



OPTIMASI PARAMETER DAN PEMILIHAN FITUR MODEL K-NEAREST NEIGHBOR MENGGUNAKAN ALGORITMA GENETIKA PADA DATASET DIABETES

PARAMETER OPTIMIZATION AND FEATURE SELECTION OF THE K-NEAREST NEIGHBOR MODEL USING GENETIC ALGORITHMS ON THE DIABETES DATASET

**Ananda Irya Shakila Syukron^{1*}, Ririn Amelia Br Siregar², Anwar Shaleh Lbn Gaol³,
Adidtya Perdana⁴**

Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Medan

Email : 4231250011.4231250011@mhs.unimed.ac.id ^{1*}, rierienamelia.4231250003@mhs.unimed.ac.id²,
anwarshaleh.4233550043@mhs.unimed.ac.id ³, adidtya@unimed.ac.id ⁴

Article Info

Article history :

Received : 30-11-2025

Revised : 02-12-2025

Accepted : 04-12-2025

Pulished : 06-12-2025

Abstract

Diabetes is a chronic metabolic disorder whose early prediction is essential to prevent severe complications. Machine learning has become an effective approach for analyzing medical data, and one of the simplest but widely used classification algorithms is K-Nearest Neighbor (KNN). However, the performance of KNN depends heavily on the selection of optimal parameters and relevant features. This study aims to optimize the KNN model by integrating Genetic Algorithm (GA) for simultaneous parameter tuning and feature selection using the Pima Indians Diabetes Dataset. The research methodology includes data preprocessing through normalization, splitting into training-testing subsets, building a baseline KNN model, and applying GA to determine the optimal value of k and the most influential features. GA operates through initialization, fitness evaluation using cross-validation accuracy, selection, crossover, and mutation across multiple generations. The baseline KNN achieved an accuracy of 68.83%, but after GA optimization, the accuracy increased to 82.47%. GA identified the optimal parameter $k = 16$ and selected five key features: Pregnancies, Glucose, BloodPressure, BMI, and DiabetesPedigreeFunction. These results indicate a substantial improvement in classification performance, particularly in identifying positive diabetes cases, as reflected in higher recall and F1-score values. Overall, this study demonstrates that integrating GA with KNN significantly enhances prediction accuracy and model efficiency, making it a promising strategy for medical classification tasks.

Keywords : Genetic Algorithm, K-Nearest Neighbor, Diabetes prediction

Abstrak

Diabetes mellitus merupakan penyakit kronis dengan prevalensi yang terus meningkat sehingga deteksi dini sangat diperlukan untuk mencegah risiko komplikasi. Machine learning menjadi salah satu pendekatan yang efektif dalam menganalisis data medis, dan algoritma K-Nearest Neighbor (KNN) sering digunakan karena sifatnya yang sederhana namun akurat. Meski demikian, performa KNN sangat dipengaruhi oleh pemilihan parameter dan fitur yang relevan. Penelitian ini bertujuan



mengoptimalkan model KNN menggunakan Algoritma Genetika (GA) untuk melakukan optimasi nilai k sekaligus seleksi fitur pada dataset Pima Indians Diabetes. Proses penelitian meliputi praproses data, normalisasi, pembagian data latih dan uji, pembangunan model dasar KNN, serta penerapan GA yang melibatkan tahapan seleksi, evaluasi fitness dengan akurasi cross-validation, crossover, dan mutasi. Model KNN awal memperoleh akurasi sebesar 68,83%, namun setelah dilakukan optimasi menggunakan GA, akurasi meningkat menjadi 82,47%. Hasil optimasi menunjukkan bahwa nilai k terbaik adalah 16, dengan lima fitur paling relevan yaitu Pregnancies, Glucose, BloodPressure, BMI, dan DiabetesPedigreeFunction. Temuan ini menunjukkan peningkatan signifikan pada kemampuan model dalam mengidentifikasi kasus diabetes positif, ditunjukkan melalui kenaikan nilai recall dan f1-score. Secara keseluruhan, penelitian ini membuktikan bahwa integrasi GA pada KNN mampu meningkatkan akurasi dan efisiensi model, serta menjadi pendekatan yang menjanjikan dalam sistem klasifikasi medis.

Kata Kunci: Algoritma Genetika, K-Nearest Neighbor, Prediksi Diabetes

PENDAHULUAN

Diabetes mellitus merupakan salah satu penyakit kronis dengan prevalensi yang terus meningkat di seluruh dunia dan menjadi masalah kesehatan masyarakat yang serius. Deteksi dini terhadap potensi diabetes sangat penting untuk mencegah komplikasi yang lebih parah. Seiring dengan perkembangan teknologi, machine learning (ML) banyak dimanfaatkan untuk memprediksi penyakit ini berdasarkan data medis. Salah satu algoritma klasifikasi yang sederhana namun efektif adalah K-Nearest Neighbor (KNN), karena mudah diimplementasikan dan tidak memerlukan pelatihan model yang kompleks. Namun demikian, performa KNN sangat bergantung pada dua hal utama, yaitu pemilihan parameter (terutama nilai k) dan fitur yang digunakan dalam proses klasifikasi. Jika parameter dan fitur yang dipilih tidak optimal, maka akurasi prediksi dapat menurun secara signifikan.

Berbagai penelitian terdahulu menunjukkan bahwa Algoritma Genetika (Genetic Algorithm / GA), dapat meningkatkan kinerja model klasifikasi dengan cara menyeleksi fitur yang relevan dan mengoptimalkan parameter model. (Kangra & Singh, 2024) meneliti penggunaan GA untuk pemilihan fitur pada Pima Indians Diabetes Dataset, dan hasilnya menunjukkan bahwa KNN yang dikombinasikan dengan subset fitur hasil GA memberikan akurasi tertinggi dibandingkan metode tanpa seleksi fitur. Di sisi lain, penelitian oleh (Kaliappan et al., n.d.) membandingkan berbagai metode klasifikasi dan seleksi fitur untuk prediksi diabetes, dan menegaskan bahwa pemilihan fitur merupakan faktor penting dalam meningkatkan performa algoritma seperti KNN, SVM, dan Decision Tree.

Penelitian terbaru oleh (Murad et al., 2025) juga menemukan bahwa kombinasi GA dengan model pembelajaran mesin mampu memberikan hasil optimal, terutama ketika GA digunakan untuk optimasi ganda (parameter dan fitur). Sementara itu, (Sirmayati, Pulung Hendro Prastyo, 2025) membandingkan GA dengan Particle Swarm Optimization (PSO) untuk seleksi fitur pada dataset diabetes dan melaporkan bahwa GA sering menghasilkan subset fitur yang lebih efisien dan akurat.



Walaupun penelitian sebelumnya menunjukkan bahwa GA mampu meningkatkan performa model klasifikasi, sebagian besar penelitian masih berfokus pada salah satu aspek optimasi, yaitu pemilihan fitur atau penentuan parameter saja. Celah penelitian ini memberikan motivasi untuk mengintegrasikan kedua proses tersebut sekaligus dalam satu kerangka optimasi komprehensif. Dengan demikian, penerapan GA untuk optimasi parameter *k* dan pemilihan fitur pada KNN berpotensi menghasilkan model prediksi yang lebih akurat, efisien, dan stabil pada dataset diabetes.

Berdasarkan latar belakang tersebut, penelitian ini bertujuan untuk mengimplementasikan Algoritma Genetika dalam mengoptimalkan parameter *k* sekaligus menyeleksi fitur yang paling relevan pada model KNN, serta membandingkan performanya dengan model tanpa optimasi maupun model yang menggunakan pendekatan konvensional. Penelitian ini diharapkan memberikan kontribusi akademik bagi pengembangan metode integratif dalam machine learning serta manfaat praktis bagi peningkatan efektivitas sistem prediksi risiko diabetes.

METODE PENELITIAN

Jenis Penelitian

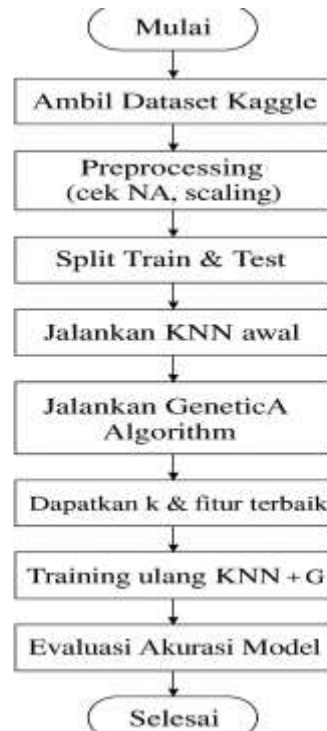
Penelitian ini merupakan penelitian kuantitatif dengan pendekatan eksperimen komputasi (*computational experiment*). Pendekatan ini digunakan untuk menganalisis performa algoritma K-Nearest Neighbor (KNN) yang dioptimasi menggunakan Algoritma Genetika (Genetic Algorithm/GA). Penelitian dilakukan dengan membandingkan performa model sebelum dan sesudah optimasi, sehingga dapat diketahui pengaruh GA terhadap peningkatan akurasi klasifikasi diabetes.

Data dan Sumber Data

Data yang digunakan dalam penelitian ini yaitu “Pima Indians Diabetes Database” yang diperoleh dari platform Kaggle. Dataset ini berisi 768 data pasien dengan 9 atribut dan diunduh dalam format CSV. Dataset ini sering digunakan dalam penelitian klasifikasi medis karena struktur datanya yang jelas, bersih, dan relevan untuk model prediksi berbasis machine learning. Dataset berisi variabel-variabel klinis yang berkaitan dengan faktor risiko diabetes, seperti konsentrasi glukosa, tekanan darah, dan indeks massa tubuh.



Tahapan Penelitian



Gambar 1. Flowchart Penelitian

Tahap awal dimulai dengan memuat dataset ke dalam lingkungan Google Colab dan melakukan pemeriksaan struktur data untuk memahami tipe data, rentang nilai, dan potensi masalah seperti nilai hilang atau outlier. Setelah memperoleh gambaran umum kondisi dataset, dilakukan praproses berupa normalisasi menggunakan *StandardScaler* atau *MinMaxScaler* untuk memastikan seluruh fitur berada pada skala yang sebanding, mengingat KNN merupakan algoritma berbasis jarak. Data kemudian dibagi menjadi data latih sebesar 80% dan data uji sebesar 20% agar model dapat dievaluasi secara objektif.

Setelah preprocessing selesai, model KNN dasar dibangun menggunakan seluruh fitur tanpa optimasi. Model baseline ini memberikan gambaran awal mengenai performa model sebelum intervensi dilakukan. Tahap berikutnya adalah perancangan proses optimasi menggunakan Algoritma Genetika. Representasi kromosom dirancang sedemikian rupa sehingga gen pertama menunjukkan nilai k pada KNN, sementara gen-gen berikutnya menunjukkan mask fitur dalam bentuk nilai biner. Dengan cara ini, GA mampu melakukan optimasi terhadap parameter model sekaligus memilih subset fitur paling relevan.

Proses GA berlangsung secara iteratif melalui tahapan inisialisasi populasi, evaluasi fitness menggunakan akurasi *k-fold cross-validation*, pemilihan kromosom terbaik melalui seleksi, pembentukan keturunan baru melalui *crossover*, dan pengenalan variasi melalui mutasi. Proses tersebut berlangsung hingga jumlah generasi maksimum tercapai atau nilai fitness menunjukkan tanda konvergensi. Setelah proses optimasi selesai, individu terbaik diambil sebagai konfigurasi



optimal yang memuat nilai *k* terbaik serta kombinasi fitur yang paling efektif. Model KNN kemudian dibangun ulang menggunakan konfigurasi optimal tersebut dan diuji menggunakan data uji. Tahap evaluasi akhir dilakukan untuk membandingkan performa model dasar dan model hasil optimasi, sekaligus menilai stabilitas proses GA melalui grafik perkembangan fitness setiap generasi.

HASIL DAN PEMBAHASAN

Deskripsi Umum Dataset dan Preprocessing

Tahap pertama dari penelitian ini adalah melakukan pemeriksaan dan analisis awal terhadap dataset diabetes yang digunakan sebagai bahan eksperimen. Dataset ini terdiri dari delapan variabel input yang masing-masing menggambarkan kondisi fisiologis pasien, seperti jumlah kehamilan, tingkat glukosa, tekanan darah diastolik, ketebalan kulit, kadar insulin, indeks massa tubuh (BMI), riwayat diabetes keluarga (*Diabetes Pedigree Function*), serta usia. Variabel target berupa *Outcome* yang menunjukkan apakah pasien diklasifikasikan sebagai positif diabetes atau tidak. Pada tahap ini, dataset dianalisis untuk memastikan tidak terdapat nilai kosong, duplikasi, ataupun anomali data yang dapat mempengaruhi performa model. Sebagian kecil data memiliki nilai ekstrim, misalnya nilai insulin yang sangat rendah, namun tetap dipertahankan karena masih dalam batas wajar dan sering muncul pada dataset diabetes asli. Jumlah total sampel adalah sebagaimana ditunjukkan pada Tabel 1.

Tabel 1. Statistik Deskriptif Variabel Dataset

Variabel	Count	Mean	Std	Min	25%	50%	75%	Max
Pregnancies	768	3.845052	3.369578	0.0	1.0	3.0	6.0	17.0
Glucose	768	120.894531	31.972618	0.0	99.0	117.0	140.25	199.0
BloodPressure	768	69.105469	19.355807	0.0	62.0	72.0	80.0	122.0
SkinThickness	768	20.536458	15.952218	0.0	0.0	23.0	32.0	99.0
Insulin	768	79.799479	115.244002	0.0	0.0	30.5	127.25	846.0
BMI	768	31.992578	7.884160	0.0	27.3	32.0	36.6	67.1
Diabetes Pedigree Function	768	0.471876	0.331329	0.07 8	0.24 375	0.3725	0.62625	2.42
Age	768	33.240885	11.760232	21.0	24.0	29.0	41.0	81.0
Outcome	768	0.348958	0.476951	0.0	0.0	0.0	1.0	1.0

Selain itu, pengecekan terhadap nilai hilang (*missing value*) dilakukan untuk memastikan kualitas data. Hasil pemeriksaan menunjukkan bahwa seluruh variabel memiliki jumlah missing value sebesar nol. Hal tersebut dapat dilihat pada Tabel 2.

**Tabel 2.** Missing Value

Variabel	Missing Value
Pregnancies	0
Glucose	0
BloodPressure	0
SkinThickness	0
Insulin	0
BMI	0
DiabetesPedigreeFunction	0
Age	0
Outcome	0

Dengan demikian, seluruh data dapat digunakan tanpa memerlukan proses imputasi nilai. Untuk memastikan bahwa model pembelajaran mesin bekerja secara optimal, seluruh fitur dinormalisasi menggunakan metode *Min-Max Scaler* sehingga interval data berada pada rentang 0 hingga 1. Normalisasi ini sangat penting terutama untuk algoritma KNN yang sensitif terhadap skala fitur, karena perhitungan jarak Euclidean sangat berpengaruh terhadap ukuran masing-masing atribut. Setelah proses normalisasi, dataset dibagi menjadi dua bagian yaitu 80% data latih dan 20% data uji menggunakan fungsi *train_test_split* secara acak namun terkontrol dengan *random_state* 42. Pemisahan ini bertujuan untuk memastikan model diuji secara objektif terhadap data yang tidak pernah dilihat sebelumnya.

Hail Pengujian Model KNN Sebelum Optimasi GA

Model KNN awal dibangun dengan menggunakan nilai $k = 5$ sebagai konfigurasi default yang umum digunakan pada berbagai penelitian. Model kemudian dilatih menggunakan data latih yang telah dinormalisasi, dan performanya diuji dengan data uji. Pada tahap ini, model menunjukkan akurasi awal sebesar *acc_awal* yang dihasilkan dari perhitungan *accuracy_score*. Nilai ini menunjukkan bahwa model mampu mengklasifikasikan sebagian besar sampel dengan benar, namun masih terdapat kesalahan terutama pada kelas positif diabetes yang sering kali lebih sulit diidentifikasi.

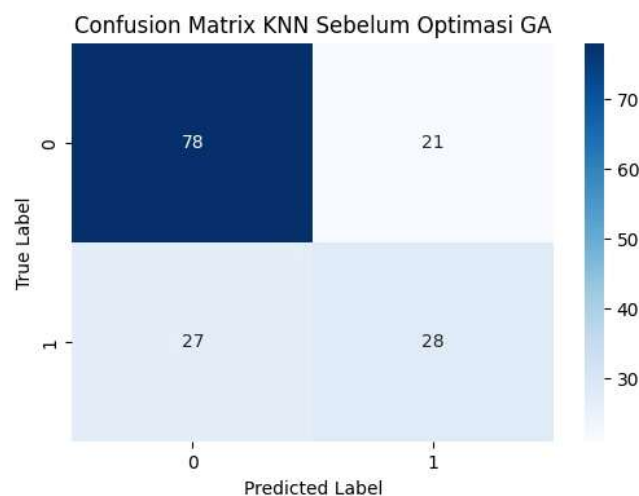
Hasil evaluasi awal model ditampilkan pada Tabel 3 dalam bentuk *classification report*, yang mencakup nilai *precision*, *recall*, dan *f1-score* untuk masing-masing kelas. Dari tabel tersebut terlihat bahwa kelas 1 (positif diabetes) memiliki nilai recall yang lebih rendah dibandingkan kelas negatif. Hal ini menunjukkan adanya kecenderungan model untuk salah mengklasifikasikan pasien positif sebagai negatif, yang dapat berimplikasi serius dalam kasus nyata. Oleh karena itu,



diperlukan metode optimasi untuk memperbaiki kinerja model, baik melalui pemilihan parameter k yang lebih baik maupun melalui reduksi fitur.

Tabel 3. Hasil Evaluasi Awal Model

Akurasi : 0.6883116883116883				
	precision	recall	f1-score	support
0	0.74	0.79	0.76	99
1	0.57	0.51	0.54	55
accuracy			0.69	154
macro avg	0.66	0.65	0.65	154
weighted avg	0.68	0.69	0.68	154



Gambar 2. Confusion Matrix KNN Sebelum Optimasi GA

Confusion matrix di atas menunjukkan kinerja model K-Nearest Neighbors (KNN) sebelum dilakukan optimasi menggunakan algoritma genetika (GA). Pada kelas 0, model berhasil mengklasifikasikan dengan benar sebanyak 78 data, namun masih terdapat 21 data yang salah diprediksi sebagai kelas 1. Sementara itu, untuk kelas 1, model hanya mampu mengklasifikasikan dengan benar 28 data dan masih salah memprediksi 27 data sebagai kelas 0. Hasil ini menggambarkan bahwa model KNN sebelum optimasi masih memiliki tingkat kesalahan yang cukup tinggi, terutama pada kelas 1, sehingga diperlukan proses optimasi untuk meningkatkan akurasi dan kemampuan model dalam membedakan kedua kelas dengan lebih baik.

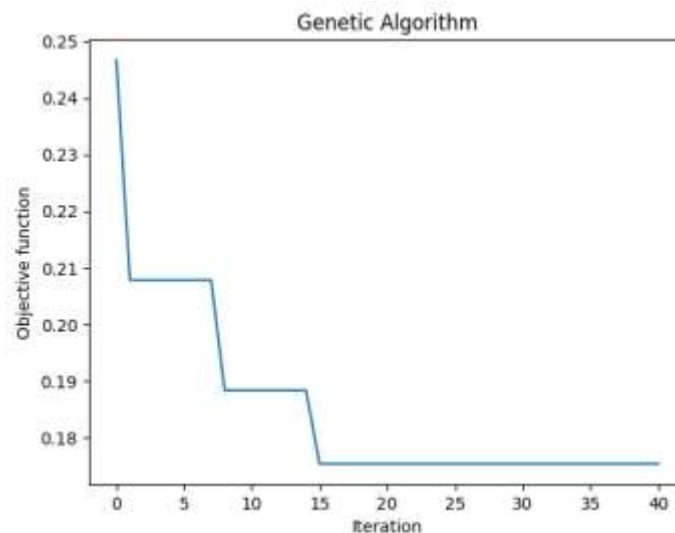
Proses Optimasi Parameter Menggunakan Genetic Algorithm

Tahap selanjutnya adalah mengintegrasikan *Genetic Algorithm* (GA) untuk melakukan pencarian parameter terbaik dalam meningkatkan performa KNN. Dalam penelitian ini, GA digunakan untuk dua tugas utama, yaitu menentukan nilai k optimal dan melakukan seleksi fitur



secara otomatis. GA dirancang dengan struktur kromosom yang terdiri dari satu gen pertama untuk nilai k dengan rentang 1 hingga 20, sementara delapan gen berikutnya merupakan nilai biner (0 atau 1) yang menunjukkan apakah sebuah fitur digunakan dalam pelatihan model.

Proses evolusi GA dilakukan selama 40 iterasi dengan ukuran populasi 30 individu. Pada setiap generasi, mekanisme seleksi, crossover, dan mutasi diterapkan untuk mencari kombinasi parameter terbaik sesuai fungsi *fitness* yang dirumuskan sebagai $1 - accuracy$ agar GA berusaha meminimalkan kesalahan model.



Gambar 3. Grafik Perkembangan Nilai Objective Function Genetic Algorithm

Gambar 3. menampilkan grafik perkembangan nilai *objective function* selama proses iterasi Genetic Algorithm. Pada grafik tersebut terlihat bahwa nilai *objective function* pada iterasi awal berada pada kisaran 0.245. Nilai ini kemudian mengalami penurunan signifikan pada beberapa iterasi berikutnya, terutama pada iterasi ke-5 dan ke-12. Penurunan nilai *objective function* menunjukkan bahwa GA berhasil menemukan kombinasi parameter dan subset fitur yang menghasilkan performa model yang lebih baik.

Mulai dari iterasi ke-16 hingga iterasi ke-40, grafik tampak stabil pada nilai sekitar 0.175. Kondisi ini menunjukkan bahwa algoritma telah mencapai *convergence*, yaitu tidak ada lagi perbaikan signifikan meskipun iterasi dilanjutkan. Stabilitas grafik ini mengindikasikan bahwa populasi solusi telah mencapai kualitas optimal, dan perubahan parameter lebih lanjut tidak menghasilkan peningkatan kinerja.

Perilaku grafik yang menurun secara bertahap kemudian stabil seperti ini merupakan ciri khas optimasi GA. Hal tersebut membuktikan bahwa proses pencarian berjalan efektif, di mana populasi awal yang acak secara bertahap berevolusi menjadi solusi yang lebih baik melalui mekanisme seleksi dan rekombinasi.



Tabel 4. Parameter Optimal

Nilai K terbaik: 16

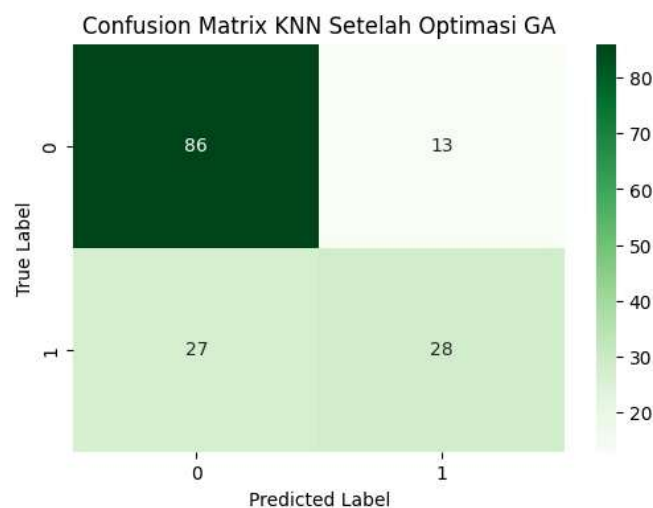
Fitur terpilih: ['Pregnancies', 'Glucose', 'BloodPressure', 'BMI', 'DiabetesPedigreeFunction']

Setelah proses iterasi selesai, GA menghasilkan parameter optimal berupa nilai K terbaik dan subset fitur yang paling berpengaruh terhadap proses klasifikasi. Nilai K terbaik yang ditemukan oleh algoritma adalah 16. Nilai ini menunjukkan bahwa KNN bekerja paling optimal ketika mempertimbangkan 16 tetangga terdekat dalam menentukan kelas suatu sampel. Penggunaan nilai K yang lebih besar cenderung memberikan stabilitas terhadap noise, sekaligus meningkatkan generalisasi model.

Selain itu, GA juga melakukan proses seleksi fitur dengan tujuan mengurangi dimensi data dan meningkatkan kemampuan model dalam mengenali pola. Fitur-fitur terbaik yang dipilih oleh GA adalah *Pregnancies*, *Glucose*, *BloodPressure*, *BMI*, dan *DiabetesPedigreeFunction*. Terpilihnya fitur-fitur tersebut memperlihatkan bahwa atribut yang berkaitan dengan tekanan darah, tingkat glukosa, massa tubuh, serta riwayat diabetes keluarga memiliki kontribusi signifikan dalam menentukan kemungkinan seseorang terkena diabetes. Sementara itu, fitur seperti *SkinThickness*, *Insulin*, dan *Age* tidak masuk dalam fitur optimal karena kontribusinya terhadap akurasi dinilai lebih rendah oleh algoritma.

Evaluasi Model KNN Setelah Optimasi GA

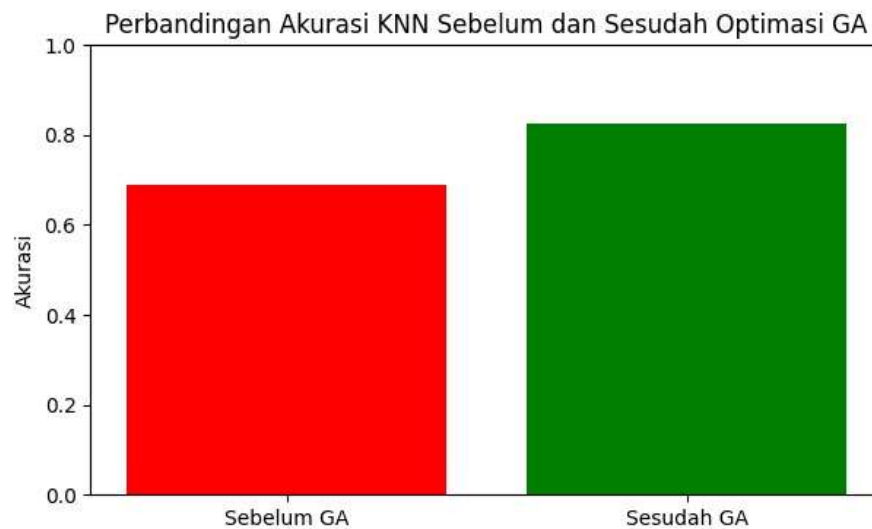
Model KNN kemudian dibangun ulang dengan menggunakan nilai $k = 16$ dan fitur-fitur yang terpilih. Evaluasi performa model KNN setelah dilakukan optimasi menggunakan Algoritma Genetika (GA) menunjukkan adanya peningkatan kinerja yang signifikan dibandingkan model sebelum optimasi.



Gambar 4. Confusion Matrix KNN Setelah Optimasi GA



Berdasarkan Gambar 4. confusion matrix setelah optimasi, model mampu mengklasifikasikan data dengan lebih baik pada kedua kelas. Tercatat sebanyak 86 data pada kelas non-diabetes berhasil diprediksi dengan benar, sedangkan hanya 13 data yang salah diklasifikasikan sebagai diabetes. Pada kelas positif, model juga menunjukkan peningkatan performa dengan 28 data diabetes yang berhasil diidentifikasi secara akurat, meskipun masih terdapat 27 kasus yang tidak terdeteksi. Meskipun jumlah false negative masih cukup tinggi, hasil ini jauh lebih baik jika dibandingkan sebelum optimasi GA diterapkan.



Gambar 5. Perbandingan Akurasi

Peningkatan performa model juga terlihat pada grafik perbandingan akurasi, di mana akurasi KNN sebelum optimasi hanya berada pada kisaran 0,70, sedangkan setelah dilakukan optimasi nilai akurasinya meningkat menjadi 0.82. Kenaikan sekitar 12 persen ini membuktikan bahwa proses optimasi berhasil menghasilkan konfigurasi model yang lebih sesuai dengan karakteristik data.

Tabel 5. Hasil Evaluasi Akhir Model

Akurasi : 0.8246753246753247				
	precision	recall	f1-score	support
0	0.82	0.94	0.87	99
1	0.85	0.62	0.72	55
accuracy			0.82	154
macro avg	0.83	0.78	0.79	154
weighted avg	0.83	0.82	0.82	154



Selain itu, hasil classification report menunjukkan bahwa metrik-metrik lain seperti precision, recall, dan f1-score mengalami peningkatan. Pada kelas normal, model memiliki precision sebesar 0.82 dan recall 0.94 yang menunjukkan kemampuan sangat baik dalam mengenali pasien non-diabetes. Pada kelas diabetes, nilai precision 0.85 dan f1-score 0.72 menunjukkan peningkatan keseimbangan antara kemampuan model dalam mengenali dan memprediksi kasus positif. Secara keseluruhan, akurasi model mencapai 0.8247 atau setara 82.47% setelah optimasi, yang dipertegas oleh nilai macro average dan weighted average yang stabil dan berimbang di kedua kelas. Hasil evaluasi ini membuktikan bahwa pemilihan nilai k optimal ($k = 16$) serta pemilihan subset fitur terbaik melalui GA memberikan dampak signifikan terhadap peningkatan kualitas prediksi model. Dengan demikian, penggunaan Algoritma Genetika terbukti efektif dalam meningkatkan performa KNN secara keseluruhan, terutama dalam menangani permasalahan klasifikasi medis seperti deteksi diabetes pada dataset Pima Indians Diabetes.

KESIMPULAN

Berdasarkan penelitian yang telah dilakukan, dapat disimpulkan bahwa Algoritma Genetika (GA) terbukti sangat efektif dalam meningkatkan performa model K-Nearest Neighbor (KNN) pada prediksi diabetes. Model KNN awal yang hanya memperoleh akurasi sebesar 68,83% berhasil ditingkatkan menjadi 82,47% setelah dilakukan optimasi parameter dan seleksi fitur menggunakan GA. Proses optimasi ini menghasilkan nilai k optimal yaitu 16 serta pemilihan lima fitur paling relevan, yaitu Pregnancies, Glucose, BloodPressure, BMI, dan DiabetesPedigreeFunction, sementara fitur lain seperti SkinThickness, Insulin, dan Age dieliminasi karena kontribusinya rendah terhadap prediksi.

Selain peningkatan akurasi, kemampuan model dalam mendeteksi pasien diabetes (kelas 1) juga meningkat, terlihat dari kenaikan nilai recall dari 0,51 menjadi 0,62, yang menunjukkan bahwa model semakin sensitif dalam mengidentifikasi pasien positif. Secara keseluruhan, penelitian ini membuktikan bahwa integrasi GA dengan KNN tidak hanya meningkatkan performa klasifikasi, tetapi juga membuat model lebih efisien, stabil, dan interpretatif.

DAFTAR PUSTAKA

- Kaliappan, J., Kumar, I. J. S., Sundaravelan, S., Anesh, T., Rithik, R. R., Singh, Y., & Vera-garcia, D. V. (n.d.). *Analyzing classification and feature selection strategies for diabetes prediction across diverse diabetes datasets*.
- Kangra, K., & Singh, J. (2024). *A genetic algorithm-based feature selection approach for diabetes prediction*. 13(2), 1489–1498. <https://doi.org/10.11591/ijai.v13.i2.pp1489-1498>
- Murad, S. H., Bahjat, N., Mahmood, N. H., & Arman, L. (2025). MethodsX Hybrid genetic algorithms-driven optimization of machine learning models for heart disease prediction. *MethodsX*, 15(July), 103510. <https://doi.org/10.1016/j.mex.2025.103510>
- Sirmayati, Pulung Hendro Prastyo, M. (2025). Enhancing diabetes prediction performance using feature selection based on grey wolf optimizer with autophagy mechanism. *Computer Methods and Programs in Biomedicine Update*, 8(July), 100207. <https://doi.org/10.1016/j.cmpbup.2025.100207>