JURIKOM (Jurnal Riset Komputer), Vol. 12 No. 3, Juni 2025
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v12i3.8615
Hal 230-240
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

Optimasi Algoritma K-Nearest Neighbors pada Prediksi Penyakit Diabetes

Sitti Arfiah*, Farid Wajidi, Nahya Nur

Teknik, Informatika, Universitas Sulawesi Barat, Majene, Indonesia
Email: [1*]sittiarfiah7@gmail.com , [2]faridwajidi@unsulbar.ac.id , [3]nahya.nur@unsulbar.ac.id
Email Penulis Korespondensi : sittiarfiah7@gmail.com
Submitted **15-05-2025;** Accepted **13-06-2025;** Published **30-06-2025**

Abstrak

Diabetes melitus merupakan salah satu penyakit kronis yang ditandai dengan tingginya kadar gula darah akibat gangguan pada sistem metabolisme tubuh, khususnya terkait produksi atau efektivitas hormon insulin. Ketidakmampuan tubuh dalam menggunakan insulin secara optimal dapat mengganggu proses penyerapan glukosa ke dalam sel, sehingga kadar gula dalam darah tetap tinggi. Jika tidak ditangani dengan tepat, kondisi ini dapat menimbulkan komplikasi serius seperti kerusakan pada jantung, ginjal, pembuluh darah, mata, dan sistem saraf. Oleh karena itu, deteksi dini dan akurat sangat diperlukan agar penderita dapat segera memperoleh penanganan medis yang tepat. Penelitian ini bertujuan untuk meningkatkan akurasi sistem klasifikasi penyakit diabetes menggunakan algoritma *K-Nearest Neighbors* (KNN). Model awal KNN dengan data tidak seimbang (tanpa SMOTE) & tanpa *GridSearchCV* hanya mampu menghasilkan akurasi sebesar 83%. Meskipun nilai ini terlihat cukup baik, namun performa terhadap kelas positif masih rendah, dengan *precision* sebesar 80%, *recall* hanya 69%, dan *F1-score* 74%, sedangkan untuk kelas negatif dengan *precision sebesar* 84%, *recall* 91%, dan *F1-score* 88% . Hal ini menunjukkan bahwa model cenderung bias terhadap kelas negatif, yang umum terjadi ketika jumlah data antar kelas tidak seimbang.Untuk memperbaiki kinerja model, dilakukan sejumlah tahapan penting, yaitu preprocessing data untuk menangani data hilang dan normalisasi fitur, optimasi *hyperparameter* menggunakan *GridSearchCV*, serta penyeimbangan data menggunakan metode *Synthetic Minority Over-sampling Technique* (SMOTE). Setelah ketiga tahapan tersebut diterapkan, model KNN menunjukkan peningkatan performa yang signifikan dengan *accuracy* mencapai menjadi 94%, dan performa terhadap kelas positif menunjukkan perbaikan besar dengan *precision* 90%, *recall* 98%, dan *F1-score* mencapai 94%, sedangkan kelas negatif dengan *precision sebesar* 98%, *recall* 89%, dan *F1-score* 93%. Hasil ini menunjukkan bahwa kombinasi preprocessing, optimasi model, dan penyeimbangan kelas dapat secara efektif meningkatkan kemampuan klasifikasi algoritma KNN dalam mendeteksi diabetes secara lebih akurat dan seimbang. Penelitian ini membuktikan bahwa pendekatan berbasis machine learning yang didukung oleh tahapan pemrosesan data yang tepat dapat membantu dalam pengembangan sistem pendukung keputusan medis, khususnya dalam proses diagnosis dini penyakit diabetes.

Kata Kunci: Optimasi K-Nearest Neighbors; GridSearchCV; Synthetic Minority Over-sampling Technique; Prediksi Diabetes

Abstract

Diabetes mellitus is a chronic disease characterized by high blood sugar levels due to metabolic system disturbances, specifically related to insulin production or effectiveness. If left untreated, it can lead to serious complications. Early and accurate detection is crucial for timely medical intervention. This research aimed to improve the accuracy of a diabetes classification system using the K-Nearest Neighbors (KNN) algorithm. An initial KNN model with imbalanced data (without SMOTE) and no GridSearchCV achieved only 83% accuracy. While seemingly good, its performance for the positive class was low (precision 80%, recall 69%, F1-score 74%), indicating bias towards the negative class due to data imbalance. To address this, several steps were implemented: data preprocessing (handling missing data and feature normalization), hyperparameter optimization using GridSearchCV, and data balancing with SMOTE. After these improvements, the KNN model showed significant performance gains, with accuracy reaching 94%. Performance for the positive class greatly improved (precision 90%, recall 98%, F1-score 94%), and for the negative class (precision 98%, recall 89%, F1-score 93%). These results demonstrate that combining preprocessing, model optimization, and class balancing effectively enhances the KNN algorithm's ability to detect diabetes more accurately and robustly, proving that machine learning with proper data processing can aid in developing medical decision support systems for early diabetes diagnosis.

Keywords: Optimization K-Nearest Neighbors; GridSearchCV; Synthetic Minority Over-sampling Technique; Diabetes Prediction

1. PENDAHULUAN

Menurut definisi yang diberikan oleh World Health Organization (WHO) , kesehatan tidak hanya diartikan sebagai ketiadaan penyakit atau kondisi lemah fisik semata, melainkan mencakup suatu keadaan yang utuh dan menyeluruh, mencakup kesejahteraan fisik, kondisi mental yang stabil, serta kehidupan sosial yang harmonis. Dengan kata lain, seseorang dapat dikatakan benar-benar sehat apabila ia berada dalam kondisi seimbang dan optimal pada aspek ketiga tersebut, bukan hanya karena tidak mengalami gangguan kesehatan secara fisik.

Seiring berjalannya waktu, struktur masyarakat mengalami perubahan yang signifikan. Perubahan ini dipicu oleh pergeseran dari kehidupan agraris menuju era industri. Dampaknya terlihat pada pola konsumsi makanan dan tingkat aktivitas fisik masyarakat. Misalnya kebiasaan makan telah berubah, dengan semakin banyaknya orang yang memilih makanan cepat saji karena kepraktisannya dan kemudahannya aksesnya. Gaya hidup yang berkembang saat ini telah menyebabkan penurunan signifikan dalam aktivitas fisik, terutama di kalangan pekerja kantoran yang cenderung menghabiskan sebagian besar waktu mereka di dalam ruangan dengan sedikit gerakan. Faktor – faktor ini berperan penting dalam meningkatkan risiko penyakit tidak menular serta penyakit degeneratif. Salah satu penyakit tidak menular yang sering dikaitkan dengan kebiasaan makan yang buruk dan kurangnya aktivitas fisik adalah diabetes melitus [1].

Diabetes menempati posisi sebagai salah satu dari tiga penyakit utama dengan tingkat kematian tertinggi di Indonesia, dengan tingkat kematian yang hanya berada di bawah stroke dan penyakit jantung, menjadikannya ancaman

JURIKOM (Jurnal Riset Komputer), Vol. 12 No. 3, Juni 2025
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v12i3.8615
Hal 230-240
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

serius bagi kesehatan masyarakat [2]. Menurut Riset Kementerian Kesehatan Republik Indonesia, Penyakit diabetes melitus kini menjadi permasalahan global yang berdampak serius terhadap kesehatan masyarakat serta perkembangan sosial dan ekonomi. Menurut data dari International Diabetes Federation, diperkirakan jumlah penderita diabetes di seluruh dunia akan mencapai 537 juta orang pada tahun 2021 dan diproyeksikan meningkat menjadi 783 juta pada tahun 2045. Indonesia menempati posisi kelima sebagai negara dengan jumlah penderita diabetes terbanyak, dengan total kasus sekitar 19,47 juta pada tahun yang sama [3]. Pada penderita diabetes melitus, sel-sel tubuh tidak lagi merespon insulin dengan efektif, atau pankreas tidak memproduksi insulin sama sekali, yang menyebabkan terjadinya hiperglikemia. Kondisi ini dapat menimbulkan komplikasi metabolik yang muncul secara tiba-tiba dalam waktu singkat. Selain itu, jika dibiarkan dalam jangka panjang, hiperglikemia dapat memicu komplikasi neuropatik yang serius [4].

Diabetes melitus terjadi akibat gangguan dalam proses metabolisme tubuh, yang menyebabkan peningkatan kadar glukosa dalam darah, padahal glukosa adalah sumber energi utama bagi sel-sel tubuh. Namun, apabila kadar glukosa tidak terjaga dengan baik, kondisi ini dapat menyebabkan komplikasi serius seperti penyakit kardiovaskuler, stroke, obesitas, serta gangguan pada mata, ginjal, dan sistem saraf [5]. Salah satu bentuk diabetes yang paling sering ditemui adalah diabetes melitus [6]. Diabetes melitus sendiri merupakan gangguan metabolik yang ditandai dengan kadar gula darah yang tinggi (hiperglikemia), yang disebabkan oleh gangguan dalam produksi insulin, kemampuan insulin bekerja, atau keduanya [7].

Penelitian ini tentunya juga tidak lepas dari kontribusi penelitian sebelumnya salah satunya adalah yang dilakukan oleh A. Yaqin, D. Kurniawan, and J. Zeniarja, Berdasarkan Analisis evaluasi model dan *confusion matrix* menunjukkan peningkatan kualitas pada aspek *hyperparameter* dan tahap preprocessing terbukti memberikan dampak yang signifikan terhadap kinerja model KNN dalam melakukan prediksi terhadap penyakit diabetes . Model ini dibangun tanpa melalui proses preprocessing menghasilkan performa yang kurang optimal. Melalui penerapan preprocessing yang tepat serta konfigurasi *hyperparameter* yang optimal seperti n_neighbors = 18, *Weight* = '*distance'*, dan *metric* = '*manhattan'*, model mampu mencapai *accuracy* sebesar 88%, dengan *precision* 75%, *recall* 89%, dan *F1-score* 82%[8]. Selain itu penelitian yang dilakukan oleh Oktaviana dkk, Evaluasi model yang dilakukan dengan menggunakan *metric accuracy*, *precision*, *recall*, dan *F1-score* menunjukkan hasil yang masing-masing mencapai 88%, 83,54%, 87,5%, dan 85,36%. Hasil ini membuktikan bahwa algoritma K-NN dapat diterapkan dengan efektif dan menghasilkan evaluasi yang memuaskan [9].

Disisi lain, penelitian yang dilakukan oleh N. Maulidah, R, dkk yang membandingkan kinerja algoritma *Support Vektor Machine* (SVM) dan *Naive Bayes* dalam memprediksi penyakit diabetes melitus, menunjukkan bahwa SVM memiliki *accuracy* lebih tinggi 78,04% dibandingkan *Naive Bayes* 76,98%, berdasarkan data rekam medis pasien sebanyak 2000 dari kaggle. Penelitian ini menekankan pentingnya pemilihan metode klasifikasi yang tepat dalam aplikasi data mining untuk mendukung diagnosa penyakit tidak menular seperti diabetes mellitus [10]. Selain itu, penelitian serupa yang dilakukan oleh J. Lemantara dan T. Lusiani juga berfokus pada upaya pencegahan penyakit diabetes pada wanita dengna membandingkan dua metode klalsifikasi, yaitu *Naïve Bayes* dan KNN, dan mendapatkan hasil menunjukan metode *Naive Bayes* memiliki tingkat *accuracy* yang lebih unggul yaitu 78,358%, sedangkan metode KNN memiliki tingkat *accuracy* 77,985% [11].

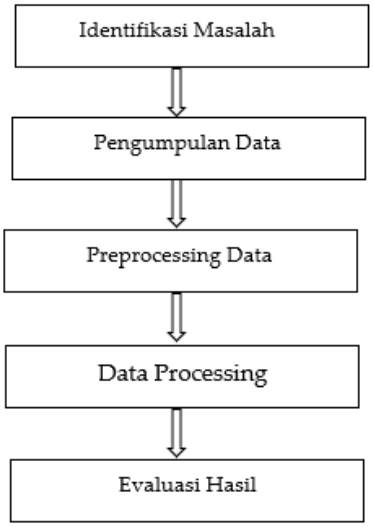
Adapun penelitian dilakukan oleh R. A. Pangestu, T. Taslim, Y. Yunefri, K. Kursiasih, and E. Sabna menggunakan dataset prakondisi pasien COVID-19 yang diperoleh dari Pemerintah Meksiko tahun 2020 melalui situs Kaggle (www.kaggle.com). Dataset dibagi dengan proporsi 70% untuk data pelatihan dan 30% untuk data pengujian, dengan kelas 'yes' dan 'no' yang menunjukkan kebutuhan perawatan ICU. Optimasi nilai k dilakukan menggunakan metode 5-fold *Cross Validation*, Evaluasi performa model menggunakan *Confusion Matrix* menghasilkan tingkat *accuracy* mencapai 86,47% pada nilai k = 16 [12]. Sementara penelitian yang dilakukan oleh E. Safitri, D. Rofianto, N. Purwati, H. Kurniawan, and S. Karnila, mengembangkan model prediksi yang akurat untuk diabetes melitus menggunakan tiga pembelajaran mesin algoritma, yaitu *Random Forest, Regresi Logistik*, dan *Decision Tree*. Hasil penelitian menunjukkan bahwa Regresi Logistik memperoleh *accuracy* tertinggi (75%) dengan kinerja yang seimbang dalam mendeteksi kasus *positif* dan *negatif*. *Decision Tree* menunjukkan kinerja terbaik dalam hal *recall*, sedangkan *Random Forest* menampilkan keseimbangan yang lebih rendah antara *precision* dan *recall*. Analisis kurva ROC menunjukkan bahwa *Random Forest* memiliki AUC tertinggi (0,82), diikuti oleh Regresi Logistik (0,81), dan Decision Tree (0,73). Penelitian ini mengkonfirmasi bahwa pembelajaran mesin algoritma dapat digunakan secara efektif untuk prediksi diabetes, memberikan alat yang berharga untuk deteksi dini dan intervensi, yang dapat mengurangi dampak global diabetes melitus [13].

Berdasarkan permasalahan yang telah diidentifikasi, tujuan dari penelitian ini untuk menerapkan model prediksi penyakit diabetes menggunakan algoritma *K- Nearest Neighbors* (KNN) yang dioptimalkan. Optimasi dilakukan dengan dua pendekatan utama, yaitu penentuan jumlah tetangga terdekat (k) terbaik menggunakan *GridSearchCV*, serta penanganan ke ketidakseimbangan kelas melalui metode *Synthetic Minority Oversampling Technique* . Kombinasi kedua teknik ini diharapkan mampu meningkatkan kinerja klasifikasi, khususnya dalam mengenali kasus positif (penderita diabetes) yang sering kali terabaikan pada dataset yang tidak seimbang. Dengan pendekatan ini, penelitian bertujuan menghasilkan model yang tidak hanya akurat secara keseluruhan, tetapi juga dapat mendeteksi kondisi yang krusial secara medis.

2. METODOLOGI PENELITIAN

JURIKOM (Jurnal Riset Komputer), Vol. 12 No. 3, Juni 2025
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v12i3.8615
Hal 230-240
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

Penelitian ini terdiri atas lima tahapan yaitu identifikasi masalah, pengumpulan data, preprocessing data,data processing, dan evaluasi hasil yang ditunjukkan pada Gambar 1 sebagai berikut:



Gambar 1. Tahapan Penelitian

2.1 Identifikasi Masalah

Diabetes melitus (DM) adalah salah satu penyakit tidak menular yang paling mematikan baik di Indonesia maupun di dunia, yang disebabkan oleh gangguan metabolisme, yaitu kelainan dalam produksi atau fungsi insulin yang menyebabkan hiperglikemia. Perubahan gaya hidup masyarakat modern, seperti tingginya konsumsi makanan cepat saji dan berkurangnya aktivitas fisik, turut berkontribusi pada peningkatan prevalensi penyakit ini. Data dari International Diabetes Federation menunjukkan peningkatan jumlah penderita diabetes di seluruh dunia, dengan Indonesia menempati urutan kelima sebagai negara dengan jumlah penderita terbanyak. Hal ini menjadi ancaman besar bagi kesehatan masyarakat dan pelayanan kesehatan.

Diabetes melitus merupakan salah satu penyakit tidak menular yang paling mematikan, baik di tingkat nasional maupun global. Penyakit ini timbul akibat gangguan pada sistem metabolisme, khususnya dalam produksi atau efektivitas kerja insulin, yang mengakibatkan peningkatan kadar gula dalam darah (hiperglikemia). Pola hidup modern, seperti konsumsi makanan cepat saji yang tinggi serta mengurangi aktifitas fisik, menjadi faktor utama yang mempercepat peningkatan angka kejadian penyakit ini . Berdasarkan data International Federation, jumlah kasus diabetes di dunia terus mengalami peningkatan, dan indonesia tercatat berada di peringkat kelima sebagai negara dengan jumlah penderita terbanyak. Kondisi ini menunjukkan adanya tantangan serius bagi sistem kesehatan masyarakat dan pelayanan medis secara keseluruhan.

Meskipun berbagai penelitian telah dilakukan untuk membangun model prediksi penyakit diabetes berbasis machine learning, permasalahan utama yang sering ditemui adalah kurang optimalnya kinerja model dalam mendeteksi kasus positif (penderita diabetes), terutama akibat penggunaan data yang belum diproses secara tepat dan ke jumlah data antar kelas. Model yang dibor menggunakan data mentah tanpa tahapan preprocessing dan optimasi hyperparameter cenderung menghasilkan performa klasifikasi yang kurang memadai, dengan nilai *recall* dan *F1-score* pada kelas positif yang masih rendah. Selain itu, dominasi jumlah data non-diabetes dalam dataset menyebabkan model bias terhadap kelas mayoritas, yang berakibat pada rendahnya sensitivitas dalam mengenali pasien yang benar-benar mengidap diabetes.

Berdasarkan hal tersebut, diperlukan pendekatan yang lebih komprehensif dalam pengembangan model prediksi diabetes, salah satunya dengan memanfaatkan algoritma (KNN) yang dioptimalkan melalui pencarian *hyperparameter* terbaik menggunakan *GridSearchCV* serta penerapan metode (SMOTE) untuk menangani ketidakseimbangan kelas. Dengan penerapan strategi ini, model yang diharapkan mampu meningkatkan akurasi sekaligus sensitivitas dalam mendeteksi kasus positif, sehingga dapat berkontribusi dalam sistem deteksi dini yang lebih efektif dan akurat.

2.2 Pengumpulan Data

Penelitian ini memanfaatkan *Pima Indian Diabetes Dataset* , yang bersumber dari kaggle "*National Institute of Diabetes and Digestive and Kidney Diseases"*, sebuah lembaga yang memiliki reputasi tinggi dan telah diakui secara luas dalam bidang penelitian kesehatan, khususnya yang berkaitan dengan penyakit metabolik dan sistem pencernaan. Dataset ini diperoleh melalui platform berbagi data *Kaggle* , yang dikenal sebagai salah satu wadah terpercaya bagi para peneliti dan praktisi data dalam mengakses dataset publik untuk keperluan analisis dan pengembangan model prediktif. Secara keseluruhan, dataset ini terdiri dari 768 sampel data individu, di mana setiap sampel mencakup delapan atribut prediktor medis yang mencerminkan kondisi dan riwayat kesehatan pasien. Selain itu, terdapat satu variabel target bernama *Outcome* yang digunakan untuk menunjukkan status diabetes pada setiap individu, di mana nilai 1 menunjukkan bahwa pasien mengidap diabetes, dan nilai 0 menunjukkan bahwa pasien tidak mengidap diabetes. Komposisi data dalam dataset

JURIKOM (Jurnal Riset Komputer), Vol. 12 No. 3, Juni 2025
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v12i3.8615
Hal 230-240
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

ini menunjukkan distribusi kelas yang cukup seimbang, dengan 268 data pasien teridentifikasi sebagai pengidap diabetes (kelas 1), sementara 500 data lainnya berasal dari pasien yang tidak mengidap diabetes (kelas 0) [6]. Informasi terkait terkait atribut yang digunakan beserta deskripsinya ditunjukkan pada Tabel 1.

Tabel 1. Informasi Dataset

No	Atribut	Deskripsi
1.	Pregnancies	Jumlah kehamilan setiap individu
2.	Glucose	jumlah gula dalam darah seseorang
3.	Blood Pressure	Tingkat glukosa dalam plasma setelah seseorang menjalani tes toleransi glukosa.
4.	Skin Thickness	Ketebalan lipatan kulit pada trisep dalam satuan milimeter.
5.	Insulin	Kadar insulin dalam serum juga dihitung sebagai salah satu variabel yang dianalisis.
6.	BMI	Indeks massa tubuh (BMI) digunakan untuk mengukur proporsi berat badan terhadap tinggi badan
7.	Diabetes Pedigree Function	Riwayat keluarga dengan diabetes menjadi indikator tambahan
8.	Age	Usia individu dalam tahun turut diperhitungkan sebagai variabel dalam analisis
9.	Outcome	Variabel target dalam dataset ini memiliki dua nilai, yaitu 0 yang menunjukkan bahwa individu tidak mengidap diabetes, dan 1 yang menandakan bahwa individu tersebut terdiagnosis menderita diabetes.

2.3 Preprocessing Data

Preprocessing data adalah langkah penting dalam data mining yang berfungsi untuk mengubah data mentah menjadi bentuk yang lebih terstruktur dan mudah dianalisis [14]. Dalam penelitian ini, preprocessing terdiri dari beberapa tahapan yaitu :

a. *Data Cleaning*
Pembersihan data digunakan untuk menghilangkan data kosong dan nilai yang hilang, mencegah kesalahan [12].

b. *Normalisasi Data*
Dalam penelitian ini, standarisasi *min-max* juga dikenal sebagai standarisasi *min-max* digunakan untuk normalisasi data; data asli diubah secara langsung. Metode ini digunakan untuk mendapatkan nilai atribut yang lebih seimbang dan memastikan bahwa data berada dalam rentang yang tepat [12]. Selanjutnya Proses normalisasi dilakukan dengan *MinMaxScaler*, yang bertujuan menstandarkan nilai fitur agar berada di kisaran 0 sampai 1[15]. Persamaan *MinMaxScaler* tertera pada persamaan 1.

$$MinMaxScaler = \frac{x_i - x_{min}}{x_{max} - x_{min}} \quad (1)$$

Keterangan :
x_i : Nilai data ke- i yang akan dinormalisasi
x_{min} : Nilai minimum dari seluruh data pada fitur tersebut
x_{max} : Nilai maksimum dari seluruh data pada fitur

c. *Pembagian Dataset*
Setelah tahap Preprocessing data selesai, langkah selanjutnya adalah memisahkan data menjadi data latih dan data uji. Tujuan dari proses ini adalah untuk menguji model menggunakan data baru yang belum pernah dipakai sebelumnya [16].

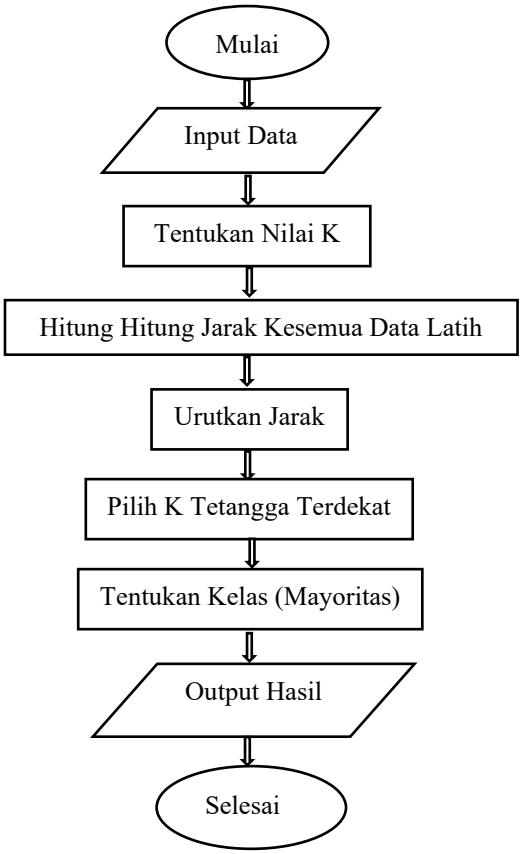
2.4 Data Processing

Pengembangan model prediksi saat ini dilakukan melalui dua metode yang berbeda dengan memanfaatkan algoritma K-Nearest Neighbors. Pada metode pertama, model dibangun menggunakan data latih tanpa melalui proses preprocessing, sehingga algoritma KNN diterapkan langsung terhadap data mentah guna melihat kinerja awalnya. Metode kedua meliputi tahapan preprocessing, termasuk penanganan data kosong, normalisasi fitur, serta penyeimbangan kelas menggunakan teknik SMOTE. Selanjutnya model dioptimalkan melalui *GridSearchCV* untuk mendapatkan kombinasi parameter yang paling efektif dalam meningkatkan kinerja. Kedua pendekatan tersebut diuji menggunakan pengujian data guna menilai sejauh mana proses preprocessing dan optimasi dapat meningkatkan akurasi akurasi serta kemampuan klasifikasi model.

a. *K-Nearest Neighbors* (KNN)
Algoritma KNN diterapkan karena memiliki sifat yang sangat nonlinier. *K-Nearest Neighbor* merupakan metode pembelajaran mesin nonparametrik yang sederhana, algoritma ini mudah dipahami serta mudah diimplementasikan [11]. Algoritma (KNN) juga dikenal memiliki keunggulan dalam menangani data pelatihan yang mengandung banyak noise serta mampu bekerja dengan baik pada dataset berukuran besar [17]. Adapun prinsip dari algoritma KNN adalah mengidentifikasi jarak terdekat antara data uji dan sejumlah *K* data pelatihan yang paling mirip atau paling dekat data

JURIKOM (Jurnal Riset Komputer), Vol. 12 No. 3, Juni 2025
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v12i3.8615
Hal 230-240
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

pelatihan yang paling mirip atau paling dekat berdasarkan jarak tertentu [18]. Adapun alur dari penerapan algoritma ini ditunjukkan pada Gambar 2.



Gambar 2. Flowchart KNN

Pada Gambar 2, proses klasifikasi menggunakan algoritma *K-Nearest Neighbors* (KNN) secara sistematis diawali dengan tahap Mulai, yang menandakan dimulainya prosedur prediksi. Sistem kemudian bersiap untuk menerima data baru yang atributnya telah diketahui namun label kelasnya belum teridentifikasi, yang menjadi objek utama klasifikasi. Algoritma selanjutnya akan menetapkan nilai K, sebuah parameter fundamental yang menentukan jumlah tetangga terdekat yang akan dijadikan referensi dalam membuat keputusan prediksi. Tahap inti melibatkan penghitungan jarak dari data baru tersebut ke setiap titik data yang sudah tersimpan dalam dataset pelatihan, menggunakan Euclidean distance untuk mengukur tingkat kemiripan atau kedekatan data. Hasil dari seluruh perhitungan jarak diurutkan dari yang paling kecil hingga paling besar, memungkinkan sistem untuk memilih K tetangga terdekat yang paling relevan. Penentuan kelas untuk data baru dilakukan berdasarkan mayoritas kelas yang dimiliki oleh K tetangga terpilih tersebut, di mana kelas yang paling sering muncul akan menjadi hasil prediksi.

b. *Synthetic Minority Oversampling Technique (SMOTE)*

Dalam dataset diabetes yang dianalisis, kelas non-diabetes merupakan kelas yang dominan, mencakup 65,1% dari total data [8]. Pada penelitian ini, SMOTE digunakan untuk menangani ketidakseimbangan kelas pada data analisis sentimen. Teknik ini bekerja dengan menghasilkan data sintetis dari kelas minoritas melalui pemanfaatan tetangga terdekat dan perbedaan vektor. Dengan metode ini, jumlah sampel dari kelas minoritas dapat diperbanyak sehingga distribusinya menjadi lebih seimbang dengan kelas mayoritas [19].

c. *GridSearchCV*

Penggunaan *Grid Search* dengan *cross-validation* (CV) cenderung menghasilkan akurasi yang lebih tinggi karena metode ini membantu menemukan kombinasi *hyperparameter* terbaik untuk model yang digunakan. *Grid Search* secara sistematis mencoba berbagai kombinasi *hyperparameter* dan memilih yang paling optimal berdasarkan matrik evaluasi seperti akurasi. Dengan menerapkan *cross-validation*, proses ini memastikan bahwa model dievaluasi secara menyeluruh pada berbagai subset data, sehingga mampu mengurangi risiko overfitting dan memberikan estimasi performa yang lebih andal untuk data yang belum pernah digunakan. Selain itu, *cross-validation* juga membantu mengurangi dampak dari ketidakseimbangan kelas atau noise dalam dataset. Dengan memvalidasi kinerja model di berbagai bagian data, *GridSearch* dengan CV memastikan bahwa hasil yang diperoleh tidak hanya baik pada satu subset, tetapi juga stabil di seluruh data. Ini menghasilkan model yang lebih general dan mampu memberikan akurasi yang lebih baik saat diuji pada data baru [20]. Dengan demikian, *GridSearchCV* berperan penting dalam menemukan parameter yang paling optimal, sehingga mampu meningkatkan model serta mendapatkan hasil yang akurat dan efisien [8].

JURIKOM (Jurnal Riset Komputer), Vol. 12 No. 3, Juni 2025
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v12i3.8615
Hal 230-240
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

2.5 Evaluasi Hasil

Pada tahap ini dilakukan analisis hasil pengujian klasifikasi menggunakan *confusion matrix* yang merepresentasikan evaluasi dari metode supervised learning. *Confusion Matrix* ini menghasilkan Nilai TP (*true positive*), TN (*true negative*), FP (*false positive),* dan FN (*false negative*) mencerminkan jumlah data yang berhasil diklasifikasikan oleh sistem. Berdasarkan nilai-nilai tersebut, dilakukan perhitungan metrik kinerja algoritma K-Nearest Neighbor menggunakan *classification* report, yang mencakup nilai *accuracy*, *precision*, *recall*, dan *F1-score* sesuai dengan rumus pada Persamaan (2), (3), (4), dan (5). Metrik ini digunakan untuk menilai seberapa baik model dalam memprediksi kondisi diabetes pada data uji.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100\% \quad (2)$$

$$Precision = \frac{TP}{TP+FN} \times 100\% \quad (3)$$

$$Recall = \frac{TP}{TP+FP} \times 100\% \quad (4)$$

$$F1-Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \times 100 \quad (5)$$

3. HASIL DAN PEMBAHASAN

Penelitian ini bertujuan untuk melakukan analisis komparatif terhadap kinerja model KNN yang dibangun dengan dua pendekatan berbeda yaitu KNN berdasarkan pengolahan data dan konfigurasi hyperparameter. Model pertama tanpa penyeimbangan data SMOTE maupun optimasi *GridSearchCV*, yakni *n_neighbors*=5, *weights='uniform'*, dan *metric='euclidean'*. Karena tidak ada penyesuaian terhadap distribusi kelas, model ini berisiko bias terhadap kelas mayoritas.Sebaliknya, model kedua menerapkan SMOTE untuk menyeimbangkan data dan *GridSearchCV* untuk menemukan kombinasi hyperparameter terbaik, yaitu *n_neighbors*=3, *weights='distance'*, dan *metric='manhattan'*. Konfigurasi ini meningkatkan kinerja model karena lebih fleksibel terhadap data seimbang dan mempertimbangkan jarak antar tetangga secara seimbang . Jadi secara keseluruhan, hasil pada Tabel 2 menunjukkan bahwa penggunaan teknik SMOTE dan optimasi hyperparameter dapat secara signifikan meningkatkan dan keseimbangan klasifikasi dibandingkan model dasar tanpa penyesuaian.

Tabel 2. Model dan Hyperparameter

No	Model	Hyperparameter
	KNN dengan data tidak seimbang (tanpa SMOTE) & tanpa *GridSearchCV*	*n_neighbors=5, weights='uniform', metric ='euclidean'*
	KNN dengan data yang seimbang (SMOTE) & *GridSearchCV*	*n_neighbors=3, weights='distance', metric='manhattan'*

Dari hasil evaluasi yang dirangkum dalam Tabel 2, menunjukkan bahwa strategi pengolahan data yang seimbang dan pemilihan *hyperparameter* yang tepat mampu meningkatkan kinerja model secara keseluruhan dibandingkan metode KNN dengan data tidak seimbang (tanpa SMOTE) dan *GridSearchCV*.

3.1 *Hasil Preprocessing Data*

Eksperimen diawali dengan proses preprocessing data, yang mencakup beberapa langkah utama. Langkah pertama adalah membersihkan data, yaitu dengan memeriksa adanya nilai yang hilang (missing values) dalam dataset. Tabel 3 menyajikan sebagian data awal sebelum dilakukan tahap preprocessing. Pada tahap ini, proses pembersihan data menjadi langkah krusial, terutama dalam mengidentifikasi nilai-nilai yang tidak logis atau dianggap hilang (missing values). Beberapa fitur seperti *Skin Thickness* dan *Insulin* menunjukkan nilai 0, yang secara medis tidak realistis dan oleh karena itu diperlakukan sebagai nilai hilang. Nilai-nilai tersebut akan ditangani pada tahap selanjutnya untuk memastikan kualitas data yang digunakan dalam pelatihan model. Pembacaan awal terhadap distribusi data ini juga memberikan gambaran bahwa terdapat ketidakseimbangan antara jumlah penderita diabetes (kelas 1) dan non-diabetes (kelas 0), yang nantinya akan diatasi dengan metode penyeimbangan data seperti SMOTE. Dengan demikian, tahap awal ini sangat penting untuk memastikan bahwa data yang digunakan benar-benar bersih dan representatif sebelum memasuki proses modeling. Adapun data awal yang digunakan pada penelitian ini ditunjukkan pada Tabel 3.

Tabel 3. Data Awal

No	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age	Out come
1.	6	148	72	35	0	33.6	0.632	50	1
2.	1	85	66	29	0	26.6	0.351	31	0

JURIKOM (Jurnal Riset Komputer), Vol. 12 No. 3, Juni 2025
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v12i3.8615
Hal 230-240
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

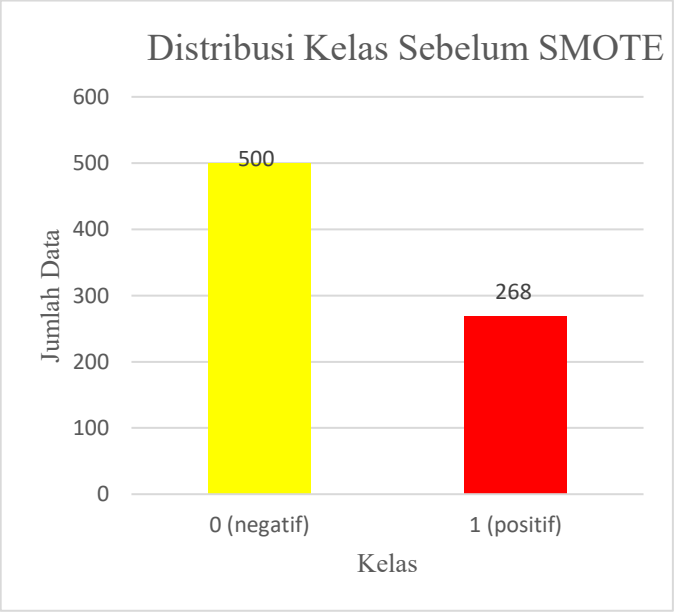
3.	8	183	64	0	0	23.3	0.772	32	1
4.	1	89	66	23	94	28.1	0.167	21	0
5.	0	137	40	35	168	43.1	2.288	33	1

Pada tahap selanjutnya, dataset yang telah dinormalisasi dibagi menjadi dua subset, yaitu data latih dan data uji, dengan proporsi 80:20. Dari proses ini, diperoleh sebanyak 800 data untuk keperluan pelatihan dan 200 data untuk pengujian. Pembagian ini bertujuan untuk memisahkan data yang digunakan dalam membangun model dari data yang digunakan untuk mengevaluasi performa model secara objektif. Namun, berdasarkan analisis awal terhadap distribusi kelas, ditemukan adanya ketidakseimbangan antara jumlah data pada masing-masing kelas, yang berpotensi menyebabkan bias dalam proses pelatihan model. Untuk mengatasi permasalahan ini, digunakan metode SMOTE sebagai solusi dalam menyeimbangkan distribusi kelas pada data pelatihan. Penerapan SMOTE dilakukan secara eksklusif pada data latih agar data uji tetap merepresentasikan distribusi kelas asli dari dataset, sehingga evaluasi model dapat mencerminkan performa yang realistis dalam kondisi dunia nyata. SMOTE bekerja dengan cara menghasilkan sampel sintetis baru untuk kelas minoritas, bukan melalui penduplikasian data yang sudah ada, melainkan dengan membentuk data baru berdasarkan interpolasi dari tetangga terdekat dalam ruang fitur. Dengan demikian, teknik ini mampu meningkatkan keberagaman data minoritas dan membantu model belajar lebih baik dalam mengenali pola dari kedua kelas secara seimbang. Adapun dataset setelah normalisasi dan smote ditunjukkan pada Tabel 4.

Tabel 4. Dataset Setelah Normalisasi dan SMOTE

Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age
0.200	0.914	0.649	0.000	0.000	0.454	0.114	0.156
0.200	0.618	0.877	0.555	0.352	0.853	0.342	0.019
0.066	0.613	0.561	0.507	0.229	0.523	0.262	0.176
0.533	0.773	0.684	0.507	0.000	0.482	0.155	0.470
0.200	0.753	0.666	0.000	0.000	0.321	0.055	0.313

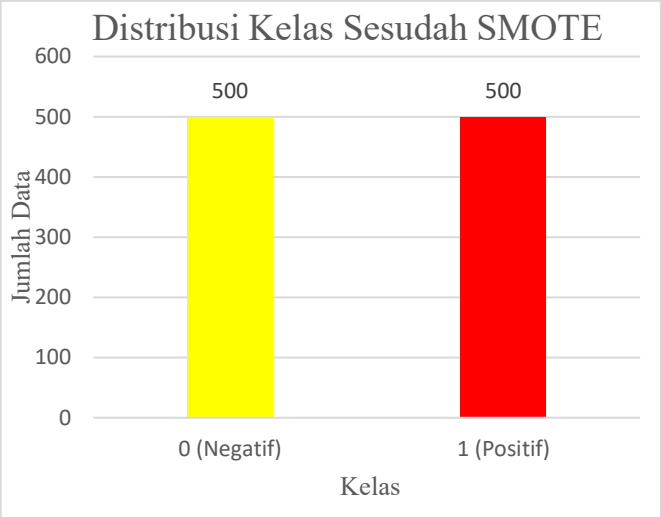
Distribusi jumlah data yang jelas menunjukkan perbedaan signifikan sebelum dan setelah proses SMOTE, distribusi kelas sebelum SMOTE ditunjukkan pada Gambar 3.



Gambar 3. Distribusi Kelas Sebelum SMOTE

Dan Gambar 4, dimana terlihat bahwa setelah pengaplikasian metode SMOTE jumlah data pada masing-masing kelas berhasil menjadi seimbang.

JURIKOM (Jurnal Riset Komputer), Vol. 12 No. 3, Juni 2025
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v12i3.8615
Hal 230-240
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom



Gambar 4. Distribusi Kelas Sesudah SMOTE

Gambar 3 dan 4 menyajikan perbandingan distribusi kelas dalam dataset sebelum dan setelah penerapan teknik SMOTE. Pada gambar 3 terlihat bahwa distribusi kelas awal menunjukkan ketidakseimbangan, dengan jumlah sampel pada kelas 0 jauh lebih banyak dibandingkan kelas 1. Setelah dilakukan penyeimbangan menggunakan SMOTE, sebagaimana ditunjukkan pada gambar 4, jumlah sampel untuk kedua kelas menjadi seimbang. Penyeimbangani ini bertujuan untuk meminimalkan potensi bias yang dapat mempengaruhi kinerja model dalam proses pembelajaran mesin. Perbandingan jumlah data pada masing-masing kelas sebelum dan sesudah penerapan teknik SMOTE. Sebelum SMOTE diterapkan, jumlah data pada kelas 0 (Negatif) sebanyak 500, sedangkan kelas 1 (Positif) hanya sebanyak 268, yang menunjukkan ketidakseimbangan data. Setelah SMOTE diterapkan, jumlah data pada kelas 1 (Positif) ditingkatkan menjadi 500, sehingga kedua kelas memiliki jumlah data yang seimbang. Teknik ini digunakan untuk mengatasi masalah data tidak seimbang yang dapat mempengaruhi kinerja model dalam klasifikasi.

3.2 *Hasil Pengujian*

Pengujian ini bertujuan untuk membandingkan hasil antar model KNN dengan data tidak seimbang (tanpa SMOTE) dan tanpa *GridSearchCV* dengan model KNN dengan data yang seimbang (SMOTE) dan *GridSearchCV*. Selain itu, dilakukan pula pencarian konfigurasi *hyperparameter* terbaik untuk meningkatkan tingkat *accuracy*, dengan memanfaatkan metode *GridSearchCV*. Adapun hasil pengujian model ditunjukkan pada Tabel 5.

Tabel 5. Hasil Pengujian Model

Model	Kelas	Accuracy	Precision	Recall	F1-Score
KNN dengan data tidak seimbang (tanpa SMOTE) & tanpa *GridSearchCV*	(0) Negatif	83%	84%	91%	88%
	(1)Positif		80%	69%	74%
KNN dengan data yang seimbang (SMOTE) & *GridSearchCV*	(0) Negatif		98%	89%	93%
	(1) Positif	94%	90%	98%	94%

Tabel 5, menunjukkan bahwa performa model *K-Nearest Neighbors* (KNN) dalam dua kondisi berbeda, yaitu saat menggunakan data yang tidak seimbang dan saat menggunakan data yang telah diseimbangkan serta dioptimasi dengan *GridSearchCV*. Pada model KNN dengan data yang tidak seimbang, *accuracy* yang dicapai sebesar 83%. Meskipun nilai ini terlihat cukup baik, namun performa terhadap kelas positif masih rendah, dengan *precision* sebesar 80%, *recall* hanya 69%, dan *F1-score* 74%, sedangkan untuk kelas negatif dengan *precision sebesar* 84%, *recall* 91%, dan *F1-score* 88% . Hal ini menunjukkan bahwa model cenderung bias terhadap kelas negatif, yang umum terjadi ketika jumlah data antar kelas tidak seimbang. Sebaliknya, setelah dilakukan penyeimbangan data dan optimasi parameter menggunakan *GridSearchCV*, performa model meningkat secara signifikan, a*ccuracy* keseluruhan naik menjadi 94%, dan performa terhadap kelas positif menunjukkan perbaikan besar dengan *precision* 90%, *recall* 98%, dan F1-*score* mencapai 94%, sedangkan kelas negatif dengan *precision sebesar* 98%, *recall* 89%, dan *F1-score* 93%.Perubahan ini menunjukkan bahwa proses penyeimbangan data dan tuning parameter model dapat meningkatkan kemampuan model dalam mengenali kedua kelas secara seimbang, terutama dalam mendeteksi kasus positif yang sangat krusial dalam konteks prediksi penyakit seperti diabetes.

3.3 *Evaluasi Hasil Model*

Hasil evaluasi kinerja model *K-Nearest Neighbors* (KNN) yang dilakukan secara cermat telah memberikan gambaran yang jelas mengenai dampak signifikan dari serangkaian proses optimasi. Dapat diamati bahwa sebelum metode penyeimbangan data SMOTE diterapkan dan optimasi *hyperparameter* dilakukan, model KNN menghasilkan *accuracy* sebesar 83%, ini menunjukkan bahwa kinerja dasar model dalam mengklasifikasikan data sebelum adanya ruang untuk

JURIKOM (Jurnal Riset Komputer), Vol. 12 No. 3, Juni 2025
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v12i3.8615
Hal 230-240
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

peningkatan, khususnya pada aspek *recall* yang menunjukkan kemampuan model dalam menangkap semua kasus positif. Namun, setelah proses penyeimbangan data yang efektif dengan SMOTE berhasil diimplementasikan untuk mengatasi ketidakseimbangan kelas, dan diikuti dengan optimasi *hyperparameter* secara sistematis menggunakan *GridSearchCV* untuk menemukan konfigurasi terbaik, kinerja model meningkat secara dramatis. Peningkatan ini terbukti dengan *accuracy* dan *F1-score* yang keduanya mencapai angka impresif sebesar 94%. Peningkatan kinerja ini terbukti bahwa model tidak hanya menjadi jauh lebih akurat dalam prediksi secara keseluruhan, tetapi juga lebih seimbang dalam mengklasifikasikan pasien yang menderita maupun tidak menderita diabetes, menunjukkan kemampuan generalisasi yang jauh lebih baik pada data yang belum pernah dilihat sebelumnya. Untuk visualisasi detail mengenai distribusi hasil prediksi dan perbandingan kinerja model sebelum SMOTE dan optimasi tersebut, termasuk informasi mengenai *True Positives, True Negatives, False Positives*, dan *False Negatives*, secara lengkap ditunjukkan pada Gambar 5, yang menjadi bukti empiris dari efektivitas metode yang diterapkan.

Confusion Matrix Sebelum SMOTE dan Optimasi

Aktual \ Prediksi	Negatif	Positif
Negatif	91	9
Positif	17	37

Gambar 5. *Confusion Matrix* Sebelum Smote dan Optimasi

Pada Gambar 5 merupakan *Confusion Matrix* yang menunjukkan bahwa sebelum dilakukan *Smote* dan optimasi menggunakan metode *GridSearchCV*, model KNN berhasil mengklasifikasikan 91 data negatif dan 37 data positif dengan benar. Namun model tersebut masih salah mengklasifikasikan 9 data negatif sebagai positif (*false positif*) dan 17 data positif sebagai negatif (*false negative*). Hal ini menghasilkan *accuracy* sebesar 83%, dengan *precision* sebesar 80%, *recall* hanya 69%, dan *F1-score* 74% untuk kelas positif diabetes sedangkan untuk kelas negatif dengan *precision sebesar* 84%, *recall* 91%, dan *F1-score* 88% . Artinya, meskipun model cukup baik dalam mengenali pasien yang tidak menderita diabetes, model masih cukup lemah dalam mendeteksi pasien yang sebenarnya menderita diabetes. Kondisi ini dapat berisiko dalam praktik nyata, karena kesalahan dalam mendeteksi penderita (*false negative*) dapat menyebabkan pasien tidak mendapat penanganan yang tepat waktu. Oleh karena itu, diperlukan strategi seperti penyeimbangan kelas menggunakan SMOTE dan optimasi parameter untuk meningkatkan kemampuan deteksi terhadap penderita diabetes secara lebih akurat dan andal. Untuk *Confusion Matrix* sesudah SMOTE dan optimasi ditunjukkan pada Gambar 6.

Confusion Matrix Susudah SMOTE dan Optimasi

Aktual \ Prediksi	Negatif	Positif
Negatif	89	11
Positif	2	98

Gambar 6. *Confusion Matrix* Setelah SMOTE dan optimasi

JURIKOM (Jurnal Riset Komputer), Vol. 12 No. 3, Juni 2025
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v12i3.8615
Hal 230-240
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

Gambar 6 menampilkan *Confusion Matrix* setelah penerapan teknik SMOTE dan optimasi *hyperparameter*, yang menunjukkan adanya peningkatan signifikan dalam kinerja model. Dari seluruh prediksi yang dilakukan, model berhasil mengklasifikasikan dengan benar 89 data dari kelas negatif dan 98 data dari kelas positif. Terdapat 11 data negatif yang salah terklasifikasi sebagai positif, serta 2 data positif yang keliru diklasifikasikan sebagai negatif. Hasil ini menunjukkan bahwa model memiliki kinerja yang baik dan seimbang dalam mengidentifikasi kedua kelas. Selain itu, model juga menunjukkan keandalan yang lebih tinggi dalam mendeteksi pasien yang benar benar menderita diabetes, berkat penanganan ketidakseimbangan data melalui SMOTE dan optimasi *hyperparameter* menggunakan *GridSearchCV*

Berdasarkan tabel 6, hasil dari proses *GridSearchCV* menunjukkan bahwa kombinasi *hyperparameter* terbaik untuk algoritma KNN diperoleh dengan nilai *n_neighbors* sebesar 3, *weights* menggunakan metode *distance*, dan *metric* jarak yang digunakan adalah *manhattan*. Hal ini berarti bahwa dalam melakukan klasifikasi, model KNN akan memperhitungkan tiga tetangga terdekat, di mana kontribusi setiap tetangga ditentukan oleh tingkat kedekatannya semakin dekat jaraknya, semakin besar pengaruhnya terhadap hasil prediksi. Pemilihan *metric Manhattan* sebagai pengukur jarak mengindikasikan bahwa selisih *absolute* antar fitur lebih relevan dibandingkan penggunaan jarak *euclidean* dalam konteks ini. Melalui validasi silang sebanyak lima lipatan (*5-fold cross-validation*), model ini mampu menghasilkan nilai *accuracy* dan *F1-score* yang sama tinggi, yaitu sebesar 94% dan performa terhadap kelas positif menunjukkan perbaikan besar dengan *precision* 90%, *recall* 98%, dan F1-*score* mencapai 94%, sedangkan kelas negatif dengan *precision sebesar* 98%, *recall* 89%, dan *F1-score* 93%. . Ini menunjukkan bahwa model memiliki performa yang optimal dan seimbang dalam mengklasifikasikan kedua kelas, baik positif maupun negatif, terlebih mengingat distribusi awal data yang tidak seimbang. Meskipun demikian, perlu ditekankan bahwa nilai ini diperoleh dari evaluasi terhadap data pelatihan, sehingga pengujian pada data uji tetap menjadi acuan utama untuk menilai efektivitas akhir dari model. Adapun hasil pengujian model *GridSearchCV* untuk kelas positif ditunjukkan pada Tabel 6.

Tabel 6 . Hasil Pengujian Model GridSearchCV Kelas Positif

Komponen	Nilai
n_neighbors	3
Weights	Distance
Metric	Manhattan
Accuracy	94%
F1 Score	94%

4. KESIMPULAN

Berdasarkan hasil evaluasi, dapat disimpulkan bahwa pelaksanaan prosedur-prosedur krusial seperti pembersihan data, normalisasi dengan *MinMaxScaler*, penyeimbangan data menggunakan metode *Synthetic Minority Over-sampling Technique* (SMOTE), dan optimasi parameter memberikan pengaruh yang sangat besar terhadap peningkatan kinerja model KNN secara umum. Model *K-Nearest Neighbors* (KNN) dalam dua kondisi berbeda, yaitu saat menggunakan data yang tidak seimbang dan menggunakan data yang telah diseimbangkan serta dioptimasi dengan *GridSearchCV*. Pada model KNN dengan data yang tidak seimbang, *accuracy* yang dicapai sebesar 83%. Meskipun nilai ini terlihat cukup baik, namun performa terhadap kelas positif masih rendah, dengan *precision* sebesar 80%, *recall* hanya 69%, dan *F1-score* 74%, sedangkan untuk kelas negatif dengan *precision sebesar* 84%, *recall* 91%, dan *F1-score* 88% . Hal ini menunjukkan bahwa model cenderung bias terhadap kelas negatif, yang umum terjadi ketika jumlah data antar kelas tidak seimbang. Sebaliknya, setelah dilakukan penyeimbangan data dan optimasi parameter menggunakan *GridSearchCV*, performa model meningkat secara signifikan. *Accuracy* keseluruhan naik menjadi 94%, dan performa terhadap kelas positif menunjukkan perbaikan besar dengan *precision* 90%, *recall* 98%, dan F1-*score* mencapai 94%, sedangkan kelas negatif dengan *precision sebesar* 98%, *recall* 89%, dan *F1-score* 93%.. Penggunaan metode SMOTE terbukti berhasil dalam mengatasi masalah ketidakseimbangan kelas pada dataset, sehingga model dapat belajar dengan lebih adil terhadap kedua kelas, baik positif maupun negatif. Selain itu, *GridSearchCV* berhasil menemukan kombinasi *hyperparameter* optimal, yaitu jumlah tetangga terdekat (*n_neighbors*) sebanyak 3, bobot (*weights*) berdasarkan jarak, serta penggunaan *metric manhattan* untuk mengukur kedekatan antar data. Kombinasi ini terbukti menghasilkan klasifikasi yang paling optimal pada model. Dengan demikian, disimpulkan bahwa penelitian ini menekankan signifikansi penerapan strategi preprocessing data yang sesuai serta pemilihan *hyperparameter* yang optimal dalam meningkatkan performa dan akurasi model prediktif. Hal ini menunjukkan bahwa proses penyeimbangan data dan tuning parameter model dapat meningkatkan kemampuan model dalam mengenali kedua kelas secara seimbang, terutama dalam mendeteksi kasus positif yang sangat krusial dalam konteks prediksi penyakit seperti diabetes.

REFERENCES

[1] I. D. A. E. C. Astutisari, A. Y. D. AAA Yuliati Darmini, and I. A. P. W. Ida Ayu Putri Wulandari, "Hubungan Pola Makan Dan Aktivitas Fisik Dengan Kadar Gula Darah Pada Pasien Diabetes Melitus Tipe 2 Di Puskesmas Manggis I," *Jurnal Riset Kesehatan Nasional*, vol. 6, no. 2, pp. 79–87, 2022, doi: 10.37294/jrkn.v6i2.350.

[2] A. Hamid and Hamdin, "Hidup Sehat Produktif Bebas Diabetes dengan Edukasi Kesehatan Tentang Penyakit Diabetes Melitus di Wilayah Kerja PKM Unit 1 Sumbawa Healthy Productive Life Free from Diabetes with Health Education about Diabetes Mellitus in the PKM Unit 1 Sumbawa Working," vol. 1, no. 3, 2024.
[3] Kementerian Kesehatan Republik Indonesia, "Review kebijakan Diabetes Melitus berbasis transformasi sistem Kesehatan dan outlook 2025," 2025.
[4] E. E. Mustofa, J. Purwono, and Ludiana, "Penerapan Senam Kaki Terhasap Kadar Glukosa Darah Pada Pasien Diabetes Melitus Di WIlayah Kerja Puskesmas Purwosari Kec. Metro Utara," *Jurnal Cendikia Muda*, vol. 2, no. 1, pp. 78–86, 2022.
[5] A. M. Argina, "Penerapan Metode Klasifikasi K-Nearest Neigbor pada Dataset Penderita Penyakit Diabetes," *Indonesian Journal of Data and Science*, vol. 1, no. 2, pp. 29–33, 2020, doi: 10.33096/ijodas.v1i2.11.
[6] S. E. Hartono, "Hubungan Tingkat Pendidikan, Lama Menderita Sakit Dengan Tingkat Pengetahuan 5 Pilar Penatalaksanaan Diabetes Mellitus Di Wilayah Kerja Puskesmas Sungai Durian Kabupaten Kbu Raya Kalimantan Barat," *Journal of TSCSIKep*, vol. 9, no. 1, pp. 2775–0345, 2024, [Online]. Available: http://ejournal.annurpurwodadi.ac.id/index.php/TSCS1Kep
[7] S. S. Fandinata and R. Darmawan, "Pengaruh Kepatuhan Minum Obat Oral Anti Diabetik Terhadap Kadar Gula Darah Pada Pasien Diabetes Mellitus Tipe II," *Jurnal Bidang Ilmu Kesehatan*, vol. 10, no. 1, pp. 23–31, 2020, doi: 10.52643/jbik.v10i1.825.
[8] A. Yaqin, D. Kurniawan, and J. Zeniarja, "Optimasi Algoritma K-Nearest Neighbors Menggunakan GridSearchCV untuk Klasifikasi Penyakit Diabetes," vol. 16, no. 01, pp. 75–84, 2025, doi: 10.35970/infotekmesin.v16i1.2557.
[9] A. Oktaviana, D. P. Wijaya, A. Pramuntadi, and D. Heksaputra, "Prediksi Penyakit Diabetes Melitus Tipe 2 Menggunakan Algoritma K-Nearest Neighbor (K-NN)," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 4, no. 3, pp. 812–818, 2024, doi: 10.57152/malcom.v4i3.1268.
[10] N. Maulidah, R. Supriyadi, D. Y. Utami, F. N. Hasan, A. Fauzi, and A. Christian, "Prediksi Penyakit Diabetes Melitus Menggunakan Metode Support Vector Machine dan Naive Bayes," *Indonesian Journal on Software Engineering (IJSE)*, vol. 7, no. 1, pp. 63–68, 2021, doi: 10.31294/ijse.v7i1.10279.
[11] J. Lemantara and T. Lusiani, "Analisis Prediksi Penyakit Diabetes Pada Wanita Menggunakan Metode Naïve Bayes Dan K-Nearest Neighbor," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 12, no. 3, 2024, doi: 10.23960/jitet.v12i3.4911.
[12] R. A. Pangestu, T. Taslim, Y. Yunefri, K. Kursiasih, and E. Sabna, "Optimasi Nilai k Pada Algoritma K-Nearest Neighbor Untuk Klasifikasi Pasien Covid-19 Yang Membutuhkan Ruangan ICU," *INOVTEK Polbeng - Seri Informatika*, vol. 7, no. 1, p. 147, 2022, doi: 10.35314/isi.v7i1.2481.
[13] E. Safitri, D. Rofianto, N. Purwati, H. Kurniawan, and S. Karnila, "Prediksi Penyakit Diabetes Melitus Menggunakan Algoritma Machine Learning Diabetes Mellitus Disease Prediction using Machine Learning Algorithms," vol. 12, no. 4, pp. 760–766, 2024, doi: 10.26418/justin.v12i4.84620.
[14] A. Anggrawan and M. Mayadi, "Application of KNN Machine Learning and Fuzzy C-Means to Diagnose Diabetes," *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 22, no. 2, pp. 405–418, 2023, doi: 10.30812/matrik.v22i2.2777.
[15] H. Syafwan, F. Siagian, P. Putri, M. Handayani, S. H. Tinggi Manajemen Informatika dan Komputer Royal Jln M Yamin No, and S. Utara, "Forecasting Jumlah Pengangguran Di Kabupaten Asahan Menggunakan Metode Weighted Moving Average," *Jurnal Teknik Informatika Kaputama (JTIK)*, vol. 5, no. 2, pp. 224–229, 2021.
[16] B. F. Rochman, A. Rahim, and T. A. Y. Siswa, "Optimasi Algoritma KNN dengan Parameter K dan PSO Untuk Klasifikasi Status Gizi Balita," *Jurnal Media Informatika Budidarma*, vol. 8, no. 3, p. 1609, 2024, doi: 10.30865/mib.v8i3.7841.
[17] S. R. Cholil, T. Handayani, R. Prathivi, and T. Ardianita, "Implementasi Algoritma Klasifikasi K-Nearest Neighbor (KNN) Untuk Klasifikasi Seleksi Penerima Beasiswa," *IJCIT (Indonesian Journal on Computer and Information Technology)*, vol. 6, no. 2, pp. 118–127, 2021, doi: 10.31294/ijcit.v6i2.10438.
[18] M. Rahmadiah and P. Suparman, "Penerapan Metode K-Nearest Neighbour Untuk Sistem Penentuan Peminjaman Modal Nasabah Bank Syariah Indonesia Cabang Cikarang Berbasis Website," *Jurnal informasi dan Komputer*, vol. 10, no. 2, pp. 189–197, 2022.
[19] A. Surya Firmansyah, A. Aziz, and M. Ahsan, "Optimasi K-Nearest Neighbor Menggunakan Algoritma Smote Untuk Mengatasi Imbalance Class Pada Klasifikasi Analisis Sentimen," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 7, no. 6, pp. 3341–3347, 2024, doi: 10.36040/jati.v7i6.7257.
[20] L. Trihardianingsih and G. S. Lasatira, "Optimasi Hyperparameter GridSearchCV pada Klasifikasi Kualitas Udara menggunakan Support Vector Machine," vol. 1, no. 2, pp. 40–47, 2024.