Journal of Artificial Intelligence and Software Engineering
Vol. 5, No. 3, September 2025, pp. 1174~1185
E-ISSN: 2777-001X, DOI: 10.30811/jaise.v5i3.7618

Feature Selection Optimization Using Genetic Algorithm for Naive Bayes-Based Diabetes Mellitus Classification

Nova Arianti Aris[1], Ade Yuliana[2*]
[1] Teknik Informatika, Politeknik TEDC Bandung, Kota Cimahi, 40513, Indonesia

Informasi Artikel

Diterima : 19 Agustus 2025
Revisi : 28 Agustus 2025
Publikasi : 30 September 2025

Kata Kunci:

Diabetes Melitus
Naive Bayes
Algoritma Genetika
Seleksi Fitur
Klasifikasi

ABSTRAK

Diabetes melitus merupakan salah satu penyakit kronis yang jumlah penderitanya terus meningkat setiap tahun dan berpotensi menimbulkan komplikasi serius apabila tidak ditangani sejak dini. Oleh karena itu, upaya deteksi dini risiko diabetes menjadi langkah krusial dalam pencegahan. Penelitian ini bertujuan untuk mengoptimalkan pemilihan fitur dengan memanfaatkan algoritma genetika pada proses klasifikasi pasien diabetes melitus menggunakan algoritma Naive Bayes. Algoritma genetika berfungsi untuk menyeleksi fitur klinis yang paling signifikan dari data pasien, sehingga diharapkan dapat meningkatkan akurasi sekaligus efisiensi model klasifikasi. Dataset yang digunakan terdiri atas 1.557 data pasien dengan 29 atribut klinis awal. Melalui tahapan persiapan dan pemilihan, diperoleh 7 fitur utama yang digunakan pada proses pelatihan model. Evaluasi dilakukan menggunakan metrik performa berupa accuracy, precision, recall, dan F1-score. Hasil eksperimen menunjukkan bahwa model dengan fitur terpilih mampu mencapai accuracy 80,99%, precision 80,99%, recall 100%, dan F1-score 89,5%. Temuan ini menegaskan bahwa penerapan algoritma genetika efektif dalam meningkatkan kinerja klasifikasi Naive Bayes untuk identifikasi risiko diabetes melitus. Hasil penelitian ini diharapkan dapat menjadi acuan dalam pengembangan sistem prediksi risiko penyakit yang lebih akurat dan efisien di masa mendatang.

ABSTRACT

Diabetes mellitus is a chronic disease with a steadily increasing prevalence each year and poses the risk of severe complications if not addressed early. Therefore, early detection of diabetes risk plays a vital role in prevention efforts. This study aims to enhance feature selection optimization through the use of a genetic algorithm in the classification of diabetes mellitus patients based on the Naive Bayes method. The genetic algorithm was applied to identify the most significant clinical features from patient data, with the expectation of improving the classification model's accuracy and efficiency. A dataset comprising 1,557 patient records with 29 initial clinical attributes was utilized. Following preparation and selection stages, 7 key features were chosen for model training. Model performance was evaluated using metrics such as accuracy, precision, recall, and F1-score. The results indicated that the model with selected features achieved an accuracy of 80.99%, precision of 80.99%, recall of 100%, and an F1-score of 89.5%. These findings confirm that genetic algorithms are effective in improving Naive Bayes classification performance for diabetes risk identification. This study is expected to serve as a foundation for the development of more accurate and efficient disease risk prediction systems in the future.

This is an open-access article under the CC BY-SA license

***Penulis Koresponden**

Email: yulianaad@poltektedc.ac.id

Cara sitasi IEEE:

N. A. Aris & A. Yuliana, "Feature Selection Optimization Using Genetic Algorithm for Naive Bayes-Based Diabetes Mellitus Classification," *Journal of Artificial Intelligence and Software Engineering (J-AISE)*, vol. 5, no. 3, pp. 1174-1185, September 2025, doi: 10.30811/jaise.v5i3.7618

1. PENDAHULUAN

Pada era digital saat ini, teknologi informasi memiliki peran yang sangat penting di berbagai bidang, termasuk sektor kesehatan. Kompleksitas serta besarnya volume data medis menuntut adanya metode analisis yang cerdas untuk menghasilkan informasi yang bernilai. Data mining hadir sebagai salah satu pendekatan yang digunakan untuk menggali pengetahuan tersembunyi dari data medis, khususnya dalam proses klasifikasi penyakit. Tantangan utama dalam pengolahan data medis adalah jumlah fitur (atribut) yang sangat banyak, yang apabila tidak dikelola dengan baik dapat menurunkan akurasi serta kinerja algoritma klasifikasi.

Dalam konteks klasifikasi penyakit, seleksi fitur merupakan tahapan penting guna meningkatkan akurasi model sekaligus menyederhanakan kompleksitas data. Pemilihan fitur yang tepat tidak hanya mempercepat proses komputasi, tetapi juga mampu meningkatkan kemampuan prediksi. Berbagai metode seleksi fitur telah dikembangkan, baik berbasis pendekatan statistik maupun algoritma evolusioner. Salah satu metode evolusioner yang cukup efektif adalah algoritma genetika (*Genetic Algorithm*/GA), yang mengadopsi mekanisme evolusi biologis untuk menemukan kombinasi fitur terbaik [1].

Di sisi lain, Diabetes Melitus (DM) termasuk salah satu penyakit kronis yang banyak dijumpai di seluruh dunia. Penyakit ini merupakan gangguan metabolik yang ditandai dengan meningkatnya kadar gula dalam darah (*hiperglikemia*), akibat kelainan pada produksi insulin, fungsi insulin, atau kombinasi keduanya [2]. Berdasarkan laporan *World Health Organization* (WHO), terdapat lebih dari 422 juta penderita diabetes secara global, dan jumlah tersebut diperkirakan akan terus meningkat. Kondisi ini menegaskan perlunya pengembangan sistem klasifikasi yang efektif untuk mendukung diagnosis dini dan pengambilan keputusan klinis yang tepat. Selain itu, klasifikasi yang akurat juga berperan dalam pemetaan risiko, penentuan strategi penanganan, hingga pemberian edukasi yang sesuai bagi masyarakat [3].

Algoritma *Naive Bayes* merupakan salah satu teknik klasifikasi yang populer karena sederhana dan memiliki kecepatan komputasi yang tinggi [4]. Namun, algoritma ini memiliki keterbatasan jika dihadapkan pada dataset dengan banyak fitur yang tidak relevan atau saling berkorelasi. Untuk mengatasi kendala tersebut, integrasi antara *Naive Bayes* dengan metode seleksi fitur seperti algoritma genetika dapat menghasilkan model klasifikasi yang lebih baik [5].

Sejumlah studi terdahulu telah mencoba mengombinasikan algoritma genetika dengan metode klasifikasi. Penelitian Somantri, Maharrani, dan Wanti [6] misalnya, berfokus pada optimasi bobot *Naive Bayes* dalam deteksi dini diabetes, dan melaporkan adanya peningkatan akurasi. Namun, penelitian tersebut masih terbatas pada variabel tertentu dan belum mengeksplorasi fitur yang lebih kompleks dari data medis pasien. Sementara itu, Yuliana dan Devianti [7] menggunakan algoritma Apriori untuk menganalisis pola gejala diabetes, dan berhasil mengidentifikasi kombinasi gejala dominan sebagai indikator awal DM. Pendekatan tersebut efektif untuk menemukan pola asosiatif, tetapi belum mengoptimalkan performa model klasifikasi secara langsung. Dengan demikian, meskipun penelitian-penelitian sebelumnya memberikan kontribusi penting, fokus mereka lebih pada optimasi parsial atau analisis pola, bukan pada integrasi seleksi fitur berbasis algoritma genetika dengan *Naive Bayes* untuk meningkatkan klasifikasi diabetes secara menyeluruh.

Berbeda dari penelitian-penelitian terdahulu, studi ini menitikberatkan pada optimasi seleksi fitur menggunakan algoritma genetika yang diterapkan langsung pada klasifikasi diabetes berbasis *Naive Bayes*. Penelitian ini menggunakan data pasien yang lebih komprehensif, mencakup identitas, riwayat penyakit, faktor risiko, pemeriksaan fisik, kadar gula darah, diagnosis, hingga tindakan lanjutan berupa terapi maupun edukasi kesehatan. Dengan ruang lingkup data yang lebih luas, diharapkan hasil penelitian ini mampu memberikan kontribusi baru dalam meningkatkan akurasi diagnosis sekaligus memperkuat keandalan sistem pendukung keputusan di bidang kesehatan.

Dataset yang dipakai berasal dari Puskesmas Cigugur Tengah dalam periode Januari 2024 hingga April 2025, dikumpulkan melalui metode observasi terhadap perilaku, aktivitas, serta kondisi pasien [8]. Tujuan akhirnya adalah meningkatkan akurasi diagnosis diabetes serta menyediakan solusi yang lebih efisien untuk praktik medis.

Dengan mengombinasikan kemampuan *Naive Bayes* dalam klasifikasi serta kekuatan algoritma genetika dalam eksplorasi ruang solusi, penelitian ini diharapkan dapat memberikan kontribusi nyata dalam

pengembangan sistem pendukung keputusan di bidang kesehatan. Selain itu, pendekatan ini juga dapat diaplikasikan pada kasus penyakit lain, sehingga bermanfaat dalam mengevaluasi kehandalan dan generalisasi metode yang digunakan.

Berdasarkan latar belakang tersebut, penelitian ini mengusung judul "Optimasi Seleksi Fitur Menggunakan Algoritma Genetika pada Klasifikasi Diabetes Melitus Berbasis *Naive Bayes*." Judul ini dipilih karena relevan dengan kebutuhan akan sistem deteksi dini diabetes yang akurat, efisien, serta adaptif dalam mengelola kompleksitas data medis.

2. METODE

Metode penelitian ini terdiri dari lima tahap utama, yaitu: penyusunan *dataset*, tahap preparasi data, penerapan algoritma genetika, penerapan algoritma *Naive Bayes*, serta tahap hasil. Alur keseluruhan penelitian digambarkan secara visual pada Gambar 1.

Gambar 1. Proses Pengolahan Data Penelitian

2.1. *Dataset*

Dataset adalah sekumpulan informasi yang terorganisir dalam format terstruktur, umumnya berbentuk tabel dengan baris sebagai representasi entri atau observasi dan kolom sebagai representasi variabel atau fitur [8]. Data yang digunakan dalam penelitian ini adalah data pasien diabetes melitus yang diperoleh secara langsung dari Puskesmas Cigugur Tengah. Jumlah keseluruhan dataset mencapai 1.557 pasien, yang dikumpulkan dalam periode Januari 2024 hingga April 2025.

Data medis pasien mencakup berbagai informasi, mulai dari data demografis, riwayat penyakit, gaya hidup, hasil pemeriksaan fisik, hingga diagnosis medis. Secara keseluruhan terdapat 29 atribut awal yang digunakan, meliputi identitas pasien, kebiasaan merokok, aktivitas fisik, pola konsumsi (gula, garam, lemak, buah, sayur, alkohol), tekanan darah (sistol dan diastol), tinggi badan, berat badan, indeks massa tubuh (IMT), lingkar perut, kadar gula darah, hingga indikasi gangguan indera seperti penglihatan dan pendengaran. Selain itu, kondisi medis seperti katarak, kelainan refraksi, congek/OMSK, serta hasil diagnosis juga termasuk di dalamnya. Rincian dari 29 atribut ini ditampilkan pada Tabel 1.

Tabel 1. Variabel pada *Dataset* Pasien Diabetes Melitus

No.	Variabel	Keterangan
1.	Nama Pasien	Nama lengkap pasien yang diperiksa.
2.	Jenis Kelamin	Jenis kelamin pasien (Laki-laki/Perempuan).
3.	Golongan Darah	Golongan darah pasien (A, B, AB, O).
4.	Merokok	Kebiasaan merokok pasien (Ya/Tidak).
5.	Kurang Aktivitas Fisik	Pasien tidak melakukan aktivitas fisik yang cukup (Ya/Tidak).
6.	Gula Berlebihan	Konsumsi gula melebihi batas normal (Ya/Tidak).
7.	Garam Berlebihan	Konsumsi garam melebihi batas normal (Ya/Tidak).
8.	Lemak Berlebihan	Konsumsi lemak melebihi batas normal (Ya/Tidak).
9.	Kurang Makan Buah dan Sayur	Asupan buah dan sayur tidak cukup (Ya/Tidak).
10.	Konsumsi Alkohol	Kebiasaan mengonsumsi alkohol (Ya/Tidak).
11.	Sistol	Tekanan darah sistolik (angka atas), satuan mmHg.
12.	Diastol	Tekanan darah diastolik (angka bawah), satuan mmHg.
13.	Tinggi Badan (cm)	Tinggi badan pasien dalam satuan sentimeter.
14.	Berat Badan (kg)	Berat badan pasien dalam satuan kilogram.
15.	IMT	Indeks Massa Tubuh, dihitung dari berat dan tinggi badan.
16.	Lingkar Perut (cm)	Ukuran lingkar perut pasien dalam sentimeter.
17.	Pemeriksaan Gula	Hasil pemeriksaan kadar gula darah pasien (angka atau kategori).
18.	Rujuk RS	Status apakah pasien dirujuk ke rumah sakit (Ya/Tidak).
19.	Katarak Mata Kanan	Kondisi katarak pada mata kanan (Ya/Tidak).
20.	Katarak Mata Kiri	Kondisi katarak pada mata kiri (Ya/Tidak).
21.	Kelainan Refraksi Mata Kanan	Gangguan pembiasan cahaya pada mata kanan (Miopi, Hipermetropi, dll.).

22.	Kelainan Refraksi Mata Kiri	Gangguan pembiasan cahaya pada mata kiri.
23.	Curiga Tuli Kongenital Telinga Kanan	Kecurigaan tuli bawaan sejak lahir di telinga kanan (Ya/Tidak).
24.	Curiga Tuli Kongenital Telinga Kiri	Kecurigaan tuli bawaan sejak lahir di telinga kiri (Ya/Tidak).
25.	OMSK/Congek Telinga Kanan	Infeksi telinga tengah kronik (congek) di telinga kanan (Ya/Tidak).
26.	OMSK/Congek Telinga Kiri	Infeksi telinga tengah kronik (congek) di telinga kiri (Ya/Tidak).
27.	Serumen Telinga Kanan	Penumpukan kotoran (serumen) di telinga kanan (Ya/Tidak).
28.	Serumen Telinga Kiri	Penumpukan kotoran (serumen) di telinga kiri (Ya/Tidak).
29.	Diagnosis	Hasil diagnosis akhir atas kondisi kesehatan pasien.

2.2. Preparasi Data

Data yang telah dikumpulkan kemudian akan melalui tahap pra-pemrosesan atau pembersihan data. Tahap ini mencakup beberapa langkah, termasuk koreksi kesalahan atau inkonsistensi, transformasi data ke format yang sesuai, pengisian nilai yang hilang, serta normalisasi atau standarisasi untuk menjaga keseragaman [9]. Sehingga nanti akan menghasilkan data yang benar-benar siap untuk diproses.Berdasarkan 29 variabel pada dataset utama, maka akan dipilih hanya menjadi 17 variabel yang akan digunakan untuk optimasi fitur dalam klasifikasi pasien diabetes melitus. Optimasi didefinisikan sebagai metode penyelesaian masalah agar memperoleh hasil terbaik, baik dalam bentuk nilai minimum maupun maksimum, tergantung dari perspektif yang digunakan [10].

Dengan memusatkan analisis pada 17 (tujuh belas) variabel ini, maka proses optimasi fitur dalam klasifikasi pasien diabetes melitus dapat dilakukan lebih efisien dan menghasilkan fitur yang optimal serta data yang bersih dan variabel yang terpilih dengan tepat akan mencegah bias, mengurangi kompleksitas berlebihan, mempercepat komputasi, serta memastikan model lebih efisien dan akurat dalam menghasilkan fitur optimal untuk klasifikasi pasien diabetes melitus. Adapun 17 (tujuh belas) variabel tersebut yang sudah di preparasi data dapat dilihat pada tabel 2 berikut.

Tabel 2. Variabel pada *Dataset* Pasien Diabetes Melitus

No.	Variabel	Keterangan
1.	Jenis Kelamin	Jenis kelamin pasien (Laki-laki/Perempuan).
2.	Golongan Darah	Golongan darah pasien (A, B, AB, O).
3.	Merokok	Kebiasaan merokok pasien (Ya/Tidak).
4.	Kurang Aktivitas Fisik	Pasien tidak melakukan aktivitas fisik yang cukup (Ya/Tidak).
5.	Gula Berlebihan	Konsumsi gula melebihi batas normal (Ya/Tidak).
6.	Garam Berlebihan	Konsumsi garam melebihi batas normal (Ya/Tidak).
7.	Lemak Berlebihan	Konsumsi lemak melebihi batas normal (Ya/Tidak).
8.	Kurang Makan Buah dan Sayur	Asupan buah dan sayur tidak cukup (Ya/Tidak).
9.	Konsumsi Alkohol	Kebiasaan mengonsumsi alkohol (Ya/Tidak).
10.	Sistol	Tekanan darah sistolik (angka atas), satuan mmHg.
11.	Diastol	Tekanan darah diastolik (angka bawah), satuan mmHg.
12.	Tinggi Badan (cm)	Tinggi badan pasien dalam satuan sentimeter.
13.	Berat Badan (kg)	Berat badan pasien dalam satuan kilogram.
14.	IMT	Indeks Massa Tubuh, dihitung dari berat dan tinggi badan.
15.	Lingkar Perut (cm)	Ukuran lingkar perut pasien dalam sentimeter.
16.	Pemeriksaan Gula	Hasil pemeriksaan kadar gula darah pasien (angka atau kategori).
17.	Diagnosis	Hasil diagnosis akhir atas kondisi kesehatan pasien.

Untuk variabel jenis kelamin, golongan darah dan diagnosis akan diinisialisasi menjadi data *numeric* atau angka yang dimulai dari 0. Berikut inisialisasi variabel jenis kelamin, golongan darah dan diagnosis.

1. Inisialisasi Data Jenis Kelamin

Data jenis kelamin diinisialisasi agar mudah dalam memproses, dan hasil inisialisasi jenis kelamin ditunjukkan pada tabel 3 dibawah ini.

Tabel 3 Inisialisasi Jenis Kelamin

Jenis Kelamin	Inisialisasi
Laki-Laki	0
Perempuan	1

2. Inisialisasi Data Golongan Darah

Data golongan darah dibagi menjadi 4 yang terdiri atas A, B, AB dan O yang di inisialisasi menjadi 0, 1, 2 dan 3. Data hasil inisialisasinya dapat diketahui dari tabel 4 dibawah ini.

Tabel 4 Inisialisasi Golongan Darah

Golongan Darah	Inisialisasi
A	0
B	1
AB	2
O	3

3. Inisialisasi Data Diagnosis

Data diagnosis dikelompokkan menjadi dua kategori, yaitu diabetes tipe 1 dan tipe 2, yang diinisialisasi menjadi 0 dan 1. Data hasil inisialisasinya dapat diketahui pada tabel 5 dibawah ini.

Tabel 5 Inisialisasi Diagnosis

Jenis Kelamin	Inisialisasi
Diabetes Melitus Tipe 1	0
Diabetes Melitus Tipe 2	1

2.3. Proses Algoritma Genetika

Algoritma genetika merupakan metode utama dalam pengoptimalan algoritma *naïve bayes* karena memiliki kemampuan untuk memilih *subset* fitur yang paling relevan secara efisien dari sejumlah besar fitur yang tersedia. Hal ini dilakukan melalui mekanisme seleksi, *crossover*, dan mutasi yang meniru proses evolusi alam, sehingga dapat meningkatkan akurasi klasifikasi dan mengurangi kompleksitas model.

Berikut langkah-langkah beserta metode yang digunakan dalam proses algoritma genetika [11].

1. Seleksi: Proses seleksi menggunakan *Tournament Selection* untuk memilih induk bagi *crossover* dan mutasi. Dua individu dipilih secara acak, lalu yang memiliki *fitness* lebih tinggi dipilih sebagai induk baru.
2. *Crossover*: *Crossover* adalah proses penggabungan informasi genetik dari dua induk terpilih untuk membangkitkan populasi baru. Metode *Single Point Crossover* pada penelitian ini, dimana menentukan titik potong secara acak sebagai acuan penggabungan genetik, menghasilkan dua individu baru (*offspring*) dengan kombinasi parameter k dan fitur yang lebih bervariasi.
3. *Mutation*: Setelah *crossover*, *offspring* mengalami mutasi untuk mencegah stagnasi dan memastikan evolusi berlanjut. Mutasi terjadi dengan probabilitas kecil, ditentukan oleh *mutation rate* (mr). Penelitian ini menggunakan *Bit-Flip Mutation*, yang membalik nilai biner pada kromosom (0 menjadi 1, dan sebaliknya).

2.4. Proses Algoritma *Naive Bayes*

Metode *Naive Bayes* adalah salah satu model klasifikasi probalistik sederhana guna menghitung kumpulan probabilitas dan menjumlahkan frekuensi dan kombinasi dari *dataset*. *Naive Bayes* mempunyai beberapa keuntungan antara lain hanya membutuhkan jumlah *training* data yang kecil. Oleh karena itu, algoritma ini cocok untuk tugas seperti klasifikasi sentimen karena dapat secara efisien mengolah data teks dan memberikan hasil yang cukup akurat [12].

Training data berperan dalam penentuan estimasi parameter untuk klasifikasi, dengan persamaan teorema *Naive Bayes* dirumuskan sebagai [13]:

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

Di mana:

X : Data kelas yang belum diketahui.
H : Kelas spesifik yang berperan sebagai hipotesis pada data.
P(H|X) : Probabilitas posteriori, yaitu probabilitas hipotesis H dengan mempertimbangkan kondisi X.
P(H) : Probabilitas dari hipotesis H.
P(X|H) : Probabilitas kondisi X diberikan hipotesis H.
P(X) : Probabilitas dari kondisi X.

Lalu untuk perhitungan juga dilakukan dengan menggunakan rumus *F1-Score* untuk dapat mengetahui nilai *accuracy*, *precision*, dan *recall* [14]. Dapat dilihat pada tabel 6 berikut.

Tabel 6 Rumus *F1-Score*

Metode	Rumus
Accuracy	$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$
Precision	$Precision = \frac{TP}{TP+FP}$
Recall	$Recall = \frac{TP}{TP+FN}$
F1-Score	$F1\text{-}Score = 2 \text{ x } \frac{Precision \text{ x } Recall}{Precision+Recall}$

Keterangan:
True Positives (TP): Jumlah entri yang benar-benar positif dan diklasifikasikan sebagai positif.
False Positives (FP): Jumlah entri yang seharusnya negatif tetapi diklasifikasikan sebagai positif.
True Negatives (TN): Jumlah entri yang benar-benar negatif dan diklasifikasikan sebagai negatif.
False Negatives (FN): Jumlah entri yang seharusnya positif namun diklasifikasikan sebagai negatif.

2.5. Hasil

Hasil optimasi seleksi fitur menggunakan algoritma genetika pada klasifikasi Diabetes Melitus berbasis *Naive Bayes* menunjukkan bahwa metode ini dapat menghasilkan *subset* fitur yang lebih optimal dibandingkan tanpa proses seleksi fitur. Dengan penerapan seleksi turnamen, persilangan titik tunggal, serta mutasi *bit-flip,* algoritma genetika berhasil mengeksplorasi ruang solusi secara efektif sehingga diperoleh kombinasi fitur yang relevan dan mendukung peningkatan akurasi klasifikasi. *Subset* fitur terpilih mampu mempertahankan atau bahkan meningkatkan performa *Naive Bayes* sambil mengurangi kompleksitas model, karena fitur yang tidak relevan berhasil dieliminasi. Hal ini membuktikan bahwa algoritma genetika efektif sebagai metode optimasi seleksi fitur dalam sistem klasifikasi Diabetes Melitus.

3. HASIL DAN PEMBAHASAN

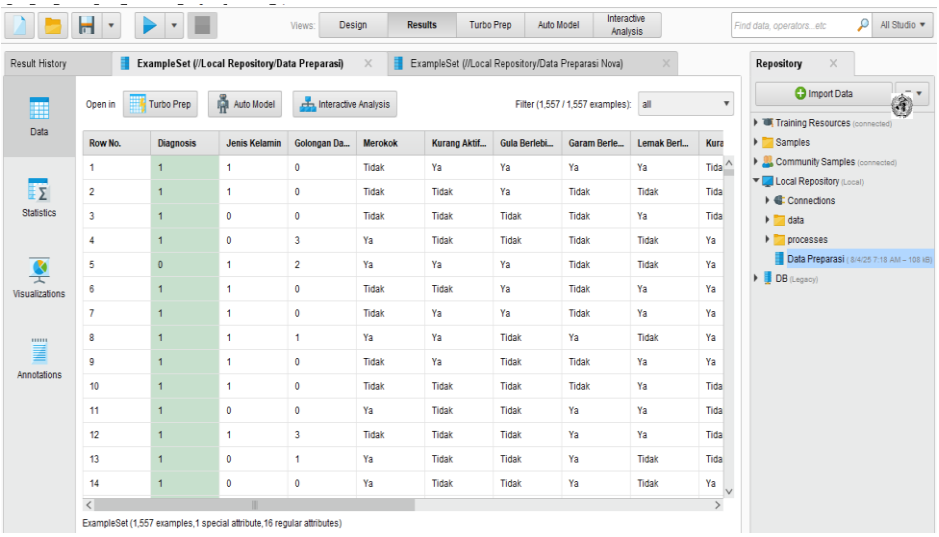
3.1. Hasil Penelitian

Hasil penelitian membahas proses pengolahan data awal, penerapan algoritma genetika untuk optimasi seleksi fitur, dan pengujian performa klasifikasi evaluasi akurasi klasifikasi *Naive Bayes* terhadap data Diabetes Melitus menggunakan teknik *cross validation* serta hasil optimasi seleksi fitur dan performa klasifikasi.

3.1.1. Proses Pengolahan Data Awal

Langkah awal pengelolaan data awal pada pengujian ini menggunakan perangkat lunak *RapidMiner*. Tahapan ini dilakukan dengan mengimpor data yang telah dipra-pemroses atau dipreparasi ke dalam perangkat lunak *RapidMiner*. Datanya berupa data rekam medis pasien penyakit diabetes melitus yang berjumlah 1557 data dari periode Januari 2024 - April 2025.

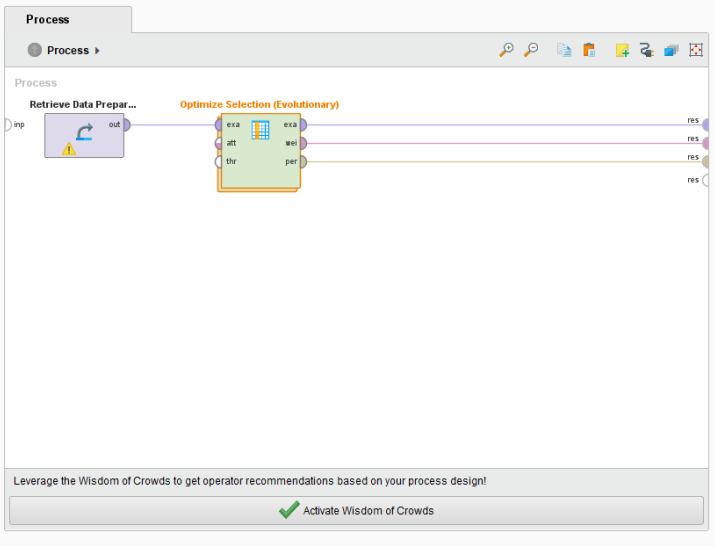
Berdasarkan data atribut yang sudah dipreparasi, maka selanjutnya akan disesuaikan formatnya agar dapat diolah pada *RapidMiner*, yaitu Jenis Kelamin, Merokok, Kurang Aktivitas Fisik, Gula Berlebihan, Garam Berlebihan, Lemak Berlebihan, Kurang Makan Buah dan Sayur, Konsumsi Alkohol serta Diagnosis dengan format *Binominal.* Lalu Golongan Darah dengan format *Polynominal*. Kemudian Sistol, Diastol, Lingkar Perut (cm) dan Pemeriksaan Gula dengan format *Integer*. Selanjutnya Tinggi Badan (cm), Berat Badan (kg) dan IMT dengan format *Real.* Data hasil preparasi yang telah di*import* pada *RapidMiner* dapat dilihat pada gambar 2 berikut



Gambar 2. *Dataset* Preparasi yang Berhasil Di*import*

3.1.2. Optimasi Seleksi Fitur Menggunakan Algoritma Genetika

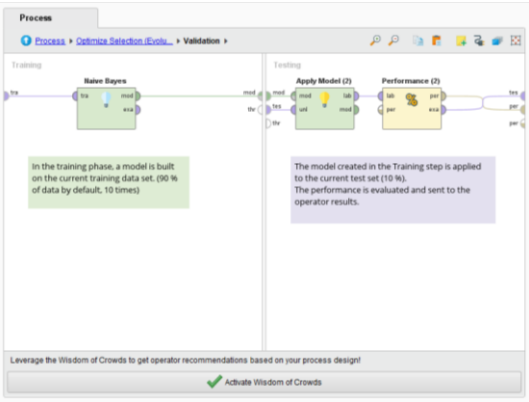
Setelah proses *impor* data, penulis melakukan pengujian menggunakan perangkat lunak *RapidMiner* untuk mengoptimasi seleksi fitur dari setiap atribut dengan algoritma genetika. Data hasil preparasi dimasukkan ke halaman *process*, kemudian ditambahkan operator *Optimize Selection (Evolutionary)* yang dihubungkan dengan data sebagai *input* dan *result* untuk *output*. Proses ini memungkinkan seleksi fitur dilakukan secara optimal menggunakan pendekatan algoritma genetika, sebagaimana terlihat pada gambar 3.



Gambar 3. Optimasi Seleksi Fitur Menggunakan Algoritma Genetika

3.1.3. Pengujian Performa Klasifikasi Menggunakan Algoritma Naïve Bayes

Proses selanjutnya adalah pengujian performa klasifikasi dengan algoritma *Naïve Bayes* menggunakan metode *k-fold cross validation* (k=10), di mana data dibagi acak menjadi k lipatan, dilatih pada k-1 lipatan, dan diuji pada lipatan tersisa, lalu hasilnya dirata-rata untuk mengurangi bias. Pengujian dilakukan dengan menempatkan operator *Cross Validation* di dalam *Optimize Selection (Evolutionary)*, menggunakan *Naïve Bayes* pada bagian *training* serta *Apply Model* dan *Performance* pada bagian *testing*. Operator dihubungkan secara berurutan dari *Naïve Bayes* ke *Apply Model*, lalu ke *Performance*, dan akhirnya ke *Result*, sebagaimana ditunjukkan pada gambar 4.



Gambar 4. Rangkaian Proses *Cross Validation*

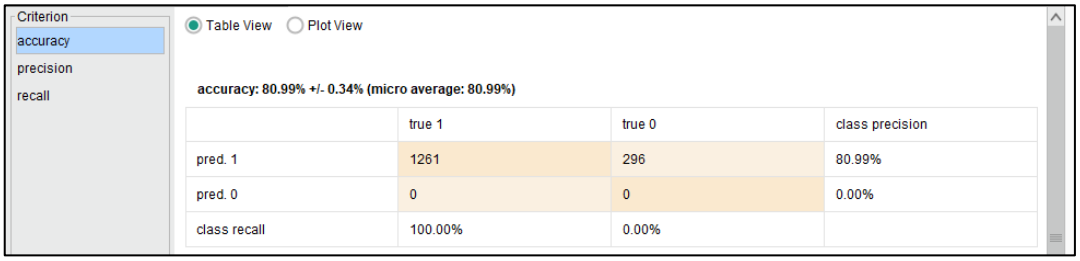
3.1.4. Hasil Optimasi Seleksi Fitur dan Performa Klasfikasi

Berdasarkan hasil optimasi seleksi fitur menggunakan algoritma genetika, diperoleh bahwa jumlah fitur optimal untuk klasifikasi penyakit diabetes melitus adalah tujuh fitur. Jumlah ini diperoleh dari 17 fitur awal yang telah melalui tahap preparasi data sebelum diuji menggunakan metode tersebut. Detail informasi mengenai fitur-fitur yang terpilih disajikan pada Tabel 7.

Tabel 7 Hasil Seleksi Fitur yang Optimal

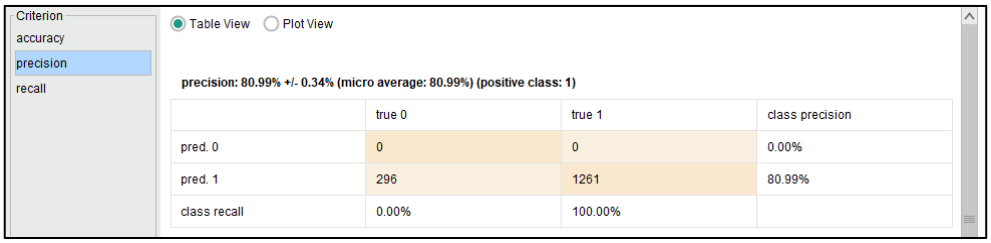
No.	Fitur
1.	Kurang Aktivitas Fisik
2.	Gula Berlebihan
3.	Lemak Berlebihan
4.	Diastol
5.	Berat Badan
6.	Lingkar Perut (cm)
7	Pemeriksaan Gula

Selanjutnya untuk hasil performa klasifikasi dengan *cross validation* menggunakan algoritma *naïve bayes* terdiri atas hasil *accuracy*, *precision*, dan *recall*. Hasil *accuracy* memperoleh nilai sebesar 80,99%, yang ditunjukkan pada gambar 5 dibawah ini.



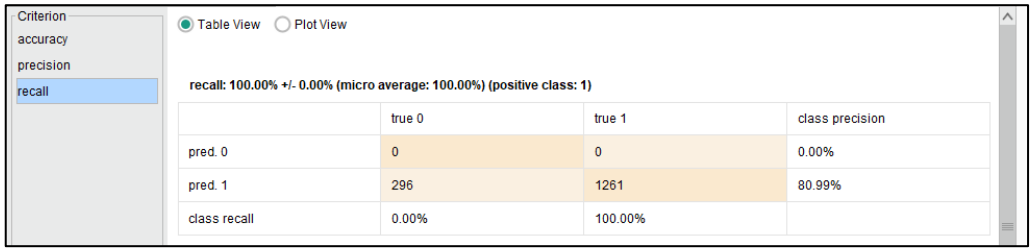
Gambar 5. Hasil Pengujian *Accuracy*

Hasil *precision* pada pengujian ini adalah sebesar 80,99%. Hasilnya sebagaimana disajikan pada gambar 7 berikut.



Gambar 7. Hasil Pengujian *Precision*

Hasil dari *Recall* pada pengujian ini adalah sebesar 100%. Hasilnya sebagaimana disajikan pada gambar 8 berikut.



Gambar 8. Hasil Pengujian *Recall*

3.2. Pembahasan

Pada bagian ini, pembahasan akan dibagi menjadi tiga tahapan, yaitu analisis optimasi seleksi fitur, analisis performa klasifikasi, dan kesimpulan dari optimasi seleksi fitur berdasarkan hasil performa klasifikasi. Setiap tahapan bertujuan untuk memberikan pemahaman yang komprehensif mengenai proses dan hasil yang diperoleh dari pengolahan data. Pembahasan akan difokuskan pada hasil pengolahan data menggunakan *RapidMiner*, yang digunakan untuk mendukung analisis secara sistematis dan terukur.

3.2.1. Analisis Optimasi Seleksi Fitur

Penelitian ini menggunakan algoritma genetika untuk optimasi seleksi fitur dalam klasifikasi penyakit diabetes melitus, dengan tujuan meningkatkan *accuracy*, *precision*, dan *recall*. *Dataset* berjumlah 1.557 data dengan 29 atribut awal.

Proses optimasi meliputi:

1. Inisialisasi (60% populasi awal acak) untuk menjaga keragaman solusi.
2. Seleksi (*Tournament Selection* 20%) untuk memilih individu terbaik.
3. *Crossover* (probabilitas 80%) untuk menggabungkan fitur unggul.
4. Mutasi (probabilitas 10%) untuk menambah variasi dan mencegah konvergensi prematur.

Hasilnya, diperoleh 7 fitur optimal dari 29 fitur awal, yaitu: kurang aktivitas fisik, gula berlebihan, lemak berlebihan, diastol, berat badan, lingkar perut, dan pemeriksaan gula. Optimasi ini efektif mengurangi dimensi data tanpa mengurangi fitur penting, sehingga berpotensi meningkatkan performa model.

3.2.2. Analisis Performa Klasifikasi

Pengujian performa klasifikasi dilakukan dengan algoritma *Naïve Bayes* menggunakan 10-*fold cross-validation* untuk evaluasi objektif dan mengurangi *overfitting*. *Dataset* dibagi menjadi 10 *subset*, di mana 9 subset digunakan untuk pelatihan dan 1 *subset* untuk pengujian, diulang hingga setiap *subset* menjadi data uji sekali.

Fitur yang digunakan merupakan hasil optimasi dari *Optimize Selection (Evolutionary)*, dengan proses yaitu pelatihan menggunakan *Naïve Bayes*, pengujian menggunakan *Apply Model*, dan evaluasi dengan *Performance (Binomial Classification)*.

Hasil perhitungan performa dari klasfikasi menggunakan *cross validation,* akan dipaparkan dengan tabel *confusion matrix. Confusion matrix* adalah representasi tabular yang menunjukkan berapa banyak data uji yang diprediksi tepat atau keliru oleh model klasifikasi [15]. Berikut tabel *confusion matrix* dari 1557 data hasil pengujian yang dapat diamati pada tabel 8 berikut.

Tabel 8 *Confusion Matrix*

Actual	Prediksi	
	Diabetes Melitus Tipe 1 (*Positive*)	Diabetes Melitus Tipe 2 (*Negative*)
Diabetes Melitus Tipe 1	1261 (TP)	0 (FN)
Diabetes Melitus Tipe 2	296 (FP)	0 (TN)

Berdasarkan tabel 8 maka *accuracy, precision, recall dan F1-Score* dapat dihitung dan diperoleh hasil.

1. *Precision*

$$Precision = \frac{1261}{1261+296} \text{ x } 100$$
$$= 0{,}8099 \text{ x } 100$$
$$= 80{,}99\ \%$$

2. *Recall*

$$Recall = \frac{1261}{1261+0} \text{ x } 100$$
$$= 1 \text{ x } 100$$

$$= 100\ \%$$

3. *F1-Score*

$$F1\text{-}Score = 2 \text{ x } \frac{0{,}8099 \text{ x } 1}{0{,}8099+1} \text{ x } 100$$
$$= 2 \text{ x } \frac{0{,}8099}{1{,}8099} \text{ x } 100$$
$$= 0{,}895 \text{ x } 100$$
$$= 89{,}5\ \%$$

4. *Accuracy*

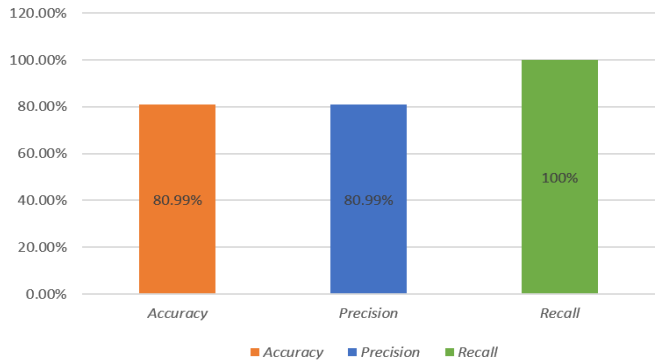
$$Accuracy = \frac{1261 + 0}{1261 + 296 + 0 + 0} \text{ x } 100$$
$$= 0{,}8099 \text{ x } 100$$
$$= 80{,}99\ \%$$

Adapun hasil pengujian kinerja model klasifikasi yang diterapkan dengan algoritma *Naive Bayes* menunjukkan nilai akurasi (*accuracy*), *precision*, dan *recall* sebagai berikut.

1. *Recall* = 100% menandakan bahwa seluruh kasus positif (pasien diabetes) berhasil terdeteksi tanpa ada satupun yang terlewat (FN = 0). Hal ini menunjukkan kekuatan model dalam mendeteksi pasien yang benar-benar mengidap diabetes, yang sangat penting dalam konteks medis karena kesalahan tipe II (gagal mendeteksi pasien sakit) dapat berakibat fatal.
2. *Precision* = 80,99% lebih rendah dibanding recall, artinya sekitar 19% dari prediksi positif sebenarnya adalah negatif (FP = 296). Kondisi ini kemungkinan dipengaruhi oleh distribusi data yang tidak seimbang (jumlah kasus positif jauh lebih banyak daripada negatif). Akibatnya, model cenderung "bermain aman" dengan mengklasifikasikan lebih banyak data sebagai positif, sehingga recall tinggi tetapi precision turun.
3. *Accuracy* = 80,99% cukup baik, tetapi dipengaruhi oleh ketidakseimbangan kelas. Pada dataset ini, dominasi kelas positif (diabetes) membuat *accuracy* lebih banyak mencerminkan keberhasilan model mengenali kelas mayoritas.
4. *F1-score* = 89,5% menjadi ukuran yang lebih seimbang, karena memperhitungkan *precision* dan *recall* sekaligus. Nilai ini menunjukkan performa model yang stabil, meskipun *trade-off* antara *precision* dan *recall* tetap ada.

Jika dibandingkan dengan penelitian Somantri dkk. [5] yang fokus pada optimasi bobot *Naïve Bayes* dengan peningkatan akurasi signifikan, hasil penelitian ini lebih menekankan pada keseimbangan recall dan *precision* melalui reduksi fitur berbasis GA. Dibanding penelitian Yuliana & Devianti [6] yang mengidentifikasi pola gejala dengan Apriori, penelitian ini memberikan kontribusi berbeda karena tidak hanya mengungkap pola gejala, tetapi juga membangun model klasifikasi yang dapat digunakan sebagai sistem pendukung keputusan medis.

Hasil pengujian ini kemudian divisualisasikan dalam bentuk grafik untuk memudahkan analisis dan interpretasi. Grafik tersebut menyajikan perbandingan nilai *accuracy*, *precision*, dan *recall* sehingga dapat dilihat tren performa model secara keseluruhan serta menentukan area di mana model bekerja dengan baik atau memerlukan perbaikan lebih lanjut. Visualisasi hasil pengujian ini dapat dilihat pada Gambar 9.



Gambar 9. Nilai *Accuracy, Precision* dan *Recall*

3.2.3. Kesimpulan Optimasi Seleksi Fitur Berdasarkan Hasil Performa Klasifikasi

Berdasarkan hasil pengujian klasifikasi dengan algoritma *Naïve Bayes* yang dioptimasi dengan metode seleksi fitur berbasis algoritma genetika, diperoleh bahwa dari 29 fitur awal yang telah melalui tahap preparasi maka diperoleh *subset* fitur terbaik yang terdiri dari 7 fitur. Optimasi tersebut berhasil meningkatkan efisiensi

model tanpa menurunkan akurasi secara signifikan. Adapun fitur terbaik hasil pengujian disajikan pada tabel 9 berikut.

Tabel 9 Fitur Terbaik

No.	Fitur
1.	Kurang Aktivitas Fisik
2.	Gula Berlebihan
3.	Lemak Berlebihan
4.	Diastol
5.	Berat Badan
6.	Lingkar Perut (cm)
7	Pemeriksaan Gula

Proses evaluasi menggunakan teknik *10-fold cross-validation*, dan menghasilkan performa model sebagai berikut:

1. Akurasi (*accuracy*) sebesar 80,99%, menunjukkan bahwa model mampu mengklasifikasikan sebagian besar data dengan benar.
2. *Precision* sebesar 80,99%, mengindikasikan bahwa sebagian besar prediksi positif adalah benar dan masih ada *false positive*, yang berarti beberapa pasien non-diabetes diklasifikasikan sebagai diabetes.
3. *Recall* mencapai 100%, menandakan bahwa semua kasus positif berhasil dikenali tanpa kesalahan dan seluruh pasien diabetes berhasil terdeteksi.
4. *F1-Score* sebesar 89,5%, menunjukkan bahwa model memiliki keseimbangan yang sangat baik antara *precision* dan *recall*.

Secara praktis, hasil ini menunjukkan bahwa model lebih menekankan keselamatan pasien (tidak ada kasus diabetes yang terlewat), meskipun dengan konsekuensi adanya beberapa prediksi berlebih (*false positive*). Kontribusi penelitian ini adalah menunjukkan bahwa integrasi GA dan *Naïve Bayes* tidak hanya meningkatkan efisiensi pemrosesan data, tetapi juga menghasilkan model yang cocok digunakan dalam skenario medis di mana sensitivitas (*recall*) lebih diutamakan dibandingkan spesifisitas (*precision*).

4. KESIMPULAN

Penelitian ini menunjukkan bahwa penerapan algoritma genetika efektif dalam memilih kombinasi fitur terbaik untuk meningkatkan performa klasifikasi pasien diabetes melitus. Proses seleksi dimulai dari inisialisasi populasi, seleksi individu terbaik, *crossover*, hingga mutasi, dengan *fitness function* berdasarkan akurasi algoritma *Naïve Bayes*. Dari 1.557 data dengan 17 fitur (1 sebagai label), diperoleh 7 fitur terpilih yang relevan, sementara fitur kurang berkontribusi berhasil dieliminasi.

Pengujian menggunakan *subset* 7 fitur tersebut menghasilkan performa klasifikasi yang baik, yaitu *Accuracy* 80,99%, *Precision* 80,99%, *Recall* 100%, dan *F1-Score* 89,5%. Nilai *recall* yang sempurna menandakan model mampu mengenai semua pasien positif diabetes tanpa kesalahan (tanpa *false negative*), yang sangat penting dalam deteksi penyakit. Dengan demikian, algoritma genetika terbukti mampu menyederhanakan fitur sambil mempertahankan tingkat deteksi tinggi terhadap pasien diabetes melitus..

REFERENSI

[1] P. Rahayu, I. G. I. Sudipa, Suryani, A. Surachman, A. Ridwan, I. G. M. Darmawiguna, M. Sutoyo, I. Slamet, S. Harlina, and I. M. May Sanjaya, *Buku Ajar Data Mining*. PT. Sonpedia Publishing Indonesia, 2024.

[2] E. Puwaningsih, L. Ludiana, and I. Immawati, "Penerapan Senam Kaki Diabetes untuk Meningkatkan Sensitivitas Kaki Pasien Diabetes Mellitus Tipe II di Puskesmas Metro," *Jurnal Cendikia Muda*, vol. 3, no. 2, pp. 235–244, 2023.

[3] I. Hidayat, S. Revo, L. Inkiriwang, and P. A. K. Pratasis, "Optimasi penjadwalan menggunakan metode algoritma genetika pada proyek rehabilitasi Puskesmas Minanga," *Jurnal Sipil Statik*, vol. 7, no. 12, pp. 1669–1680, 2019.

[4] S. F. Tahir and C. A. Sugianto, "Optimasi Naive Bayes Menggunakan Algoritma Genetika Pada Klasifikasi Komentar Cyberbullying Pada Media Sosial X," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 12, no. 3, 2024.

[5] E. Manalu, F. A. Siantur, and M. R. Manalu, "Penerapan algoritma Naive Bayes untuk memprediksi jumlah produksi barang berdasarkan data persediaan dan jumlah pemesanan pada CV Papadan Mama Pastries," *Jurnal Mantik Penusa*, vol. 1, no. 2, pp. 16–21, 2017.

[6] O. Somantri, R. H. Maharrani, and L. P. Wanti, "An optimize weights Naïve Bayes model for early detection of diabetes," *Telematika*, vol. 15, no. 1, pp. 14–22, 2022.

[7] A. Yuliana and F. Devianti, "Analisis pola diabetes melitus menggunakan algoritma Apriori," *Journal of Informatics and Electronics Engineering (JIEE),* vol. 5, no. 1, pp. 28–37, 2023.

[8] A. Yuliana, *Pengantar Metodologi Penelitian Kualitatif*. Bandung, Indonesia: CV. Gita Letera, 2024.

[9] A. Maulana and A. Yuliana, "Analisis Sentimen Opini Publik Terkait Judi Online Pada Pengguna Aplikasi X Menggunakan Algoritma Naïve Bayes Dan Support Vector Mechine," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 12, no. 3S1, 2024.

[10] W. D. Septiani and U. Rohwadi, "Optimasi Algoritma Genetika Pada Algoritma C4.5 untuk Deteksi Dini Penyakit Diabetes," *Akrab Juara: Jurnal Ilmu-ilmu Sosial*, vol. 6, no. 5, pp. 221–229, 2021.

[11] M. A. V. Darmawan, M. M. Al Haromainy, and A. Junaidi, "Optimasi algoritma k-nearest neighbor dengan algoritma genetika pada

deteksi penyakit diabetes mellitus," *Jurnal Teknik Informatika dan Sistem Informasi*, vol. 12, no. 2, pp. 247–259, 2025.
[12] M. R. Herdiansyah and A. Yuliana, "Analisis Sentimen Kebijakan Kampus Merdeka Menggunakan Naïve Bayes Berdasarkan Komentar Pada Youtube," *Jurnal Mahasiswa Teknik Informatika*, vol. 8, no. 6, pp. 12454–12459, 2024.
[13] W. E. Nugroho, A. Sofyan, and O. Somantri, "Metode Naive Bayes dalam menentukan program studi bagi calon mahasiswa baru," *Infotekmesin*, vol. 12, no. 1, pp. 59–64, 2021.
[14] F. R. Mashfia, "Prediksi ketepatan waktu kelulusan mahasiswa menggunakan metode Naïve Bayes classifier," Doctoral dissertation, Universitas Islam Negeri Maulana Malik Ibrahim, 2022.
[15] A. Yuliana and D. B. Pratomo, "Algoritma Decision Tree (C4.5) Untuk Memprediksi Kepuasan Mahasiswa Terhadap Kinerja Dosen Politeknik TEDC Bandung," *Seminar Nasional Inovasi Teknologi*, vol. 1, no. 1, pp. 377–384, 2017.