNB adalah asumsi bahwa setiap fitur bersifat independen. Dalam GNB, semua fitur dianggap tidak saling bergantung. Meskipun begitu, teknik ekstraksi fitur dapat diterapkan untuk menghasilkan fitur yang lebih independen dari data yang ada (Ampomah dkk, 2021). Berikut merupakan bentuk dari GNB yang ditunjukkan pada persamaan 2.2:

$$P(x_i|C_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \cdot \exp\left(-\frac{(x_i - \mu_k)^2}{2\sigma_k^2}\right) \quad (2.2)$$

Disini, x_i merupakan nilai fitur ke- i dari data uji, μ_k merupakan rata-rata fitur ke- i pada kelas C_k , kemudian σ_k^2 merupakan varians fitur ke- i pada kelas C_k , dan C_k merupakan kelas ke- k (misalnya 0 = non-diabetes dan 1 = diabetes). *Gaussian Naïve Bayes* telah terbukti sebagai metode klasifikasi yang sangat efektif, terutama saat menangani kumpulan data yang kompleks dan jumlah atribut yang besar. Namun, perlu dicatat bahwa asumsi kemandirian antara atribut mungkin tidak akurat pada kumpulan data di mana atribut-atribut tersebut ditemukan memiliki korelasi yang kuat (Chen dkk, 2021).

2.1.2. Algoritma *Bernoulli Naïve Bayes*

Algoritma *Bernoulli Naïve Bayes* adalah metode klasifikator probalistik yang didasarkan pada Teorema Bayes, cocok untuk menangani *output* dua kategori atau *biner*. Algoritma ini mengasumsikan bahwa setiap fitur bersifat independen dan umumnya diterapkan dalam klasifikasi teks, terutama ketika fitur biner lebih dominan (Dirgantara dkk, 2024). Algoritma *Bernoulli Naïve Bayes* mengklasifikasikan data berdasarkan distribusi *multivariate Bernoulli*. Keputusan klasifikasi dalam metode ini yang ditunjukkan pada persamaan 2.3 ditentukan menggunakan aturan berikut (Azizah dkk, 2022) :

$$P(A_i|B) = P(i|B)A_i + (1 - P(i|B))(1 - A_i) \quad (2.3)$$

Algoritma *Bernoulli Naïve Bayes* merupakan varian dari algoritma *Naïve Bayes* yang menggunakan vektor biner untuk merepresentasikan sebuah dokumen, dengan menunjukkan apakah suatu kata muncul atau tidak dalam dokumen tersebut (Yusran dkk, 2024). Dalam konteks klasifikasi, nilai 1 menunjukkan bahwa suatu atribut muncul atau bernilai positif, sedangkan nilai 0 menunjukkan ketiadaan atribut tersebut. Model ini menghitung probabilitas posterior dari suatu kelas dengan mempertimbangkan keberadaan atau ketidakberadaan fitur pada setiap sampel data (Gan dkk, 2021). Kelebihan dari *Bernoulli Naïve Bayes* adalah kesederhanaannya dalam menangani data diskrit biner dan kemampuannya yang cukup baik untuk permasalahan klasifikasi teks, terutama pada kasus *document classification* atau *spam filtering*. Namun demikian, ketika diterapkan pada data yang bersifat kontinu, performa algoritma ini cenderung menurun karena asumsi distribusi biner tidak sesuai dengan karakteristik data (Ahmed & Haruna, 2025).

2.1.3. Algoritma *Multinomial Naïve Bayes*

Algoritma *Multinomial Naïve Bayes* (MNB) adalah metode klasifikasi berbasis probabilitas yang memprediksi peluang di masa depan berdasarkan kejadian yang telah terjadi, yang juga dikenal *Teorema Bayes* (Zulfikar dkk, 2023). Algoritma MNB menganggap bahwa setiap fitur bersifat kondisional independen satu sama lain berdasarkan label kelas dan mengikuti distribusi *multinomial*. Metode ini menghitung probabilitas suatu dokumen termasuk dalam kelas tertentu berdasarkan frekuensi atau jumlah kemunculan fitur (kata) dalam dokumen (Dewi dkk, 2023).

Algoritma MNB beroperasi berdasarkan konsep *term frequency*, yaitu menghitung seberapa sering suatu kata muncul dalam sebuah dokumen. Metode ini menjelaskan dua fakta yaitu keberadaan kata dalam dokumen serta seberapa sering kata tersebut muncul di dalamnya. Untuk rumus dari MNB dapat dilihat pada persamaan 2.4 :

$$P(p|n) \propto P_{(p)} \prod_{1 \leq k \leq nd} P(tk|p) \quad (2.4)$$

Disini, $P(tk|p)$ adalah probabilitas munculnya teks dokumen (tk), n adalah jumlah dokumen dan p adalah polaritas (Yuyun dkk, 2021). Kelebihan dari algoritma ini adalah kesederhanaannya dalam komputasi serta performa yang cukup baik untuk data dengan karakteristik diskrit berfrekuensi. Namun demikian, ketika diterapkan pada data numerik kontinu tanpa proses transformasi atau diskretisasi, performa MNB cenderung menurun karena asumsi distribusinya tidak sesuai dengan karakteristik data tersebut (Wang dkk, 2023).

2.2. Klasifikasi

Klasifikasi merupakan suatu metode yang digunakan untuk mengelompokkan sebuah entitas ke dalam salah satu kelas yang telah ditentukan sebelumnya. Tujuan dari klasifikasi adalah memprediksi kelas dari suatu item yang belum diketahui kategorinya (Yolanda Paramitha dkk, 2023). Dalam proses klasifikasi data, sangat penting untuk berhati-hati dalam membedakan antara data pelatihan dan data uji. Data pelatihan didefinisikan sebagai data yang proporsinya lebih besar daripada data uji. Data tersebut kemudian diproses oleh komputer dengan tujuan menghasilkan model (Ramadhon dkk, 2024).

Proses klasifikasi terdiri dari tiga tahap, yaitu pembangunan model, penerapan model, dan evaluasi hasil. Pada tahap pembangunan model, sistem dibentuk menggunakan data pelatihan yang telah memiliki atribut dan label kelas. Informasi ini digunakan sebagai dasar untuk menentukan kelas dari data atau objek baru. Selanjutnya, data hasil klasifikasi dievaluasi untuk mengukur tingkat akurasi dari model yang telah dikembangkan dan diterapkan pada data yang belum pernah dilihat sebelumnya (Pebdika dkk, 2023).

Berbagai algoritma telah dikembangkan untuk menyelesaikan permasalahan klasifikasi, di antaranya adalah *Decision Tree*, *Support Vector Machine (SVM)*, *K-Nearest Neighbor (KNN)*, dan *Naïve Bayes* (Sheth dkk, 2022). Setiap algoritma memiliki kelebihan dan kekurangan bergantung pada karakteristik data yang digunakan. *Naïve Bayes* dikenal sederhana dalam perhitungan probabilitasnya dan efisien dalam menangani dataset berukuran besar, sehingga sering dijadikan sebagai salah satu metode dasar dalam penelitian klasifikasi (Blanquero dkk, 2021).

2.3. Confusion Matrix

Confusion matrix merupakan *tools* evaluasi yang umum digunakan dalam *machine learning* untuk menilai performa model klasifikasi. Matriks ini menyajikan perbandingan antara label kelas yang dihasilkan oleh model dengan label kelas sebenarnya dalam bentuk tabel. Meskipun awalnya dirancang untuk klasifikasi biner, penggunaannya dapat diperluas untuk mengakomodasi kasus klasifikasi multikelas (Berliana & Riassetiawan, 2024). *Confusion matrix* menyediakan berbagai metrik evaluasi untuk menilai kinerja model klasifikasi, di antaranya adalah akurasi, presisi, *recall*, dan *f1-score* (Simanjuntak dkk, 2023).

$$Accuracy = \frac{TN + TP}{TN + FP + TP + FN} \quad (2.5)$$

$$Precision = \frac{TP}{TP + FP} \quad (2.6)$$

$$Recall = \frac{TP}{TP + FN} \quad (2.7)$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \quad (2.8)$$

Confusion Matrix berbentuk tabel dua dimensi yang menyajikan informasi mengenai jumlah prediksi model yang benar maupun salah dengan membandingkan hasil prediksi terhadap nilai aktual (Miller dkk, 2024). Pada kasus klasifikasi, *Confusion Matrix* terdiri dari empat komponen utama, yaitu *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN). Nilai-nilai tersebut digunakan untuk menghitung berbagai metrik evaluasi seperti akurasi, presisi, *recall*, dan *f1-score* (Rainio dkk, 2024).

Kelebihan penggunaan *Confusion Matrix* adalah kemampuannya memberikan gambaran yang lebih detail dibandingkan hanya menggunakan akurasi semata. Dengan *Confusion Matrix*, peneliti dapat mengetahui distribusi kesalahan model, misalnya apakah model lebih sering salah memprediksi kelas positif sebagai negatif atau sebaliknya (Okwudili dkk, 2025). Hal ini sangat penting terutama pada kasus klasifikasi di bidang kesehatan, seperti prediksi penyakit diabetes, di mana kesalahan prediksi pada salah satu kelas dapat menimbulkan konsekuensi yang signifikan (Ansari dkk, 2025).

Secara umum, *Confusion Matrix* menjadi *tool* evaluasi standar yang dapat dijadikan acuan untuk menganalisis kualitas model klasifikasi. Informasi yang dihasilkan membantu dalam menentukan apakah model sudah cukup baik dalam melakukan generalisasi atau masih memerlukan perbaikan pada aspek tertentu, seperti pemilihan algoritma, *tuning* parameter, atau teknik pra-pemrosesan data (Markoulidakis & Markoulidakis, 2024).

2.4. Penyakit Diabetes

Penyakit diabetes atau *diabetes melitus* (DM) adalah suatu penyakit kronis terjadi ketika pankreas tidak mampu memproduksi insulin dalam jumlah yang cukup, atau ketika tubuh tidak dapat memanfaatkan insulin yang dihasilkan secara efektif. Insulin sendiri merupakan hormon yang berperan dalam mengatur kadar gula dalam darah (Lissentiya Armal dkk, 2023). Penyebab pasti dari penyakit ini masih belum sepenuhnya diketahui. Namun, para ahli medis menduga bahwa faktor gaya hidup dan pola makan memiliki peransignifikan dalam perkembangan penyakit tersebut. Penderita diabetes juga berisiko tinggi mengalami berbagai

komplikasi kesehatan sekunder, seperti penyakit jantung, kerusakan saraf, dan gangguan lainnya (Wafa dkk, 2022).

Komplikasi diabetes dapat bersifat akut maupun kronis. Komplikasi akut meliputi *hipoglikemia* dan *ketoasidosis diabetik*, sedangkan komplikasi kronis mencakup kerusakan pada organ vital seperti jantung, ginjal, saraf, dan mata (Jha dkk, 2024). Kondisi *hiperglikemia*, yaitu peningkatan kadar gula darah, merupakan dampak umum dari diabetes yang tidak terkontrol dan dapat menyebabkan kerusakan serius pada berbagai sistem tubuh, terutama pada sistem saraf dan pembuluh darah (Aminuddin dkk, 2023).

Oleh karena itu, deteksi dini serta klasifikasi yang tepat terhadap kondisi diabetes menjadi langkah penting dalam upaya pencegahan maupun penanganan penyakit ini.

2.5. Penelitian Terdahulu

Penelitian terdahulu dibutuhkan sebagai acuan untuk melihat sejauh mana topik yang diteliti telah dikaji sebelumnya. Berikut rangkuman dari penelitian terdahulu yang dapat dilihat pada tabel 2.1 :

Tabel 2.1. Rangkuman Penelitian Terdahulu

Penulis (Tahun)	Judul	Metode	Hasil
(Simanjuntak, Muhathir, Fadlisyah, Safira, 2023)	<i>Performance Analysis of Naïve Bayes Variation Method in Spice Image Classification Using Histogram of Gradient Oriented (HOG) Feature Extraction</i>	<i>Gaussian Naïve Bayes, Bernoulli Naïve Bayes, Multinomial Naïve Bayes, & Complement Naïve Bayes</i>	Metrik kinerja penelitian menunjukkan bahwa metode <i>Gaussian Naïve Bayes</i> menunjukkan hasil terbaik, dengan akurasi sebesar 0,946, presisi 0,95, <i>recall</i> 0,945, <i>f1-score</i> 0,947, <i>f-beta score</i> 0,946, dan <i>Jaccard score</i> 0,90. Sementara itu, metode <i>Complement Naïve Bayes</i> menunjukkan kinerja terendah dibandingkan metode lainnya.

(Muhathir, Muliono, Hafni, 2022)	<i>Image Classification of Autism Spectrum Disorder Children Using Naïve Bayes Method With Hog Feature Extraction</i>	<i>Bernoulli Naïve Bayes, Multinomial Naïve Bayes, & Gaussian Naïve Bayes</i>	Diantara skor untuk index evaluasi kinerja yang diperoleh, skor <i>Bernoulli Naïve Bayes</i> adalah yang tertinggi dan <i>accuracy</i> , <i>precision</i> , <i>recall</i> dan <i>f1-score</i> masing-masing, 89,725%, 90,54%, 89.725%, 89.9%.
(Siregar & Hasudungan Lubis, 2025)	Analisis Sentimen Komentar Pengunjung Terhadap Tempat Wisata <i>Tjong A Fie</i> Mansion Menggunakan Metode <i>Naïve Bayes Classifier</i>	<i>Multinomial Naïve Bayes</i>	Berdasarkan hasil penelitian, metode <i>Multinomial Naïve Bayes</i> terbukti efektif dalam mengklasifikasikan komentar pengunjung ke dalam kategori positif dan negatif. Pengujian dilakukan dalam tiga skenario pembagian data, dengan hasil akurasi tertinggi sebesar 80% pada skenario pelatihan 80% dan pengujian 20%, dan hasil sentimen positif sebesar 100%.
(Dhuhita & Zone., 2023)	Perbandingan Kinerja Algoritma <i>Naïve Bayes Model Multinomial dan Bernoulli</i> pada Pengklasifikasian Komentar <i>Cyberbullying</i>	<i>Multinomial Naïve Bayes & Bernoulli Naïve Bayes</i>	Penelitian ini membandingkan <i>Multinomial</i> dan <i>Bernoulli Naïve Bayes</i> dalam klasifikasi komentar <i>cyberbullying</i> . Dari 650 data uji, <i>Multinomial Naïve Bayes</i> dengan Bag of Words pada fold 10 mencatat akurasi tertinggi sebesar 0,769 dan waktu pemrosesan tercepat 0,35 detik. Dalam pengujian manual, <i>Multinomial</i> memprediksi 10 dari 10 data dengan benar, sedangkan <i>Bernoulli</i> hanya 9 dari 10.
(Ihwana As'ad, 2023)	<i>Advancing Healthcare Diagnostics: A Study on Gaussian Naïve Bayes Classification of Blood Samples</i>	<i>Gaussian Naïve Bayes</i>	Performa pengujian dievaluasi menggunakan <i>5-fold cross-validation</i> dan menghasilkan nilai rata-rata akurasi 88.10%, presisi 89.00%, <i>recall</i> 88.10%, dan <i>f-measure</i> 87.92%

BAB III

METODOLOGI PENELITIAN

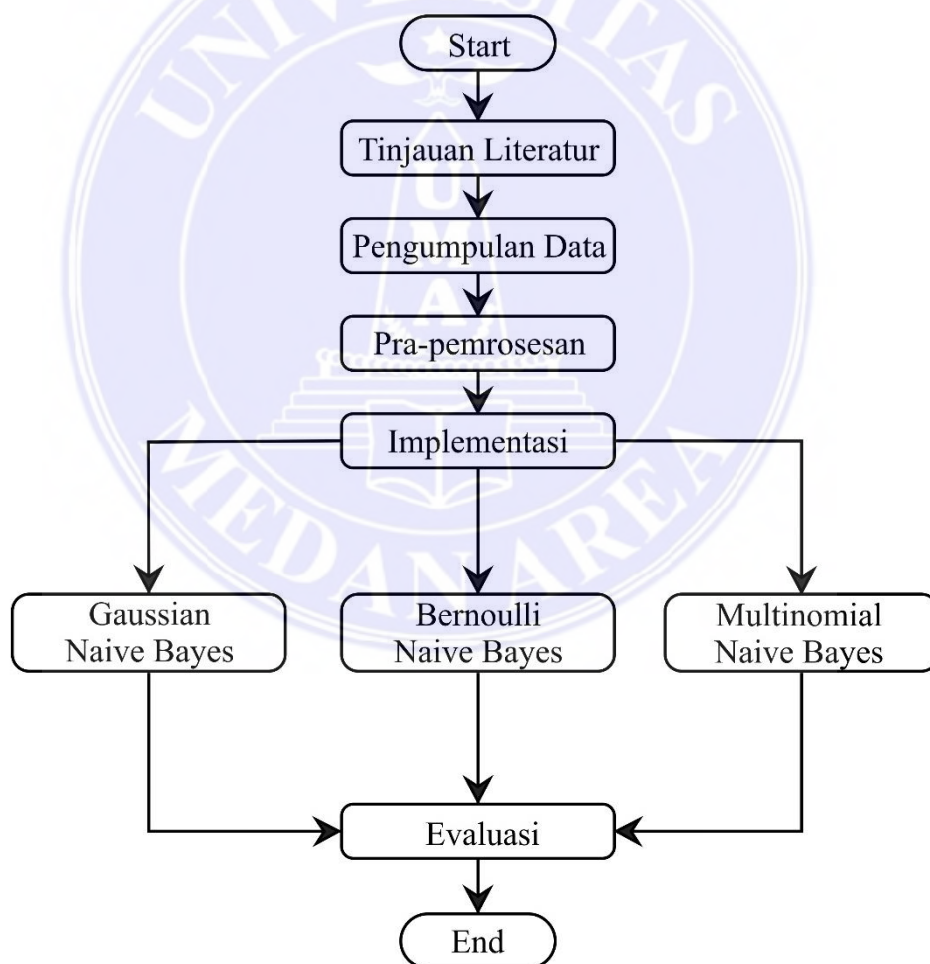
3.1 Kebutuhan Sistem, Perangkat Keras dan Perangkat Lunak

Penelitian ini membutuhkan dukungan perangkat keras dan perangkat lunak yang memadai agar proses pengolahan data, penerapan algoritma, serta analisis hasil dapat berjalan dengan optimal. Pemilihan perangkat yang tepat sangat penting karena akan memengaruhi kecepatan komputasi, efisiensi pemrosesan, serta akurasi dalam menghasilkan temuan penelitian. Spesifikasi perangkat keras yang digunakan berperan dalam menyediakan kapasitas penyimpanan, daya pemrosesan, serta stabilitas sistem selama penelitian berlangsung. Sementara itu, perangkat lunak diperlukan sebagai sarana utama untuk melakukan implementasi metode klasifikasi, pengolahan dataset, visualisasi hasil, dan evaluasi model. Adapun kebutuhan perangkat keras dan perangkat lunak yang diperlukan dalam pembuatan sistem ini terdiri dari :

1. Untuk perangkat keras ialah sebagai berikut :
 - a. Laptop,
 - b. Prosesor *Intel Core i3* generasi ke-7,
 - c. Ram 8GB, dan
 - d. *SSD* 256GB.
2. Untuk perangkat lunak ialah sebagai berikut :
 - a. Bahasa pemrograman *Python*, dan
 - b. *Spyder*.

3.2 Tahapan Penelitian

Pada tahapan penelitian ini dijelaskan mengenai alur penelitian yang dilakukan, mulai dari tinjauan literatur hingga evaluasi hasil klasifikasi. Metode yang digunakan meliputi *Gaussian Naïve Bayes*, *Bernoulli Naïve Bayes*, dan *Multinomial Naïve Bayes* yang masing-masing diimplementasikan untuk mengklasifikasikan penyakit diabetes berdasarkan atribut yang tersedia pada dataset. Untuk gambaran yang lebih jelas mengenai alur penelitian yang dilakukan, dapat dilihat pada Gambar 3.1 berikut.



Gambar 3.1. Alur Tahapan Penelitian

3.2.1. Tinjauan Literatur

Pada tahap ini, peneliti melakukan kajian literatur yang bertujuan untuk memperoleh pemahaman mendalam mengenai konsep dan pendekatan yang berkaitan dengan analisis kinerja metode *Gaussian Naïve Bayes*, *Bernoulli Naïve Bayes*, dan *Multinomial Naïve Bayes* dalam proses klasifikasi data. Analisis terhadap penelitian terdahulu dilakukan untuk mengetahui bagaimana algoritma tersebut digunakan, keunggulan dan keterbatasan masing-masing varian, serta teknik evaluasi yang diterapkan seperti akurasi, presisi, *recall*, dan *f1-score*. Selain itu, kajian ini juga membantu peneliti dalam mengidentifikasi variabel-variabel penting yang berperan dalam proses klasifikasi dan menentukan pendekatan paling tepat yang dapat diadopsi dalam penelitian ini.

3.2.2. Pengumpulan Data

Data yang digunakan dalam penelitian ini merupakan data sekunder yang diperoleh melalui *website* Mendeley Data pada tautan berikut (<https://data.mendeley.com/datasets/7zcc8v6hvp/1>). Dataset yang digunakan terdiri dari 9 atribut yang menjelaskan mengenai umur, tekanan darah, indeks massa tubuh, dan glukosa. Berikut merupakan rangkuman dari data pada penelitian ini yang dapat dilihat pada tabel 3.1 :

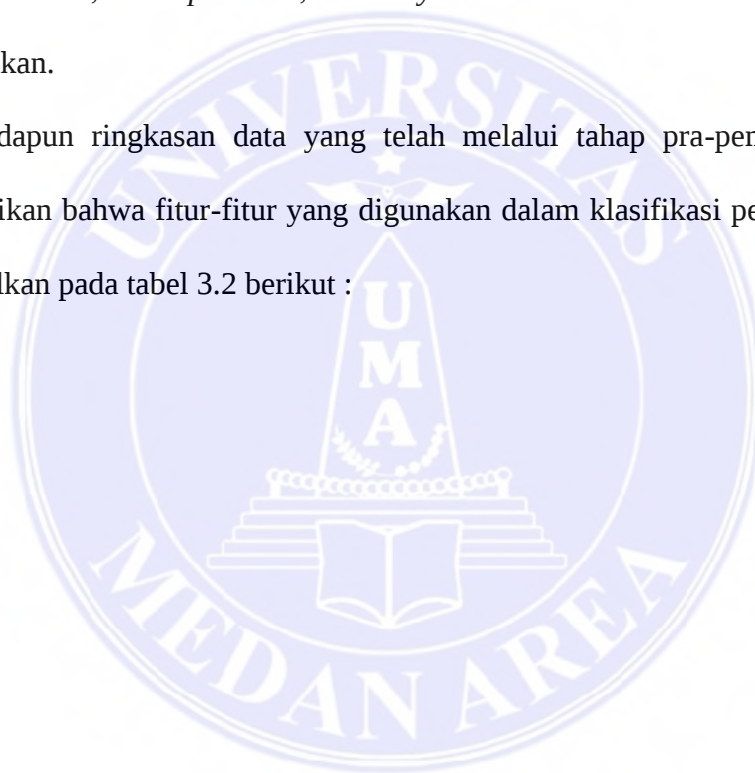
Tabel 3.1. Rangkuman Dataset

Pregnancies	Glucose	Blood pressure	Skin thickness	Insulin	Body mass index	Diabetes pedigree function	Age	Outcome
6	148	72	35	0	33,6	0,627	50	1
1	85	66	29	0	26,6	0,351	31	0
8	183	64	0	0	23,3	0,672	32	1
1	89	66	23	94	28,1	0,167	21	0
0	137	40	35	168	43,1	2,288	33	1
5	116	74	0	0	25,6	0,201	30	0
3	78	50	32	88	31	0,248	26	1
10	115	0	0	0	35,3	0,134	29	0
2	197	70	45	543	30,5	0,158	53	1
8	125	96	0	0	0	0,232	54	1
4	110	92	0	0	37,6	0,191	30	0
10	168	74	0	0	38	0,537	34	1
10	139	80	0	0	27,1	1,441	57	0
1	189	60	23	846	30,1	0,398	59	1
5	166	72	19	175	25,8	0,587	51	1
7	100	0	0	0	30	0,484	32	1
0	118	84	47	230	45,8	0,551	31	1
7	107	74	0	0	29,6	0,254	31	1
1	103	30	38	83	43,3	0,183	33	0
1	115	70	30	96	34,6	0,529	32	1
3	126	88	41	235	39,3	0,704	27	0
8	99	84	0	0	35,4	0,388	50	0
7	196	90	0	0	39,8	0,451	41	1

3.2.3. Pra-pemrosesan Data

Pra-pemrosesan data dalam analisis kinerja metode varian *Naïve Bayes* mencakup sejumlah tahapan yang bertujuan untuk memastikan bahwa data yang digunakan berada dalam format yang sesuai dan layak untuk masing-masing varian algoritma *Naïve Bayes*, yaitu *Gaussian*, *Bernoulli*, dan *Multinomial*. Pada dataset yang digunakan terdapat beberapa data kosong/nilai nol tidak wajar, seperti pada kolom *Glucose*, *Blood pressure*, dan *Body mass index*. Nilai-nilai ini nantinya perlu dibersihkan.

Adapun ringkasan data yang telah melalui tahap pra-pemrosesan untuk memastikan bahwa fitur-fitur yang digunakan dalam klasifikasi penyakit diabetes ditampilkan pada tabel 3.2 berikut :



Tabel 3.2. Dataset Setelah Prapemrosesan

Pregnancies	Glucose	Blood pressure	Skin thickness	Insulin	Body mass index	Diabetes pedigree function	Age	Outcome
1	89	66	23	94	28,1	0,167	21	0
0	137	40	35	168	43,1	2,288	33	1
3	78	50	32	88	31	0,248	26	1
2	197	70	45	543	30,5	0,158	53	1
1	189	60	23	846	30,1	0,398	59	1
5	166	72	19	175	25,8	0,587	51	1
0	118	84	47	230	45,8	0,551	31	1
1	103	30	38	83	43,3	0,183	33	0
1	115	70	30	96	34,6	0,529	32	1
3	126	88	41	235	39,3	0,704	27	0
11	143	94	33	146	36,6	0,254	51	1
10	125	70	26	115	31,1	0,205	41	1
1	97	66	15	140	23,2	0,487	22	0
13	145	82	19	110	22,2	0,245	57	0
3	158	76	36	245	31,6	0,851	28	1
3	88	58	11	54	24,8	0,267	22	0
4	103	60	33	192	24	0,966	33	0
4	111	72	47	207	37,1	1,39	56	1
3	180	64	25	70	34	0,271	26	0
9	171	110	24	240	45,4	0,721	54	1
1	103	80	11	82	19,4	0,491	22	0
1	101	50	15	36	24,2	0,526	26	0
5	88	66	21	23	24,4	0,342	30	0

3.2.4. Implementasi

Pada tahap ini, peneliti melakukan pengujian terhadap tiga metode klasifikasi varian *Naïve Bayes*, yaitu *Gaussian Naïve Bayes*, *Bernoulli Naïve Bayes*, dan *Multinomial Naïve Bayes* pada dataset yang telah melalui tahap pra-pemrosesan. Dataset yang digunakan sebelumnya telah dibersihkan dari nilai yang hilang, dinormalisasi, serta dipisahkan antara fitur dan label, sehingga siap untuk diolah lebih lanjut.

Implementasi metode dilakukan dengan memanfaatkan bahasa pemrograman *Python*, yang dikenal luas dalam bidang *data science* karena memiliki *library* yang kaya dan mendukung analisis data secara komprehensif. Lingkungan pengembangan yang digunakan adalah *Spyder*, bagian dari distribusi *Anaconda*, yang menyediakan *interface* interaktif serta kemudahan dalam menulis, menjalankan, dan menganalisis kode program.

Masing-masing metode *Naïve Bayes* diterapkan untuk mempelajari pola dari data latih, kemudian diuji pada data uji dengan proporsi tertentu (90:10, 80:20, dan 70:30). Tujuan dari implementasi ini adalah untuk memperoleh gambaran perbandingan efektivitas klasifikasi antar varian, baik dari segi akurasi maupun keseimbangan prediksi pada kelas positif dan negatif.

Hasil dari proses klasifikasi dievaluasi menggunakan *Confusion Matrix* beserta metrik turunannya, yaitu akurasi, presisi, *recall*, dan *f1-score*. Evaluasi ini memberikan informasi yang lebih detail mengenai kemampuan masing-masing model dalam mengklasifikasikan data diabetes. Dengan demikian, implementasi ini tidak hanya menampilkan hasil klasifikasi, tetapi juga memberikan dasar yang kuat untuk melakukan analisis perbandingan antar varian metode *Naïve Bayes*.

3.2.5. Evaluasi

Tahap evaluasi bertujuan untuk menilai kinerja masing-masing algoritma yang diuji, yaitu *Gaussian Naïve Bayes*, *Bernoulli Naïve Bayes*, dan *Multinomial Naïve Bayes* dalam mengklasifikasikan penyakit diabetes. Proses evaluasi ini dilakukan menggunakan dataset uji yang telah dipisahkan pada tahap pra-pemrosesan, sehingga hasil pengujian benar-benar mencerminkan kemampuan model dalam melakukan generalisasi terhadap data baru.

Tool utama yang digunakan dalam tahap evaluasi adalah *Confusion Matrix*, yang mampu menyajikan informasi detail mengenai jumlah prediksi benar maupun salah pada masing-masing kelas. Dari *Confusion Matrix* kemudian diturunkan sejumlah metrik evaluasi, yaitu akurasi, presisi, *recall*, dan *f1-score*. Akurasi digunakan untuk melihat sejauh mana model mampu memberikan prediksi yang tepat secara keseluruhan. Presisi menilai sejauh mana model dapat memprediksi kelas positif secara benar dibandingkan seluruh prediksi positif yang dihasilkan. *Recall* digunakan untuk mengukur kemampuan model dalam mengenali seluruh data yang benar-benar termasuk ke dalam kelas positif. Sedangkan *f1-score* menjadi ukuran keseimbangan antara presisi dan *recall*, terutama ketika distribusi data antar kelas tidak seimbang.

Semakin tinggi nilai dari keempat metrik tersebut, maka semakin baik pula kinerja metode yang diuji. Setelah metrik evaluasi dihitung, hasilnya dibandingkan antar varian metode *Naïve Bayes* untuk menentukan metode yang paling optimal. Dengan demikian, tahap evaluasi ini tidak hanya berfokus pada satu ukuran kinerja saja, melainkan memberikan gambaran yang lebih menyeluruh mengenai kelebihan dan kelemahan masing-masing metode.

Temuan dari tahap evaluasi kemudian digunakan sebagai dasar dalam penarikan kesimpulan mengenai efektivitas setiap metode *Naïve Bayes* dalam klasifikasi penyakit diabetes. Hasil ini juga menjadi acuan dalam memberikan rekomendasi terhadap metode yang paling sesuai untuk digunakan pada penelitian sejenis di masa mendatang.

3.3. Perhitungan Manual

Pada penelitian ini, perhitungan manual digunakan untuk memberikan gambaran tentang bagaimana proses klasifikasi dilakukan. Perhitungan manual ini akan menggunakan data *dummy* atau data contoh yang terdiri dari tiga fitur yaitu *glucose*, *body mass index*, dan, *age*, serta satu kolom kelas (*outcome*) yang bernilai 0 untuk non-diabetes dan 1 untuk diabetes. Adapun perhitungan manual ini akan dibagi menjadi tiga bagian, yaitu perhitungan untuk metode *Gaussian Naïve Bayes*, *Bernoulli Naïve Bayes*, dan *Multinomial Naïve Bayes*. Berikut ini adalah langkah-langkah perhitungannya :

Tabel 3.3. Data Latih *Dummy*

Glucose	Body Mass index	Age	Outcome
150	30.0	50	1
130	28.0	45	1
100	25.0	30	0
90	24.0	25	0

Tabel 3.4. Data Uji *Dummy*

Glucose	Body Mass index	Age
120	27.0	40

3.3.1. Perhitungan Manual untuk GNB

Gaussian Naïve Bayes mengasumsikan bahwa setiap fitur numerik mengikuti distribusi normal. Probabilitas dihitung dengan rumus distribusi *Gaussian* :

$$P(X_i|C) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left(-\frac{(X_i - \mu)^2}{2\sigma^2}\right) \quad (3.1)$$

1. Menghitung nilai rata-rata (μ) dan varians (σ^2):

Tabel 3.5. Rata-rata dan Varians per-Kelas

Fitur	μ (C=1)	σ^2 (C=1)	μ (C=0)	σ^2 (C=0)
Glucose	140	100	95	25
BMI	29.0	1.0	24.5	0.25
Age	47.5	6.25	27.5	6.25

2. Menghitung probabilitas *Likelihood*

Kelas 1 :

$$P(\text{Glucose} = 120|C) = \frac{1}{\sqrt{2\pi \cdot 100}} \cdot \exp\left(-\frac{(120 - 140)^2}{2 \cdot 100}\right) \approx 0.0147$$

$$P(\text{BMI} = 27|C) = \frac{1}{\sqrt{2\pi \cdot 1}} \cdot \exp\left(-\frac{(27 - 29)^2}{2 \cdot 1}\right) \approx 0.05399$$

$$P(\text{Age} = 40|C) = \frac{1}{\sqrt{2\pi \cdot 6.25}} \cdot \exp\left(-\frac{(40 - 47.5)^2}{2 \cdot 6.25}\right) \approx 0.05513$$

$$P(C = 1|X) \propto 0.5 \cdot 0.0147 \cdot 0.05399 \cdot 0.05513 \approx 6.40 \times 10^{-6}$$

Kelas 0 :

$$P(\text{Glucose} = 120|C = 0) \approx 2.97 \times 10^{-6}$$

$$P(\text{BMI} = 27|C = 0) \approx 6.07 \times 10^{-5}$$

$$P(\text{Age} = 40|C = 0) \approx 0.0141$$

$$P(C = 0|X) \propto 0.5 \cdot 2.97 \times 10^{-6} \cdot 6.07 \times 10^{-5} \cdot 0.0141 \approx 1.29 \times 10^{-11}$$

Sehingga, data dapat diklasifikasikan kedalam **Kelas 1** (Diabetes).

3.3.2. Perhitungan Manual untuk BNB

Metode *Bernoulli Naïve Bayes* digunakan ketika fitur berupa data biner. Oleh karena itu, dilakukan proses binarisasi fitur berdasarkan ambang batas tertentu (threshold), yaitu median dari masing-masing fitur.

1. Threshold binarisasi (berdasarkan median):

- a. Glucose $\geq 120 \rightarrow 1$
- b. BMI $\geq 26 \rightarrow 1$
- c. Age $\geq 35 \rightarrow 1$

Tabel 3.6. Data Setelah Binarisasi

Glucose	Body Mass index	Age	Outcome
1	1	1	1
1	1	1	1
0	0	0	0
0	0	0	0

2. Menghitung Bernoulli dengan *Laplace Smoothing*

$$P(x'_i = 1|C_k) = \frac{N_{ik} + 1}{N_k + 2}, \quad P(x'_i = 0|C_k) = \frac{N_k - N_{ik} + 1}{N_k + 2} \quad (3.2)$$

Kelas 1 :

$$P(\text{Glucose} = 1|C = 1) = \frac{2 + 1}{2 + 2} = 0.75$$

$$P(BMI = 1|C = 1) = \frac{2 + 1}{2 + 2} = 0.75$$

$$P(Age = 1|C = 1) = \frac{2 + 1}{2 + 2} = 0.75$$

$$P(C = 1 | X) = 0.5 \cdot 0.75^3 = 0.2109$$

Kelas 0 :

$$P(Glucose = 0|C = 0) = \frac{0 + 1}{2 + 2} = 0.25$$

$$P(BMI = 0|C = 0) = \frac{0 + 1}{2 + 2} = 0.25$$

$$P(Age = 0|C = 0) = \frac{0 + 1}{2 + 2} = 0.25$$

$$P(C = 0 | X) = 0.5 \cdot 0.25^3 = 0.0078$$

Sehingga, data dapat diklasifikasikan kedalam **Kelas 1** (Diabetes).

3.3.3. Perhitungan Manual untuk MNB

Metode *Multinomial Naïve Bayes* biasanya digunakan untuk data berupa jumlah atau frekuensi (seperti teks). Dalam kasus ini, fitur numerik digunakan langsung sebagai representasi jumlah, dan probabilitas dihitung berdasarkan proporsi setiap fitur dalam total nilai per kelas. Berikut adalah langkah-langkah perhitungan manualnya :

1. Menghitung total nilai per-kelas:

Tabel 3.7. Data Total Nilai per-Kelas

Fitur	$\Sigma(C=1)$	$\Sigma(C=0)$
Glucose	280	190

BMI	58	49
Age	95	55
Total	433	294

2. Menghitung probabilitas dengan *Laplace Smoothing*

$$P(x_i|C_k) = \frac{f_{ik} + 1}{\sum_{j=1}^n f_{jk} + n} \quad (3.3)$$

Kelas 1 :

$$P(Glucose = 120|C = 1) = \frac{120 + 1}{433 + 3} = \frac{121}{436}$$

$$P(BMI = 27|C = 1) = \frac{27 + 1}{433 + 3} = \frac{28}{436}$$

$$P(Age = 40|C = 1) = \frac{40 + 1}{433 + 3} = \frac{41}{436}$$

$$P(C = 1|X) \approx 0.5 \cdot \frac{121}{436} \cdot \frac{28}{436} \cdot \frac{41}{436} \approx 0.00084$$

Kelas 0 :

$$P(Glucose = 120|C = 0) = \frac{120 + 1}{294 + 3} = \frac{121}{297}$$

$$P(BMI = 27|C = 0) = \frac{27 + 1}{294 + 3} = \frac{28}{297}$$

$$P(Age = 40|C = 0) = \frac{40 + 1}{294 + 3} = \frac{41}{297}$$

$$P(C = 0|X) \approx 0.5 \cdot \frac{121}{294} \cdot \frac{28}{294} \cdot \frac{41}{294} \approx 0.00265$$

Sehingga, data dapat diklasifikasikan kedalam **Kelas 0** (non-diabetes).

Berdasarkan hasil perhitungan manual, *Gaussian Naïve Bayes* dan *Bernoulli Naïve Bayes* mengklasifikasikan bahwa data uji termasuk ke dalam kelas 1 (diabetes), karena nilai probabilitas posteriornya lebih besar dibanding kelas 0. Sedangkan untuk *Multinomial Naïve Bayes* justru mengklasifikasikan kelas 0 (non-diabetes).



DAFTAR PUSTAKA

- A'an Prasetyo, M., Khanif Zyen, A., & Kusumodestoni, R. H. (2024). Optimasi Algoritma Naive Bayes Berbasis Kernel Untuk Klasifikasi Penyakit Hati. *Jurnal Informatika Teknologi Dan Sains*, 6;3, 748–756. <https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset?resource>.
- Ahmed, A. B., & Haruna, K. (2025). Enhanced Sms Spam Detection Using Bernoulli Naive Bayes With Tf-Idf. *Fudma Journal Of Sciences*, 9(1), 393–399. <https://doi.org/10.33003/fjs-2025-0901-3226>
- Aminuddin, A., Sima, Y., Cholifatul Izza, N., Syamsi Norma Lalla, N., Arda, D., Kemenkes Palu, P., Amanah Makassar, S., Karya Husada, P., & Sandi Karsa, P. (2023). Edukasi Kesehatan Tentang Penyakit Diabetes Melitus Bagi Masyarakat. *Abdimas Polsaka: Jurnal Pengabdian Kepada Masyarakat*, 2(1), 7–12. <https://doi.org/10.35816/abdimaspolsaka>
- Ampomah, E. K., Nyame, G., Qin, Z., Addo, P. C., Gyamfi, E. O., & Gyan, M. (2021). Stock Market Prediction With Gaussian Naïve Bayes Machine Learning Algorithm. *Informatica (Slovenia)*, 45(2), 243–256. <https://doi.org/10.31449/inf.v45i2.3407>
- Ansari, G. A., Shafi, S., Ansari, M. D., & Shadab, A. (2025). Advanced Supervised Machine Learning Methods For Precise Diabetes Mellitus Prediction Using Feature Selection. *Frontiers In Medicine*, 12. <https://doi.org/10.3389/fmed.2025.1620268>
- As'ad, I. (2023). Advancing Healthcare Diagnostics: A Study On Gaussian Naive Bayes Classification Of Blood Samples. *International Journal Of Artificial*

Intelligence In Medical Issues, 1(2), 115–123.

<https://doi.org/10.56705/Ijaimi.V1i2.120>

Azizah, H., Rintyarna, B. S., & Cahyanto, T. A. (2022). Sentimen Analisis Untuk Mengukur Kepercayaan Masyarakat Terhadap Pengadaan Vaksin Covid-19 Berbasis Bernoulli Naive Bayes. *BIOS: Jurnal Teknologi Informasi Dan Rekayasa Komputer*, 3(1), 23–29. <https://doi.org/10.37148/Bios.V3i1.36>

Berliana, E. V., & Riasetiawan, M. (2024). Comparative Analysis Of Naïve Bayes Classifier, Support Vector Machine And Decision Tree In Rainfall Classification Using Confusion Matrix. *IJACSA) International Journal Of Advanced Computer Science And Applications*, 15(7), 2024. www.ijacsa.thesai.org

Blanquero, R., Carrizosa, E., Ramírez-Cobo, P., & Sillero-Denamiel, M. R. (2021). Variable Selection For Naïve Bayes Classification. *Computers And Operations Research*, 135. <https://doi.org/10.1016/J.Cor.2021.105456>

Chen, H., Hu, S., Hua, R., & Zhao, X. (2021). Improved Naive Bayes Classification Algorithm For Traffic Risk Management. *Eurasip Journal On Advances In Signal Processing*, 2021(1). <https://doi.org/10.1186/S13634-021-00742-6>

Dewi, C., Chen, R. C., Christanto, H. J., & Cauteruccio, F. (2023). Multinomial Naïve Bayes Classifier For Sentiment Analysis Of Internet Movie Database. *Vietnam Journal Of Computer Science*, 10(4), 485–498. <https://doi.org/10.1142/S2196888823500100>

Dhuhita, W. M. P., & Zone, F. (2023). Perbandingan Kinerja Algoritma Multinomial Dan Bernoulli Naïve Bayes Dalam Mengklasifikasikan Komentar

- Cyberbullying. *Komputika : Jurnal Sistem Komputer*, 12(2), 109–117.
<https://doi.org/10.34010/Komputika.V12i2.9767>
- Dirgantara, W., Iqbal Maulana, F., Subairi, S., & Arifuddin, R. (2024). The Performance Of Machine Learning Model Bernoulli Naïve Bayes, Support Vector Machine, And Logistic Regression On COVID-19 In Indonesia Using Sentiment Analysis. *Techné Jurnal Ilmiah Elektroteknika*, 23;1, 153–162.
- Elistiana, K. M., Bagus Adhi Kusuma, Subarkah, P., & Awal Rozaq, H. A. (2023). Improvement Of Naive Bayes Algorithm In Sentiment Analysis Of Shopee Application Reviews On Google Play Store. *Jurnal Teknik Informatika (Jutif)*, 4(6), 1431–1436.
- Gan, S., Shao, S., Chen, L., Yu, L., & Jiang, L. (2021). Adapting Hidden Naive Bayes For Text Classification. *Mathematics*, 9(19).
<https://doi.org/10.3390/Math9192378>
- Hassan, S. U., Ahamed, J., & Ahmad, K. (2022). Analytics Of Machine Learning-Based Algorithms For Text Classification. *Sustainable Operations And Computers*, 3, 238–248.
- Ige, T., & Kiekintveld, C. (2023). *Performance Comparison And Implementation Of Bayesian Variants For Network Intrusion Detection*.
- Jha, R., Lopez-Trevino, S., Kankanamalage, H. R., & Jha, J. C. (2024). Diabetes And Renal Complications: An Overview On Pathophysiology, Biomarkers And Therapeutic Interventions. *Biomedicines*, 12(5).
<https://doi.org/10.3390/Biomedicines12051098>
- Lissentiya Armal, H., Annisa, I., Nurcahya, H., & Wiryanti, I. (2023). Korelasi Glukosa Darah Puasa Pasien Diabetes Mellitus Tipe 2 Yang Mengalami

- Obesitas Terhadap Kadar Amilase, Insulin Dan Indeks Massa Tubuh. *Anakes: Jurnal Ilmiah Analisis Kesehatan*, 9(1), 70–78.
[Http://Journal.Thamrin.Ac.Id/Index.Php/Anakes/Issue/View/83](http://Journal.Thamrin.Ac.Id/Index.Php/Anakes/Issue/View/83)
- Markoulidakis, I., & Markoulidakis, G. (2024). Probabilistic Confusion Matrix: A Novel Method For Machine Learning Algorithm Generalized Performance Analysis. *Technologies*, 12(7).
- Miller, C., Portlock, T., Nyaga, D. M., & O’Sullivan, J. M. (2024). A Review Of Model Evaluation Metrics For Machine Learning In Genetics And Genomics. *Frontiers In Bioinformatics*, 4. <https://doi.org/10.3389/fbinf.2024.1457619>
- Mulyani, A., Kurniadi, D., Nashrulloh, M. R., Julianto, I. T., & Regita, M. (2022). The Prediction Of Ppa And Kip-Kuliah Scholarship Recipients Using Naive Bayes Algorithm. *Jurnal Teknik Informatika (Jutif)*, 3(4), 821–827.
<https://doi.org/10.20884/1.Jutif.2022.3.4.297>
- Okwudili, U. H., Ukachukwu, O. O., Chijindu, V. C., Ezea, M. O., & Ishaq, B. (2025). An Improved Performance Model For Artificial Intelligence-Based Diabetes Prediction. *Journal Of Electrical Systems And Information Technology*, 12(1). <https://doi.org/10.1186/S43067-025-00224-X>
- Ong, K. L., Stafford, L. K., Mclaughlin, S. A., Boyko, E. J., Vollset, S. E., Smith, A. E., Dalton, B. E., Duprey, J., Cruz, J. A., Hagins, H., Lindstedt, P. A., Aali, A., Abate, Y. H., Abate, M. D., Abbasian, M., Abbasi-Kangevari, Z., Abbasi-Kangevari, M., Abd ElHafeez, S., Abd-Rabu, R., ... Vos, T. (2023). Global, regional, and national burden of diabetes from 1990 to 2021, with projections of prevalence to 2050: a systematic analysis for the Global Burden of Disease

- Study 2021. *The Lancet*, 402(10397), 203–234.
[https://doi.org/https://doi.org/10.1016/S0140-6736\(23\)01301-6](https://doi.org/https://doi.org/10.1016/S0140-6736(23)01301-6)
- Pebdika, A., Herdiana, R., & Solihudin, D. (2023). Klasifikasi Menggunakan Metode Naive Bayes Untuk Menentukan Calon Penerima Pip. In *Jurnal Mahasiswa Teknik Informatika* (Vol. 7, Issue 1).
- Putra, A. W., Kosumo, K., Sitho, A., Ratu, R., Mujayanto, R. R., Rafly, M., Mintarum, M. M., & Nurcahyawati, V. (2024). Perbandingan Klasifikasi antara Naives Bayes dan Decision Tree dalam Prediksi Penyakit Diabetes Tahap Awal. *Jurnal Ilmu Komputer*, 17(1), 46–54.
- Rainio, O., Teuho, J., & Klén, R. (2024). Evaluation metrics and statistical tests for machine learning. *Scientific Reports*, 14(1). <https://doi.org/10.1038/s41598-024-56706-x>
- Ramadhon, R. N., Ogi, A., Agung, A. P., Putra, R., Febrihartina, S. S., & Firdaus, U. (2024). Implementasi Algoritma Decision Tree untuk Klasifikasi Pelanggan Aktif atau Tidak Aktif pada Data Bank. *Karimah Tauhid*, 3(2), 1860–1874.
- Sheth, V., Tripathi, U., & Sharma, A. (2022). A Comparative Analysis of Machine Learning Algorithms for Classification Purpose. *Procedia Computer Science*, 215, 422–431. <https://doi.org/10.1016/j.procs.2022.12.044>
- Simanjuntak, T. I., Muhathir, M., Fadlisyah, F., & Safira, I. (2023). Performance Analysis of Naive Bayes Variation Method in Spice Image Classification Using Histogram of Gradient Oriented (HOG) Feature Extraction. *Journal Of Informatics And Telecommunication Engineering*, 7(1), 282–291.
<https://doi.org/10.31289/jite.v7i1.7957>

- Taslim, T., Handayani, S., & Fajrizal, F. (2023). Kinerja Komparatif Optimasi Algoritma Naive Bayes dalam Klasifikasi Teks untuk Uji Klinis Kanker. *Jurnal Eksplora Informatika*, 13(1), 113–123. <https://doi.org/10.30864/eksplora.v13i1.994>
- Wafa, H. S., Hadiana, A. I., & Umbara, F. R. (2022). Prediksi Penyakit Diabetes Menggunakan Algoritma Support Vector Machine (SVM) Informasi Artikel Abstrak. *Informatics And Digital Expert (INDEX)*, 4(1), 40–45. <https://e-journal.unper.ac.id/index.php/informatics>
- Wang, S., Ren, J., & Bai, R. (2023). A Semi-Supervised Adaptive Discriminative Discretization Method Improving Discrimination Power of Regularized Naive Bayes. <https://doi.org/10.1016/j.eswa.2023.120094>
- Yolanda Paramitha, N., Nuryaman, A., Faisol, A., Setiawan, E., Dina Eka Nurvazly, dan, Matematika, J., Mipa, F., Lampung Jl Soemantri Brojonegoro No, U., Lampung, B., Kunci, K., & Bayes, N. (2023). Klasifikasi Penyakit Stroke Menggunakan Metode Naïve Bayes. In *Jurnal Siger Matematika* (Vol. 04, Issue 01).
- Yusran, M., Siswanto, S., & Islamiyati, A. (2024). SISTEMASI: Jurnal Sistem Informasi Comparison of Multinomial Naïve Bayes and Bernoulli Naïve Bayes on Sentiment Analysis of Kurikulum Merdeka with Query Expansion Ranking. *SISTEMASI: Jurnal Sistem Informasi*, 13;1, 96–106. <http://sistemasi.ftik.unisi.ac.id>
- Yuyun, Nurul Hidayah, & Supriadi Sahibu. (2021). Algoritma Multinomial Naïve Bayes Untuk Klasifikasi Sentimen Pemerintah Terhadap Penanganan Covid-

- 19 Menggunakan Data Twitter. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, 5(4), 820–826. <https://doi.org/10.29207/resti.v5i4.3146>
- Zulfikar, W. B., Atmadja, A. R., & Pratama, S. F. (2023). Sentiment Analysis on Social Media Against Public Policy Using Multinomial Naive Bayes. *Scientific Journal of Informatics*, 10(1), 25–34. <https://doi.org/10.15294/sji.v10i1.39952>



LAMPIRAN

1. Kode Program

```
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score,
classification_report, confusion_matrix
import matplotlib.pyplot as plt
import seaborn as sns

# dataset
file_path = "C:/Users/nasut/Pima_diabetes_normalized.xlsx"
df = pd.read_excel(file_path)

X = df.drop(columns='Outcome')
y = df['Outcome']
X_discrete = X.apply(lambda col: pd.cut(col, bins=5,
labels=False))

X_train, X_test, y_train, y_test = train_test_split(
    X_discrete, y, test_size=0.1, random_state=42, stratify=y
)

class NaiveBayesOriginal:
    def fit(self, X, y):
        self.classes = np.unique(y)
        self.class_priors = {c: np.mean(y == c) for c in
self.classes}

        self.likelihoods = {}
        for c in self.classes:
            X_c = X[y == c]
            self.likelihoods[c] = {}
            for col in X.columns:
                value_counts =
X_c[col].value_counts(normalize=True)
                self.likelihoods[c][col] =
value_counts.to_dict()

    def predict_row(self, row):
        posteriors = {}
        for c in self.classes:
            prior = np.log(self.class_priors[c])
            likelihood = 0
            for col, val in row.items():
                prob = self.likelihoods[c][col].get(val, 1e-
6) # Laplace smoothing
                likelihood += np.log(prob)
            posteriors[c] = prior + likelihood
        return max(posteriors, key=posteriors.get)

    def predict(self, X):
        return X.apply(self.predict_row, axis=1)

model_ori = NaiveBayesOriginal()
```

```

model_ori.fit(X_train, y_train)
y_pred = model_ori.predict(X_test)

acc = accuracy_score(y_test, y_pred)
print(f"\nNaiveBayes (manual) - Akurasi: {acc:.4f}")
print(classification_report(y_test, y_pred,
target_names=["Negatif", "Positif"]))

cm = confusion_matrix(y_test, y_pred)
print(f"Confusion Matrix NaiveBayes (angka):\n{cm}\n")

plt.figure(figsize=(5, 4))
sns.heatmap(cm, annot=True, fmt='d', cmap='Blues',
            xticklabels=["Negatif", "Positif"],
            yticklabels=["Negatif", "Positif"])
plt.title("Confusion Matrix NaiveBayes (80:20)")
plt.xlabel("Prediksi")
plt.ylabel("Aktual")
plt.tight_layout()
plt.show()

import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score,
classification_report, confusion_matrix
import matplotlib.pyplot as plt
import seaborn as sns

# dataset
file_path = "C:/Users/nasut/Pima_diabetes_normalized.xlsx"
df = pd.read_excel(file_path)

# fitur dan label
X = df.drop(columns='Outcome')
y = df['Outcome']

# diskretisasi fitur (5 bins)
X_discrete = X.apply(lambda col: pd.cut(col, bins=5,
labels=False))

# split data
X_train, X_test, y_train, y_test = train_test_split(
    X_discrete, y, test_size=0.2, random_state=42, stratify=y
)

class NaiveBayesOriginal:
    def fit(self, X, y):
        self.classes = np.unique(y)
        self.class_priors = {c: np.mean(y == c) for c in
self.classes}

        self.likelihoods = {}
        for c in self.classes:
            X_c = X[y == c]
            self.likelihoods[c] = {}
            for col in X.columns:

```

```

        value_counts =
X_c[col].value_counts(normalize=True)
        self.likelihoods[c][col] =
value_counts.to_dict()

    def predict_row(self, row):
        posteriors = {}
        for c in self.classes:
            prior = np.log(self.class_priors[c])
            likelihood = 0
            for col, val in row.items():
                prob = self.likelihoods[c][col].get(val, 1e-
6) # Laplace smoothing
                likelihood += np.log(prob)
            posteriors[c] = prior + likelihood
        return max(posteriors, key=posteriors.get)

    def predict(self, X):
        return X.apply(self.predict_row, axis=1)

model_ori = NaiveBayesOriginal()
model_ori.fit(X_train, y_train)
y_pred = model_ori.predict(X_test)

acc = accuracy_score(y_test, y_pred)
print(f"\nNaiveBayes (manual) - Akurasi: {acc:.4f}")
print(classification_report(y_test, y_pred,
target_names=["Negatif", "Positif"]))

# Confusion Matrix
cm = confusion_matrix(y_test, y_pred)
print(f"Confusion Matrix NaiveBayes (angka):\n{cm}\n")

# Visualisasi
plt.figure(figsize=(5, 4))
sns.heatmap(cm, annot=True, fmt='d', cmap='Blues',
            xticklabels=["Negatif", "Positif"],
            yticklabels=["Negatif", "Positif"])
plt.title("Confusion Matrix NaiveBayes (90:10)")
plt.xlabel("Prediksi")
plt.ylabel("Aktual")
plt.tight_layout()
plt.show()

import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score,
classification_report, confusion_matrix
import matplotlib.pyplot as plt
import seaborn as sns

# dataset
file_path = "C:/Users/nasut/Pima_diabetes_normalized.xlsx"
df = pd.read_excel(file_path)

# fitur dan label
X = df.drop(columns='Outcome')
y = df['Outcome']

```

```
# diskretisasi fitur (5 bins)
X_discrete = X.apply(lambda col: pd.cut(col, bins=5,
labels=False))

# split data
X_train, X_test, y_train, y_test = train_test_split(
    X_discrete, y, test_size=0.3, random_state=42, stratify=y
)

class NaiveBayesOriginal:
    def fit(self, X, y):
        self.classes = np.unique(y)
        self.class_priors = {c: np.mean(y == c) for c in
self.classes}

        self.likelihoods = {}
        for c in self.classes:
            X_c = X[y == c]
            self.likelihoods[c] = {}
            for col in X.columns:
                value_counts =
X_c[col].value_counts(normalize=True)
                self.likelihoods[c][col] =
value_counts.to_dict()

        def predict_row(self, row):
            posteriors = {}
            for c in self.classes:
                prior = np.log(self.class_priors[c])
                likelihood = 0
                for col, val in row.items():
                    prob = self.likelihoods[c][col].get(val, 1e-
6) # Laplace smoothing
                    likelihood += np.log(prob)
                posteriors[c] = prior + likelihood
            return max(posteriors, key=posteriors.get)

        def predict(self, X):
            return X.apply(self.predict_row, axis=1)

model_ori = NaiveBayesOriginal()
model_ori.fit(X_train, y_train)
y_pred = model_ori.predict(X_test)

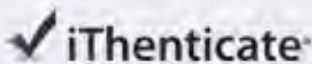
acc = accuracy_score(y_test, y_pred)
print(f"\nNaiveBayes (manual) - Akurasi: {acc:.4f}")
print(classification_report(y_test, y_pred,
target_names=["Negatif", "Positif"]))

# Confusion Matrix
cm = confusion_matrix(y_test, y_pred)
print(f"Confusion Matrix NaiveBayes (angka):\n{cm}\n")

# Visualisasi
plt.figure(figsize=(5, 4))
sns.heatmap(cm, annot=True, fmt='d', cmap='Blues',
            xticklabels=["Negatif", "Positif"],
            yticklabels=["Negatif", "Positif"])
```

```
plt.title("Confusion Matrix NaiveBayes (90:10)")
plt.xlabel("Prediksi")
plt.ylabel("Aktual")
plt.tight_layout()
plt.show()
```





Similarity Report ID: otd:29477:115668157

PAPER NAME

ILYAS NASUTION_Analisis Kinerja Metod
e Naive Bayes Dalam Mengklasifikasikan
Penyakit Diabetes_d6SLH0Sf5nKWf5Ch
VxsoMDC9U4kJl4Qq4L4yIK1p.pdf

AUTHOR

ILYAS NASUTION

WORD COUNT

11765 Words

CHARACTER COUNT

67635 Characters

PAGE COUNT

81 Pages

FILE SIZE

3.8MB

SUBMISSION DATE

Oct 6, 2025 3:03 PM GMT+7

REPORT DATE

Oct 6, 2025 3:05 PM GMT+7

19% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

19% Internet database

Crossref database

9% Publications database

Crossref Posted Content database

Excluded from Similarity Report

Submitted Works database

Cited material

Small Matches (Less than 15 words)

Bibliographic material

Abstract

UNIVERSITAS MEDAN AREA

© Hak Cipta Di Lindungi Undang-Undang

57

Document Accepted 22/12/25

1. Dilarang Mengutip sebagian atau seluruh dokumen ini tanpa mencantumkan sumber
2. Pengutipan hanya untuk keperluan pendidikan, penelitian dan penulisan karya ilmiah
3. Dilarang memperbanyak sebagian atau seluruh karya ini dalam bentuk apapun tanpa izin Universitas Medan Area

Access From (repositori.uma.ac.id)22/12/25



UNIVERSITAS MEDAN AREA

FAKULTAS TEKNIK

Kampus I : Jalan Kolam Nomor 1 Medan Estate / Jalan Gedung PBSI, Medan 20223
Kampus II : Jalan Sei Serayu Nomor 70 A / Jalan Setia Budi Nomor 79 B, Medan 20112 Telepon : (061) 8225602, 8201994
Fax : (061) 8226331 HP : 0811 607 259 website: www.uma.ac.id Email : univ_medanarea@uma.ac.id

Nomor : 600/FT/01.10/III/2025

10 Maret 2025

Lampiran : -

Hal : Pembimbing Tugas Akhir

Yth. Pembimbing Tugas Akhir
Dr Hartono S.Kom, M.Kom (Sebagai Pembimbing)
di Tempat

Dengan hormat, sehubungan telah dipenuhinya persyaratan untuk memperoleh Tugas Akhir dari mahasiswa atas :

Nama : ILYAS NASUTION
NIM : 218160039
Jurusan : TEKNIK INFORMATIKA

Maka dengan hormat kami mengharapkan kesediaan saudara :

Dr Hartono S.Kom, M.Kom (Sebagai Pembimbing)

Adapun Tugas Akhir Skripsi berjudul :

Analisis Kinerja Metode Naive Bayes Dalam Mengklasifikasikan Penyakit Diabetes

SK Pembimbing ini berlaku selama enam bulan terhitung sejak SK ini diterbitkan. Jika proses pembimbing melebihi batas waktu yang telah ditetapkan, SK ini dapat ditinjau ulang.

Demikian kami sampaikan, atas kesediaan saudara diucapkan terima kasih.

Dekan,



Dr Eng. Supriatno.ST, MT.





UNIVERSITAS MEDAN AREA

FAKULTAS TEKNIK

Kampus I : Jalan Kolam Nomor 1 Medan Estate ☎ (061) 7360168, Medan, 20223
Kampus II : Jalan Seliabudi Nomor 79 / Jalan Sei Serayu Nomor 70 A ☎ (061) 42402994, Medan, 20122
Website: www.teknik.uma.ac.id E-mail: univ_medanarea@uma.ac.id

Nomor
Tempat
Hal

: 191/FT.6/01.10/VII/2025
:-
: Penelitian Dan Pengambilan Data Tugas Akhir

1 Juli 2025

Yth. Wakil Rektor Bidang Mutu Sumber Daya Manusia dan Perekonomian
Jln. Kolam No.1
Di
Medan

Dengan hormat, kami mohon kesediaan bapak kiranya berkenan untuk memberikan izin dan kesempatan kepada mahasiswa kami tersebut dibawah ini :

NO	NAMA	NPM	PRODI
1	Ilyas Nasution	218160039	Teknik Informatika

Untuk melaksanakan Penelitian dan Pengambilan Data Tugas Akhir di **Laboratorium Komputer Program Studi Teknik Informatika Fakultas Teknik Universitas Medan Area**.

Perlu kami jelaskan bahwa Pengambilan Data tersebut adalah semata-mata untuk tujuan Ilmiah dan kriptisi, yang merupakan salah satu syarat bagi mahasiswa tersebut untuk mengikuti ujian sarjana pada Fakultas Teknik Universitas Medan Area dan tidak untuk dipublikasikan, dengan judul :

Analisis Kinerja Metode Naive Bayes dalam Mengklasifikasikan Penyakit Diabetes.

Mohon kiranya tanggal Surat Izin Pengambilan Data Tugas Akhir agar disesuaikan dengan tanggal terbitnya Surat ini.

Atas perhatian dan kerja sama yang baik diucapkan terima kasih.

Dekan


Dekan, ST, MT

Tembusan :

1. Ka. BPMPP
2. Mahasiswa
3. File



UNIVERSITAS MEDAN AREA

Kampus I : Jalan Kolam Nomor 1 Medan Estate ☎ (061) 7360168, 7366878, 7364348 ☎ (061) 7368012 Medan 20223
Kampus II : Jalan Setiabudi Nomor 79 / Jalan Sei Serayu Nomor 70 A ☎ (061) 8225602 ☎ (061) 8226331 Medan 20122
Website: www.uma.ac.id E-Mail: univ_medanarea@uma.ac.id

SURAT KETERANGAN
Nomor : 1320/UMA/B/01.7/VIII/2025

Rektor Universitas Medan Area dengan ini menerangkan bahwa :

Nama : Ilyas Nasution
No.Pokok Mahasiswa : 218160039
Fakultas : Teknik Informatika

Benar telah selesai Pengambilan Data di Fakultas Teknik Universitas Medan Area dengan
Judul Skripsi” Analisis Kinerja Metode Nave Bayes dalam mengklasifikasikan Penyakit
Diabetes ”.

Dan kami harapkan Data tersebut kiranya dapat membantu yang bersangkutan dalam
penyusunan skripsi dan dapat bermanfaat bagi mahasiswa khususnya Fakultas Teknik
Universitas Medan Area.

Demikian surat ini diterbitkan untuk dapat digunakan seperlunya.

Medan, 04 Agustus 2025.
an Rektor
Wakil Rektor Bidang Mutu SDM &
Perekonomian,

Dedi Sahputra, S.Sos, MA

CC :
- Arsip

