Journal of Information System Research (JOSH)
Volume 6, No. 4, Juli 2025, pp 2238−2247
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i4.7237

Klasifikasi Risiko Diabetes Mellitus Menggunakan K-Nearest Neighbors dengan Peningkatan Performa Melalui Teknik Oversampling ADASYN

Muhammad Bagir[1,*], Hendra Mayatopani[2], Umbar Riyanto[3], Dedy Alamsyah[3]

[1]Program Studi Sistem Informasi, Sekolah Tinggi Teknologi Informasi NIIT, Jakarta
Jl. Asem II No.22, Cipete Sel., Kec. Cilandak, Kota Jakarta Selatan, Daerah Khusus Ibukota Jakarta, Indonesia
[2]Program Studi Sistem Informasi, Fakultas Sains dan Teknologi, Universitas Pradita, Tangerang
Jl. Gading Serpong Boulevard No.1, Curug Sangereng, Kec. Kelapa Dua, Kabupaten Tangerang, Banten, Indonesia
[3]Program Studi Teknik Informatika, Fakultas Teknik, Universitas Muhammadiyah Tangerang, Tangerang
Jl. Perintis Kemerdekaan I No.33, Babakan, Cikokol, Kec. Tangerang, Kota Tangerang, Banten, Indonesia
Email: [1,*]bagir.ui@gmail.com, [2]hendra.mayatopani@pradita.ac.id, [3]umbar@ft-umt.ac.id,[4]dedy.alamsyah@umt.ac.id
Email Penulis Korespondensi: bagir.ui@gmail.com
Submitted: **29/04/2025;** Accepted: **31/07/2025;** Published: **31/07/2025**

Abstrak–Diabetes mellitus merupakan penyakit metabolik kronis yang prevalensinya terus meningkat secara global. Deteksi dini terhadap risiko diabetes sangat penting untuk mengurangi komplikasi jangka panjang dan beban biaya perawatan. Namun, tantangan utama dalam penerapan model machine learning pada data medis adalah ketidakseimbangan distribusi kelas, yang dapat menyebabkan bias model terhadap kelas mayoritas. Penelitian ini bertujuan mengembangkan model klasifikasi risiko Diabetes Mellitus dengan mengintegrasikan algoritma K-Nearest Neighbors (KNN) dan teknik Adaptive Synthetic Sampling (ADASYN) untuk menangani masalah ketidakseimbangan data. Dataset yang digunakan berasal dari platform Kaggle, berisi 2000 sampel pasien dengan sembilan fitur prediktor. Pra-pemrosesan data dilakukan melalui imputasi nilai hilang, penanganan outlier menggunakan winsorizing, dan normalisasi menggunakan StandardScaler. Teknik ADASYN diterapkan untuk menghasilkan sampel sintetis adaptif pada data minoritas, dan model KNN dilatih serta dievaluasi menggunakan confusion matrix, precision, recall, F1-Score, accuracy, dan ROC-AUC. Hasil penelitian menunjukkan bahwa penerapan ADASYN meningkatkan nilai ROC-AUC sebesar 5,48% (dari 91,34% menjadi 96,82%) dan akurasi sebesar 2,50% (dari 81,50% menjadi 84,00%). F1-Score untuk kelas Diabetes juga meningkat sebesar 0,40%. Integrasi KNN dan ADASYN terbukti meningkatkan performa model dalam mendeteksi pasien berisiko diabetes serta memperbaiki sensitivitas terhadap kelas minoritas.

Kata Kunci: Adaptive Synthetic Sampling (ADASYN); Diabetes Mellitus; K-Nearest Neighbors (KNN); Ketidakseimbangan Data; Klasifikasi Risiko; Machine Learning

Abstract–Diabetes mellitus is a chronic metabolic disease with a continuously increasing global prevalence. Early detection of diabetes risk is crucial to reduce long-term health complications and the associated healthcare costs. However, a major challenge in applying machine learning models to medical data is the issue of class imbalance, which can lead to model bias toward the majority class. This study aims to develop a diabetes risk classification model by integrating the K-Nearest Neighbors (KNN) algorithm with the Adaptive Synthetic Sampling (ADASYN) technique to address the class imbalance problem. The dataset used was obtained from the Kaggle platform, containing 2,000 patient samples with nine predictive features. Data preprocessing was performed through missing value imputation, outlier handling using winsorizing, and feature normalization using StandardScaler. ADASYN was applied to generate adaptive synthetic samples for the minority class, and the KNN model was trained and evaluated using confusion matrix, precision, recall, F1-Score, accuracy, and ROC-AUC metrics. The results indicate that the implementation of ADASYN improved the ROC-AUC Score by 5.48% (from 91.34% to 96.82%) and the overall accuracy by 2.50% (from 81.50% to 84.00%). The F1-Score for the Diabetes class also increased by 0.40%. The integration of KNN and ADASYN has proven effective in enhancing model performance for detecting high-risk diabetes patients and improving sensitivity toward the minority class.

Keywords: Adaptive Synthetic Sampling (ADASYN); Diabetes Mellitus; K-Nearest Neighbors (KNN); Data Imbalance; Risk Classification; Machine Learning

1. PENDAHULUAN

Diabetes mellitus merupakan penyakit metabolik kronis yang menjadi salah satu penyebab utama kematian di dunia. Menurut laporan International Diabetes Federation (IDF), lebih dari 537 juta orang dewasa hidup dengan diabetes pada tahun 2021, dan jumlah ini diproyeksikan akan terus meningkat dalam beberapa dekade mendatang [1]. Deteksi dini terhadap risiko diabetes sangat penting untuk memungkinkan intervensi medis secara tepat waktu, sehingga dapat mengurangi komplikasi kesehatan jangka panjang dan menekan beban biaya perawatan [2]. Dalam mendukung deteksi dini tersebut, berbagai pendekatan berbasis machine learning telah dikembangkan untuk membangun model prediktif yang mampu mengklasifikasikan individu berdasarkan tingkat risiko diabetes.

Namun demikian, salah satu tantangan utama dalam penerapan machine learning pada data medis adalah ketidakseimbangan distribusi kelas (imbalanced data), di mana jumlah sampel non-diabetes jauh lebih banyak dibandingkan sampel diabetes. Ketidakseimbangan ini menyebabkan model klasifikasi cenderung bias terhadap kelas mayoritas, sehingga mengurangi akurasi deteksi terhadap pasien yang berisiko [3]. Oleh karena itu, diperlukan metode yang tidak hanya mampu mengklasifikasikan secara akurat, tetapi juga efektif dalam menangani ketidakseimbangan distribusi data.

Journal of Information System Research (JOSH)
Volume 6, No. 4, Juli 2025, pp 2238−2247
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i4.7237

Berbagai penelitian sebelumnya telah berkontribusi dalam membangun model klasifikasi risiko Diabetes Mellitus. Maulidah et al. (2021) membandingkan kinerja Support Vector Machine (SVM) dengan Naive Bayes, dengan hasil bahwa SVM memperoleh akurasi lebih tinggi sebesar 78,04% dibandingkan dengan Naive Bayes yang mencapai 76,98% [4]. Penelitian oleh Abdurrahman (2022) menggunakan algoritma Adaboost Classifier, dengan hasil akurasi sebesar 80,09% pada dataset setelah imputasi mean, 76,19% setelah imputasi median, dan akurasi yang sama dengan dataset default setelah imputasi modus [5]. Safitri dan Fatah (2023) menerapkan algoritma Decision Tree dan memperoleh akurasi sebesar 65,06%, dengan precision sebesar 32,53% dan recall sebesar 50,00% [6]. Penelitian oleh Nainggolan dan Sinaga (2023) membandingkan algoritma Random Forest dan Gradient Boosting Classifier, yang masing-masing mencapai akurasi sebesar 79% dan 81% [7]. Sementara itu, Putri dan Alit (2024) mengusulkan model Support Vector Machine (SVM) yang menghasilkan akurasi sebesar 70,78% [8].

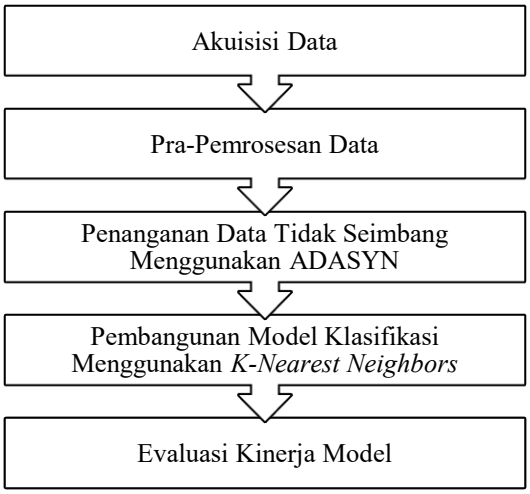
Meskipun hasil-hasil tersebut menunjukkan performa yang kompetitif, masih terdapat sejumlah keterbatasan yang menjadi dasar penelitian ini. Sebagian besar studi sebelumnya belum secara eksplisit menangani masalah ketidakseimbangan data, sehingga menurunkan sensitivitas model terhadap deteksi pasien diabetes. Untuk itu, penelitian ini mengusulkan integrasi algoritma K-Nearest Neighbors (KNN) dengan teknik oversampling Adaptive Synthetic Sampling (ADASYN) untuk meningkatkan performa klasifikasi risiko diabetes pada data yang tidak seimbang. Algoritma KNN dipilih karena kesederhanaannya serta kemampuannya dalam mengklasifikasikan data berdasarkan kedekatan antar sampel [9]. Meskipun demikian, KNN memiliki kelemahan berupa sensitivitas terhadap ketidakseimbangan kelas, di mana data mayoritas dapat mendominasi hasil klasifikasi, serta kinerjanya sangat bergantung pada distribusi data latih yang tersedia [10]. Sementara itu, ADASYN merupakan metode oversampling berbasis adaptif yang secara otomatis menghasilkan data sintetis di sekitar sampel minoritas yang sulit diklasifikasikan [11]. Keunggulan utama ADASYN dibandingkan metode oversampling konvensional adalah kemampuannya beradaptasi terhadap kompleksitas distribusi data minoritas, sehingga tidak hanya menyeimbangkan jumlah data antar kelas, tetapi juga memperkuat representasi lokal dari kelas minoritas [12].

Berdasarkan latar belakang tersebut, penelitian ini bertujuan untuk mengembangkan model klasifikasi risiko diabetes mellitus melalui integrasi algoritma K-Nearest Neighbors (KNN) dan teknik Adaptive Synthetic Sampling (ADASYN). Kontribusi utama dari penelitian ini adalah mengusulkan integrasi KNN dan ADASYN untuk meningkatkan performa klasifikasi risiko diabetes, membandingkan performa model sebelum dan sesudah penerapan teknik oversampling menggunakan evaluasi berbasis confusion matrix, precision, recall, F1-Score, accuracy, dan ROC-AUC, serta memberikan analisis empiris terhadap efektivitas pendekatan hybrid sederhana dalam konteks klasifikasi medis berbasis data yang tidak seimbang, yang diharapkan dapat menjadi rujukan untuk pengembangan model klasifikasi serupa pada domain kesehatan lainnya.

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Pengembangan model klasifikasi risiko Diabetes Mellitus berbasis machine learning dalam penelitian ini dilakukan melalui sejumlah tahapan sistematis yang saling terkait. Rangkaian tahapan ini dirancang secara metodologis untuk memastikan bahwa pendekatan yang diterapkan dapat direproduksi dan diterapkan dalam skenario serupa [13]. Alur tahapan penelitian yang diimplementasikan ditampilkan pada Gambar 1.



Gambar 1. Alur Penelitian

Merujuk pada Gambar 1, berikut adalah penjelasan rinci dari masing-masing tahapan penelitian yang telah diimplementasikan:

Journal of Information System Research (JOSH)
Volume 6, No. 4, Juli 2025, pp 2238−2247
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i4.7237

1) Akuisisi Data
Tahap awal penelitian dimulai dengan pengumpulan dataset yang relevan untuk membangun model klasifikasi Diabetes Mellitus. Dataset yang digunakan diperoleh dari platform Kaggle dengan nama "Diabetes" dan file sumber bernama diabetes.csv [14]. Dataset ini berisi rekam medis pasien yang mencakup sembilan atribut, yaitu jumlah kehamilan (Pregnancies), kadar glukosa (Glucose), tekanan darah (BloodPressure), ketebalan kulit (SkinThickness), kadar insulin (Insulin), indeks massa tubuh (BMI), riwayat genetik diabetes (DiabetesPedigreeFunction), usia (Age), serta label keluaran (Outcome), di mana nilai 1 menunjukkan pasien menderita diabetes dan nilai 0 menunjukkan sebaliknya. Dataset terdiri dari 2000 baris data dan dipilih karena memiliki fitur yang representatif serta kualitas data yang memadai untuk mendukung proses analisis dan klasifikasi dalam konteks penelitian ini.
2) Pra-pemrosesan Data
Pra-pemrosesan dilakukan untuk memastikan kualitas dan konsistensi data sebelum digunakan dalam tahap pelatihan model [15]. Beberapa fitur seperti Glucose, BloodPressure, SkinThickness, Insulin, dan BMI mengandung nilai nol yang secara medis tidak valid, sehingga nilai-nilai tersebut diperlakukan sebagai data hilang dan diimputasi menggunakan nilai median guna menjaga kestabilan distribusi terhadap outlier. Selanjutnya, dilakukan penanganan terhadap nilai-nilai ekstrem pada fitur numerik menggunakan metode winsorizing, dengan membatasi nilai di kedua ujung distribusi pada persentil ke-1 dan ke-99. Langkah ini bertujuan untuk mengurangi pengaruh outlier yang dapat mempengaruhi performa algoritma berbasis jarak. Setelah penanganan nilai ekstrem, seluruh fitur numerik dinormalisasi menggunakan metode StandardScaler sehingga memiliki distribusi dengan rata-rata nol dan standar deviasi satu. Normalisasi ini penting untuk menghilangkan dominasi skala antar fitur dan meningkatkan efektivitas algoritma berbasis jarak seperti K-Nearest Neighbors (KNN). Terakhir, dataset dibagi menjadi dua bagian, yaitu 80% untuk data latih dan 20% untuk data uji, menggunakan metode stratified split. Rasio ini dipilih untuk memastikan keseimbangan antara ukuran data latih dan data uji serta mendukung evaluasi model yang lebih adil [16].
3) Penanganan Imbalanced Data Menggunakan ADASYN
Dataset yang digunakan memiliki ketidakseimbangan kelas, di mana jumlah sampel penderita diabetes (Outcome = 1) lebih sedikit dibandingkan non-diabetes. Ketidakseimbangan ini dapat menyebabkan model bias terhadap kelas mayoritas dan menurunkan kemampuan deteksi terhadap pasien berisiko tinggi. Untuk mengatasi hal ini, digunakan metode Adaptive Synthetic Sampling (ADASYN), yang merupakan teknik oversampling berbasis adaptif. ADASYN bekerja dengan menghasilkan sampel sintetis tambahan di sekitar sampel minoritas yang sulit diklasifikasikan, sehingga meningkatkan fokus model pada area kompleks dari distribusi data minoritas [17]. Dengan demikian, ADASYN tidak hanya menyeimbangkan jumlah data antar kelas, tetapi juga memperbaiki representasi lokal dari kelas minoritas.
4) Pembangunan Model Klasifikasi Menggunakan K-Nearest Neighbors
Model klasifikasi yang digunakan dalam penelitian ini adalah K-Nearest Neighbors (KNN), sebuah algoritma supervised learning berbasis instance yang menentukan kelas sebuah data baru berdasarkan mayoritas kelas dari sejumlah tetangga terdekatnya [18]. KNN dipilih karena kesederhanaannya, efektivitas dalam menangani hubungan non-linear antar fitur, serta tidak memerlukan asumsi distribusi data. Dalam penelitian ini, model KNN dilatih pada data latih baik dalam kondisi asli maupun setelah dilakukan oversampling menggunakan ADASYN. Nilai parameter k (jumlah tetangga) diatur ke 5 berdasarkan praktik umum dalam penerapan KNN, untuk menjaga keseimbangan antara bias dan varians.
5) Evaluasi Kinerja Model
Evaluasi dilakukan untuk mengukur performa model dalam mengklasifikasikan pasien menjadi kelompok diabetes dan non-diabetes. Metrik evaluasi yang digunakan meliputi confusion matrix, precision, recall (sensitivitas), F1-Score, serta ROC Curve dan ROC-AUC Score. Confusion matrix memberikan informasi tentang jumlah klasifikasi benar dan salah untuk masing-masing kelas, sementara precision dan recall digunakan untuk mengukur ketepatan serta kelengkapan deteksi [19]. F1-Score dihitung untuk mengevaluasi keseimbangan antara precision dan recall. Selain itu, ROC Curve digunakan untuk menggambarkan performa model pada berbagai ambang klasifikasi, dan nilai AUC digunakan untuk mengukur kemampuan diskriminatif model [20]. Evaluasi ini dilakukan dengan membandingkan performa KNN sebelum dan sesudah penerapan teknik ADASYN, untuk menilai kontribusi oversampling terhadap peningkatan performa model.

2.2 Adaptive Synthetic Sampling (ADASYN)

Adaptive Synthetic Sampling (ADASYN) adalah salah satu teknik oversampling berbasis adaptif yang dikembangkan untuk mengatasi ketidakseimbangan kelas dalam pembelajaran mesin [21]. Teknik ini merupakan pengembangan dari metode Synthetic Minority Oversampling Technique (SMOTE) yang bertujuan tidak hanya untuk menyeimbangkan jumlah sampel antar kelas, tetapi juga memperhatikan kompleksitas distribusi lokal dari kelas minoritas [22]. Berbeda dengan SMOTE yang secara seragam membuat data sintetis di seluruh data minoritas, ADASYN secara dinamis menghasilkan lebih banyak sampel sintetis di area di mana minoritas lebih sulit diklasifikasikan, berdasarkan jumlah tetangga mayoritas di sekitarnya [12].

Secara umum, prosedur ADASYN (Adaptive Synthetic Sampling) melibatkan beberapa tahap yang saling berhubungan. Tahap pertama adalah menghitung jumlah tetangga mayoritas di sekitar setiap sampel minoritas

Journal of Information System Research (JOSH)
Volume 6, No. 4, Juli 2025, pp 2238−2247
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i4.7237

menggunakan algoritma K-Nearest Neighbors. Setelah itu, tingkat kesulitan klasifikasi untuk masing-masing sampel minoritas ditentukan berdasarkan seberapa sulit sampel tersebut untuk diklasifikasikan. Terakhir, sampel sintetis dihasilkan secara proporsional terhadap tingkat kesulitan klasifikasi yang telah ditentukan sebelumnya, dengan tujuan untuk meningkatkan representasi sampel minoritas dalam dataset. Jumlah sampel sintetis yang harus dihasilkan untuk setiap data minoritas dihitung berdasarkan tingkat kesulitan klasifikasinya dengan persamaan (1).

$$G_i = \frac{r_i}{\sum_{i=1}^{n_{minority}} r_i} \times G \quad (1)$$

di mana G adalah total jumlah sampel sintetis yang akan dibuat, r_i adalah rasio tetangga mayoritas di sekitar sampel minoritas i, G_i adalah jumlah sampel sintetis untuk data i. Sampel sintetis dibuat menggunakan interpolasi linier antara data minoritas dan tetangga terdekatnya, dirumuskan melalui persamaan (2).

$$x_{new} = x_i + \delta \times (x_{zi} - x_i) \quad (2)$$

di mana δ adalah bilangan acak dalam interval [0,1], x_i adalah sampel minoritas, dan x_{zi} adalah salah satu tetangganya. Dengan pendekatan ini, ADASYN mampu memperkaya representasi data minoritas terutama pada wilayah yang sulit diklasifikasikan, sekaligus mengurangi risiko overfitting yang sering terjadi pada metode oversampling konvensional.

2.3 Metode K-Nearest Neighbors

K-Nearest Neighbors (KNN) merupakan algoritma klasifikasi non-parametrik berbasis instance yang sangat populer dalam berbagai tugas pembelajaran mesin, termasuk klasifikasi medis [23]. Algoritma ini mengasumsikan bahwa objek-objek yang berdekatan dalam ruang fitur cenderung memiliki label kelas yang sama. Oleh karena itu, klasifikasi terhadap suatu data baru dilakukan dengan melihat mayoritas kelas dari k tetangga terdekatnya di ruang fitur [24].

Dalam implementasinya, langkah-langkah dasar K-Nearest Neighbors (KNN) meliputi beberapa tahapan yang saling berurutan. Pertama, jarak antara data baru dengan semua data dalam dataset latih dihitung. Selanjutnya, berdasarkan jarak terpendek, ditentukan k tetangga terdekat yang paling relevan dengan data baru tersebut. Terakhir, dilakukan proses voting mayoritas terhadap label dari tetangga-tetangga terdekat untuk menentukan label prediksi dari data baru, yang merupakan hasil klasifikasi berdasarkan mayoritas label yang ada di antara tetangga-tetangga tersebut. Jarak antar data umumnya dihitung menggunakan jarak Euclidean, yang dirumuskan melalui persamaan (3).

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2} \quad (3)$$

di mana $d(x_i, x_j)$ adalah jarak Euclidean antara dua titik data x_i dan x_j, x_{ik} dan x_{jk} adalah nilai fitur ke-k dari masing-masing data, n adalah jumlah fitur. Setelah diperoleh k tetangga terdekat, keputusan kelas (C) dari data baru x diambil berdasarkan voting mayoritas menggunakan persamaan (4).

$$C(x) = arg\,\max_c \sum_{i \in N_k(x)} 1(y_i = c) \quad (4)$$

di mana $N_k(x)$ adalah himpunan tetangga terdekat ke-k dari x, $1(y_i = c)$ adalah fungsi indikator yang bernilai 1 jika label y_i sama dengan kelas c, dan 0 sebaliknya. KNN memiliki keunggulan dalam fleksibilitas terhadap bentuk distribusi data, kemudahan implementasi, dan interpretabilitas [9]. Namun, kelemahannya adalah kebutuhan memori yang tinggi untuk dataset besar dan sensitivitas terhadap ketidakseimbangan kelas, sehingga penerapan metode oversampling seperti ADASYN sangat penting untuk meningkatkan kinerjanya. Selain itu, performa KNN sangat dipengaruhi oleh pemilihan nilai k dan teknik pra-pemrosesan data, yang berperan krusial dalam mengoptimalkan akurasi klasifikasi [25].

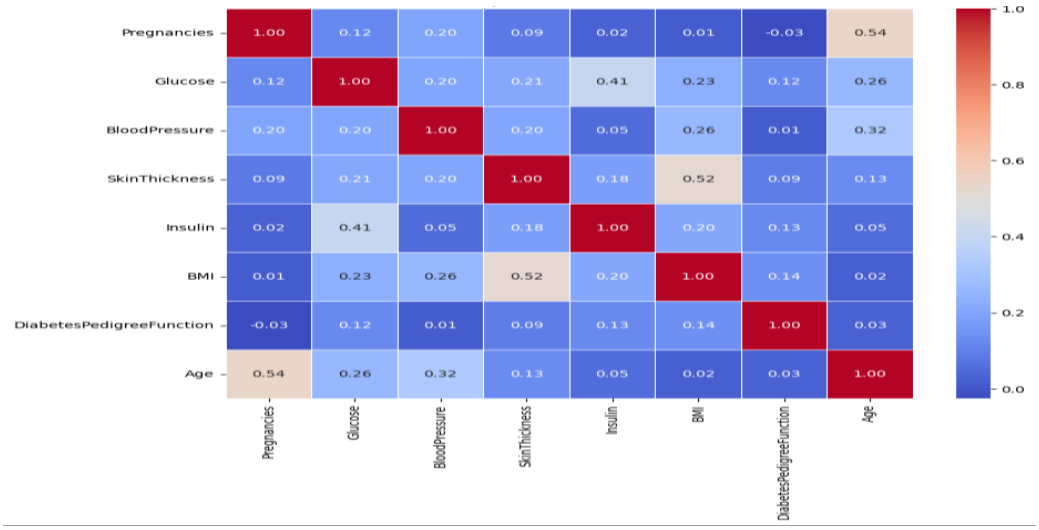
3. HASIL DAN PEMBAHASAN

Pembangunan model klasifikasi risiko Diabetes Mellitus menggunakan pendekatan K-Nearest Neighbors (KNN) diawali dengan tahapan persiapan dataset yang relevan. Dataset yang digunakan dalam penelitian ini diperoleh dari platform publik Kaggle dengan judul "Diabetes" [14]. Dataset ini memuat informasi rekam medis dari 2000 pasien, mencakup sembilan atribut prediktor, yaitu jumlah kehamilan (Pregnancies), kadar glukosa (Glucose), tekanan darah diastolik (BloodPressure), ketebalan kulit (SkinThickness), kadar insulin (Insulin), indeks massa tubuh (BMI), riwayat genetik diabetes (DiabetesPedigreeFunction), usia (Age), serta label keluaran (Outcome) yang menunjukkan status diabetes pasien. Label Outcome bernilai 1 untuk pasien yang menderita Diabetes dan 0 untuk pasien Non-Diabetes. Sebelum digunakan dalam proses pembangunan model, data melalui tahap pembersihan awal yang mencakup imputasi nilai nol pada fitur-fitur medis penting dengan nilai median, serta normalisasi seluruh fitur numerik menggunakan metode StandardScaler. Proses ini bertujuan untuk memastikan bahwa semua fitur memiliki skala yang seragam dan untuk mengurangi bias dalam

Journal of Information System Research (JOSH)
Volume 6, No. 4, Juli 2025, pp 2238−2247
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i4.7237

perhitungan jarak antar data yang menjadi dasar dalam algoritma KNN. Normalisasi sangat diperlukan mengingat perbedaan skala antar fitur dapat mempengaruhi performa model berbasis jarak.

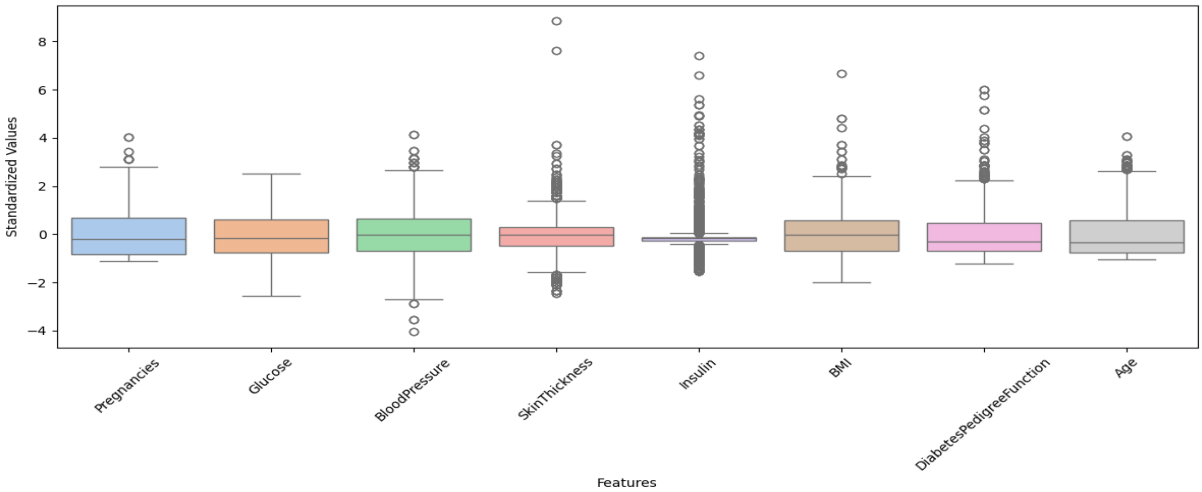
Setelah tahap pembersihan data selesai, penelitian dilanjutkan dengan eksplorasi data untuk memahami karakteristik dasar dataset yang digunakan. Eksplorasi data bertujuan untuk mengidentifikasi distribusi masing-masing fitur, mendeteksi adanya ketidakseimbangan kelas, serta mengevaluasi potensi korelasi antar fitur. Langkah awal dalam eksplorasi ini adalah menganalisis distribusi masing-masing atribut, baik fitur input maupun target, untuk memahami pola penyebaran data dan mendeteksi adanya outlier yang dapat memengaruhi proses pelatihan model. Pemahaman terhadap distribusi ini penting dalam menentukan strategi preprocessing dan pemilihan metode penguatan data. Selanjutnya, hubungan antar fitur dianalisis menggunakan heatmap korelasi, yang disajikan pada Gambar 2.



Gambar 2. Heatmap Korelasi Antar Fitur

Gambar 2 menunjukkan heatmap korelasi antar fitur numerik dalam dataset yang digunakan. Secara umum, hubungan antar fitur relatif lemah, dengan sebagian besar nilai korelasi berada di bawah 0,5. Korelasi positif yang paling kuat terlihat antara fitur SkinThickness dan BMI (0,52), serta antara Pregnancies dan Age (0,54). Selain itu, terdapat korelasi moderat antara Insulin dan Glucose (0,41), serta BloodPressure dan Age (0,32). Nilai korelasi yang rendah menunjukkan bahwa sebagian besar fitur dalam dataset memiliki hubungan yang cukup independen satu sama lain, sehingga penggunaan kombinasi fitur menjadi penting untuk membangun model klasifikasi yang efektif. Heatmap ini juga menunjukkan bahwa tidak ada fitur tunggal yang secara dominan berkorelasi dengan fitur lain, sehingga pendekatan berbasis keseluruhan fitur lebih tepat diterapkan dalam penelitian ini.

Selanjutnya, dilakukan analisis terhadap karakteristik fitur numerik dalam dataset. Tujuan dari langkah ini adalah untuk memahami distribusi nilai masing-masing fitur, mengidentifikasi pola distribusi, serta mendeteksi kemungkinan adanya outlier yang dapat memengaruhi proses pelatihan model. Informasi ini sangat penting sebagai dasar dalam memilih teknik pra-pemrosesan yang tepat, seperti pemilihan metode skala, serta memberikan wawasan awal tentang kontribusi masing-masing fitur terhadap proses klasifikasi. Visualisasi boxplot dari fitur numerik dalam dataset yang digunakan dapat dilihat pada Gambar 3.

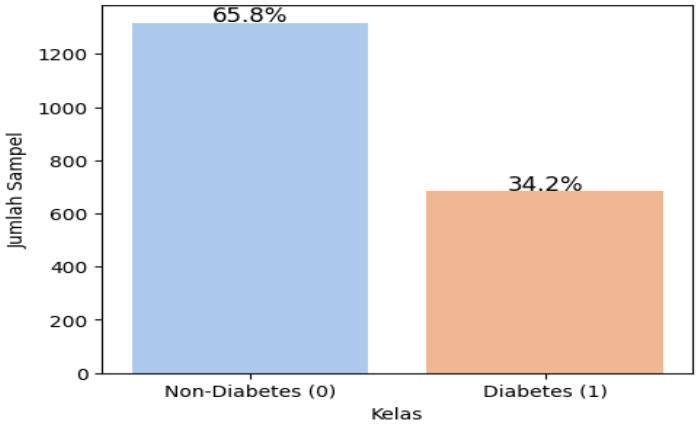


Gambar 3. Boxplot Fitur-Fitur Numerik

Journal of Information System Research (JOSH)
Volume 6, No. 4, Juli 2025, pp 2238−2247
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i4.7237

Gambar 3 menunjukkan boxplot gabungan seluruh fitur numerik setelah normalisasi. Terlihat bahwa fitur Insulin memiliki penyebaran nilai dan jumlah outlier yang paling ekstrem, sementara fitur SkinThickness, BloodPressure, dan BMI menunjukkan outlier dalam jumlah moderat. Mayoritas fitur memiliki distribusi yang relatif simetris, kecuali Insulin yang lebih menyebar. Mengingat sensitivitas algoritma K-Nearest Neighbors (KNN) terhadap outlier, dilakukan winsorizing pada fitur Insulin dengan membatasi nilai pada persentil ke-1 dan ke-99. Langkah ini bertujuan mengurangi pengaruh outlier ekstrem tanpa mengubah karakter distribusi utama, sebelum seluruh fitur dinormalisasi menggunakan StandardScaler untuk memastikan keseragaman skala data.

Langkah berikutnya dalam eksplorasi ini adalah memvisualisasikan distribusi kelas target. Tujuan dari visualisasi ini adalah untuk mendapatkan pemahaman mengenai keseimbangan antar kelas, yang merupakan faktor krusial dalam klasifikasi biner. Distribusi data target dapat dilihat pada Gambar 4.



Gambar 4. Distribusi Kelas Target

Gambar 4 menunjukkan bahwa terdapat ketidakseimbangan distribusi antara kelas Non-Diabetes (0) dan Diabetes (1), di mana 65,8% sampel merupakan pasien Non-Diabetes, sedangkan 34,2% sisanya adalah pasien Diabetes. Ketidakseimbangan ini berpotensi menyebabkan model bias terhadap kelas mayoritas, sehingga diperlukan strategi khusus, seperti teknik oversampling menggunakan ADASYN, untuk memperbaiki distribusi data sebelum membangun model klasifikasi.

Sebagai ilustrasi untuk memperjelas prinsip kerja ADASYN, disajikan sebuah studi kasus perhitungan manual sederhana yang menunjukkan proses penentuan jumlah sampel sintetis, alokasi berdasarkan tingkat kesulitan klasifikasi, serta mekanisme pembuatan sampel baru melalui interpolasi antara sampel minoritas dan tetangganya. Contoh ini menggunakan dua fitur dari dataset Diabetes, yaitu Glucose dan BMI, sebagaimana ditampilkan dalam Tabel 1.

Tabel 1. Sampel Data

Sampel	Glucose	BMI	Outcome
A	150	35	1 (Diabetes)
B	140	33	1 (Diabetes)
C	160	38	1 (Diabetes)
D	100	28	0 (Non-Diabetes)
E	95	27	0 (Non-Diabetes)
F	105	29	0 (Non-Diabetes)

Langkah pertama dalam ADASYN adalah mencari tetangga terdekat untuk setiap sampel minoritas (A, B, dan C) berdasarkan jarak Euclidean antara nilai Glucose dan BMI. Berikut perhitungan jarak antara sampel A dan D:

$$d(A,D) = \sqrt{(150-100)^2 + (35-28)^2} = \sqrt{2500+49} = \sqrt{2549} = 50{,}49$$

Setelah menghitung jarak Euclidean antar semua kombinasi data, diperoleh bahwa tetangga terdekat untuk sampel A adalah D (Non-Diabetes), B (Diabetes), dan F (Non-Diabetes). Untuk sampel B, tetangga terdekatnya adalah D (Non-Diabetes), A (Diabetes), dan E (Non-Diabetes). Sedangkan untuk sampel C, tetangganya adalah F (Non-Diabetes), B (Diabetes), dan D (Non-Diabetes). Selanjutnya, langkah berikutnya adalah menghitung rasio r_i, yang merupakan proporsi jumlah tetangga dari kelas mayoritas dibandingkan dengan jumlah total tetangga. Berdasarkan susunan tetangga tersebut, nilai rasio untuk masing-masing sampel adalah sebagai berikut:

$$r_A = \frac{2}{3} = 0{,}666$$

$$r_B = \frac{2}{3} = 0{,}666$$

$$r_C = \frac{2}{3} = 0{,}666$$

Journal of Information System Research (JOSH)
Volume 6, No. 4, Juli 2025, pp 2238−2247
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i4.7237

Sehingga jumlah total rasio r adalah:

$$\sum r = 0{,}666 + 0{,}666 + 0{,}666 = 2{,}0$$

Langkah ketiga adalah menentukan jumlah total sampel sintetis yang perlu dihasilkan (G). Dalam contoh ini, untuk meningkatkan jumlah sampel minoritas dari 3 menjadi 4, diperlukan 1 sampel sintetis:

$$G = 4 - 3 = 1$$

Langkah keempat adalah menghitung berapa banyak sampel sintetis yang dialokasikan untuk masing-masing sampel minoritas berdasarkan rasio r_i. Menggunakan persamaan (1), berikut perhitungannya:

$$G_A = \frac{0{,}666}{2{,}0} = 0{,}333$$

$$G_B = \frac{0{,}666}{2{,}0} = 0{,}333$$

$$G_C = \frac{0{,}666}{2{,}0} = 0{,}333$$

Langkah kelima dalam penerapan ADASYN adalah pembuatan data sintetis berdasarkan alokasi jumlah sampel yang telah dihitung. Pada kasus ini, karena nilai G_A, G_B, dan G_C masing-masing sebesar 0,333, dan jumlah total sampel sintetis yang dibutuhkan hanya satu, maka satu sampel minoritas dipilih secara acak untuk dibuatkan data sintetis. Misalnya, sampel A terpilih untuk proses pembuatan sampel baru. Selanjutnya, dilakukan interpolasi linier antara sampel A dan salah satu tetangga minoritas terdekatnya, dalam hal ini sampel B. Nilai interpolasi dihitung dengan memilih faktor acak δ sebesar 0,5 untuk menghasilkan nilai fitur baru. Perhitungan interpolasi untuk fitur Glucose dan BMI dilakukan menggunakan persamaan (2). Berikut hasil perhitungannya:

$$Glucose_{New} = 150 + 0{,}5 \times (140 - 150) = 145$$

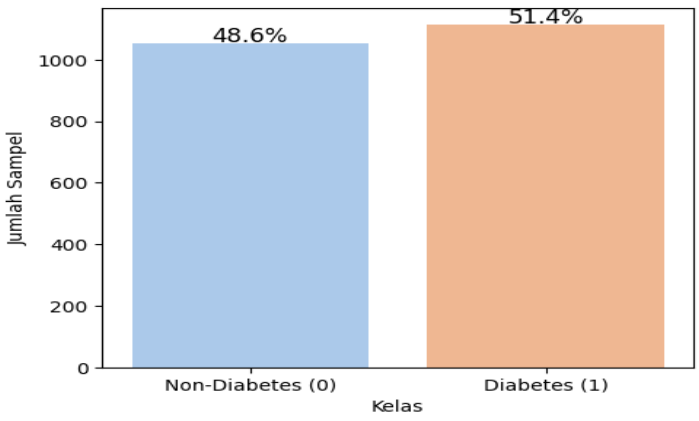
$$BMI_{New} = 35 + 0{,}5 \times (33 - 35) = 34$$

Hasil dari interpolasi tersebut menghasilkan satu data sintetis baru. Hasil keseluruhan data setelah dilakukan teknik oversampling dengan ADASYN ditampilkan pada Tabel 2.

Tabel 2. Hasil Oversampling Menggunakan ADASYN

Sampel	Glucose	BMI	Outcome
A	150	35	1 (Diabetes)
B	140	33	1 (Diabetes)
C	160	38	1 (Diabetes)
D	100	28	0 (Non-Diabetes)
E	95	27	0 (Non-Diabetes)
F	105	29	0 (Non-Diabetes)
Synthetic 1	145	34	1 (Diabetes)

Tabel 2 menyajikan hasil perhitungan interpolasi linier dalam proses pembentukan data sintetis menggunakan teknik ADASYN. Contoh ini menggambarkan bagaimana ADASYN bekerja secara adaptif dengan menghasilkan sampel baru di sekitar area minoritas yang lebih sulit diklasifikasikan, berdasarkan tingkat kepadatan tetangga mayoritas. Berbeda dengan metode oversampling standar, ADASYN berfokus memperkuat representasi lokal data minoritas sehingga dapat meminimalkan bias klasifikasi tanpa menciptakan data sintetis secara acak di seluruh ruang fitur. Hasil distribusi kelas target setelah penerapan oversampling menggunakan ADASYN pada dataset ini ditampilkan pada Gambar 5.

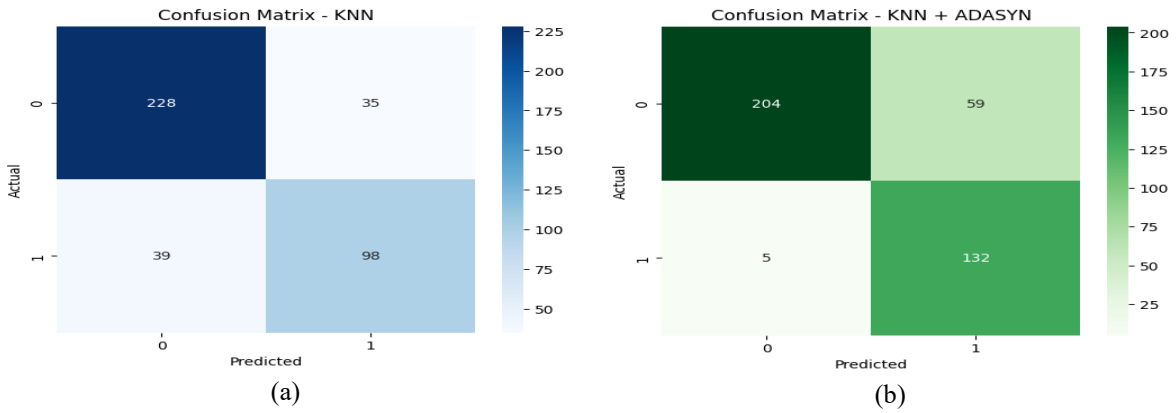


Gambar 5. Distribusi Kelas Target Setelah Dilakukan Oversampling Menggunakan ADASYN

Gambar 5 menunjukkan distribusi kelas target setelah dilakukan oversampling menggunakan ADASYN. Setelah proses ini, distribusi kelas menjadi lebih seimbang, dengan 48,6% sampel pada kelas Non-Diabetes (0) dan 51,4% pada kelas Diabetes (1). Keseimbangan ini berperan penting dalam mengurangi potensi bias model terhadap kelas mayoritas, serta meningkatkan sensitivitas model dalam mendeteksi kasus diabetes.

Journal of Information System Research (JOSH)
Volume 6, No. 4, Juli 2025, pp 2238−2247
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i4.7237

Proses selanjutnya adalah membangun model klasifikasi menggunakan algoritma K-Nearest Neighbors (KNN), yang diterapkan untuk proses pelatihan dan pengujian. Dataset dibagi menjadi data pelatihan dan data pengujian dengan rasio 80:20 menggunakan teknik stratified sampling, untuk memastikan bahwa proporsi kelas target tetap seimbang pada kedua subset data. KNN bekerja dengan mengklasifikasikan sampel baru berdasarkan mayoritas kelas dari k tetangga terdekatnya dalam ruang fitur, menggunakan perhitungan jarak, seperti jarak Euclidean. Sifat non-parametrik ini menjadikan KNN sangat sesuai untuk kasus klasifikasi biner berbasis instance, seperti dalam studi klasifikasi risiko Diabetes Mellitus. Proses pelatihan pada KNN melibatkan penyimpanan seluruh data latih, di mana model kemudian menentukan kelas untuk data uji dengan mengidentifikasi k tetangga terdekat dan melakukan voting mayoritas. Untuk menguji performa model pertama, digunakan confusion matrix, yang bertujuan untuk mengevaluasi hasil klasifikasi model. Evaluasi ini dilakukan pada model KNN tanpa oversampling dan model KNN dengan ADASYN. Hasil dari confusion matrix untuk kedua model dapat dilihat pada Gambar 6.

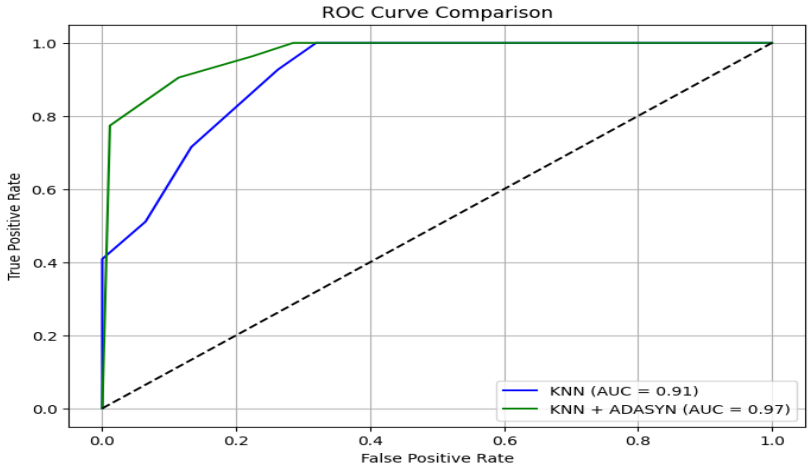


(a) (b)

Gambar 6. (a) Confusion Matrix Menggunakan KNN Tanpa Oversampling dan (b) Confusion Matrik Menggunakan KNN Dengan ADASYN

Gambar 6(a) menunjukkan confusion matrix untuk model KNN tanpa oversampling, di mana terdapat 228 prediksi benar untuk kelas Non-Diabetes dan 98 prediksi benar untuk kelas Diabetes. Namun, model masih menunjukkan tingkat kesalahan yang cukup besar dalam mengklasifikasikan pasien diabetes, dengan 39 kasus positif yang salah diklasifikasikan sebagai negatif. Sebaliknya, Gambar 6(b) memperlihatkan confusion matrix setelah penerapan teknik oversampling ADASYN. Terjadi peningkatan yang signifikan dalam pendeteksian kasus Diabetes, dengan 132 prediksi benar dan hanya 5 kesalahan klasifikasi pada kelas positif. Meskipun terjadi sedikit penurunan pada akurasi kelas Non-Diabetes, hasil ini menunjukkan bahwa ADASYN berhasil meningkatkan sensitivitas model terhadap kelas minoritas, sehingga lebih efektif dalam mendeteksi pasien yang berisiko diabetes.

Uji performa selanjutnya dilakukan menggunakan ROC Curve, yang bertujuan untuk mengevaluasi kemampuan model dalam membedakan antara kelas positif dan negatif pada berbagai ambang batas keputusan. Hasil perbandingan ROC Curve untuk model KNN tanpa oversampling dan KNN dengan ADASYN ditampilkan pada Gambar 7.



Gambar 7. Perbandingan ROC Curve Menggunakan KNN Tanpa Oversampling dan KNN Dengan ADASYN

Gambar 6 menunjukkan perbandingan kurva ROC antara model KNN tanpa oversampling dan model KNN dengan ADASYN. Model KNN tanpa oversampling menghasilkan nilai AUC sebesar 0,91, sedangkan model KNN dengan ADASYN mencapai nilai AUC sebesar 0,97. Peningkatan nilai AUC ini menunjukkan bahwa

Journal of Information System Research (JOSH)
Volume 6, No. 4, Juli 2025, pp 2238−2247
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i4.7237

penerapan ADASYN berhasil meningkatkan kemampuan model dalam membedakan antara kelas positif dan negatif secara lebih akurat.

Berdasarkan confusion matrix dan ROC Curve, diperoleh metrik evaluasi tambahan berupa precision, recall, F1-Score, akurasi, dan ROC-AUC. Hasil evaluasi tersebut untuk model klasifikasi risiko Diabetes Mellitus menggunakan KNN tanpa oversampling dan KNN dengan ADASYN ditampilkan pada Tabel 3.

Tabel 3. Hasil Evaluasi Model Klasifikasi Risiko Diabetes Mellitus Menggunakan KNN Tanpa Oversampling dan KNN Dengan ADASYN

Model	Kelas	Precision	Recall	F1-Score	Accuracy	ROC-AUC Score
KNN	Non-Diabetes	73,68%	71,53%	72,59%	81,50%	91,34%
	Diabetes	85,39%	86,69%	86,04%		
KNN + ADASYN	Non-Diabetes	69,11%	96,35%	80,49%	84,00%	96,82%
	Diabetes	97,61%	77,57%	86,44%		

Tabel 3 menunjukkan hasil evaluasi model klasifikasi risiko Diabetes Mellitus menggunakan algoritma K-Nearest Neighbors (KNN) tanpa oversampling dan dengan penerapan teknik Adaptive Synthetic Sampling (ADASYN). Secara umum, penerapan ADASYN menunjukkan peningkatan performa yang signifikan pada sejumlah metrik evaluasi.

Peningkatan paling mencolok terlihat pada nilai ROC-AUC, di mana model KNN tanpa oversampling memiliki skor 91,34%, sedangkan model KNN dengan ADASYN meningkat menjadi 96,82%, atau mengalami kenaikan sebesar 5,48%. Selain itu, akurasi keseluruhan model juga meningkat dari 81,50% menjadi 84,00%, mencatatkan peningkatan sebesar 2,50%. Perbaikan juga terjadi pada F1-Score kelas Diabetes, dari 86,04% menjadi 86,44%, meskipun peningkatannya relatif kecil sebesar 0,40%, menunjukkan bahwa model menjadi lebih seimbang dalam mengklasifikasikan pasien diabetes setelah oversampling.

Keberhasilan ini terutama disebabkan oleh kemampuan ADASYN dalam mengatasi ketidakseimbangan kelas. ADASYN secara adaptif menghasilkan sampel sintetis tambahan di sekitar sampel minoritas yang sulit diklasifikasikan, sehingga model lebih mampu mengenali pola-pola penting dari kelas Diabetes tanpa mengorbankan performa klasifikasi untuk kelas Non-Diabetes secara berlebihan.

Namun demikian, masih terdapat aspek yang perlu ditingkatkan. Meskipun recall untuk kelas Non-Diabetes meningkat secara signifikan dari 71,53% menjadi 96,35%, precision untuk kelas ini justru mengalami penurunan dari 73,68% menjadi 69,11%. Hal ini menunjukkan bahwa model menjadi lebih sensitif dalam mendeteksi Non-Diabetes, namun dengan kecenderungan menghasilkan lebih banyak false positive. Oleh karena itu, pengembangan selanjutnya perlu mempertimbangkan teknik balancing yang lebih cermat atau mengoptimalkan threshold klasifikasi agar trade-off antara precision dan recall lebih ideal, khususnya pada kelas mayoritas.

4. KESIMPULAN

Penelitian ini berhasil membangun model klasifikasi risiko Diabetes Mellitus menggunakan algoritma K-Nearest Neighbors (KNN) dengan pendekatan penanganan ketidakseimbangan data melalui teknik Adaptive Synthetic Sampling (ADASYN). Berdasarkan hasil evaluasi, penerapan ADASYN terbukti meningkatkan performa model secara signifikan dibandingkan KNN tanpa oversampling. Model KNN dengan ADASYN mampu meningkatkan nilai ROC-AUC sebesar 5,48% (dari 91,34% menjadi 96,82%) dan akurasi keseluruhan sebesar 2,50% (dari 81,50% menjadi 84,00%). Selain itu, F1-Score untuk kelas Diabetes juga mengalami peningkatan sebesar 0,40%, menunjukkan bahwa model menjadi lebih efektif dalam mendeteksi pasien berisiko diabetes. Keberhasilan ini terutama didorong oleh kemampuan ADASYN dalam memperkuat representasi kelas minoritas secara adaptif, sehingga model menjadi lebih sensitif terhadap pola klasifikasi pasien diabetes. Secara keseluruhan, integrasi KNN dan ADASYN dalam penelitian ini menunjukkan potensi yang kuat untuk meningkatkan akurasi dan keadilan klasifikasi pada data medis yang tidak seimbang, serta memberikan kontribusi nyata dalam pengembangan sistem klasifikasi risiko penyakit berbasis machine learning. Meskipun demikian, precision untuk kelas Non-Diabetes mengalami sedikit penurunan, yang mengindikasikan adanya peningkatan jumlah false positive. Oleh karena itu, penelitian selanjutnya disarankan untuk mengeksplorasi kombinasi teknik oversampling lain atau mengoptimalkan threshold klasifikasi guna mencapai keseimbangan yang lebih ideal antara precision dan recall.

REFERENCES

[1] D. Priyantini, N. A. Sari, and P. A. Hanggitriana, “Indeks Massa Tubuh Pada Penderita Diabetes Melitus Dengan Nilai Ankle Brachial Index,” J. Ilm. Keperawatan Stikes Hang Tuah Surabaya, vol. 17, no. 2, pp. 144–149, 2022.

[2] A. Aminuddin, Yenny Sima, Nurril Cholifatul Izza, Nur Syamsi Norma Lalla, and Darmi Arda, “Edukasi Kesehatan Tentang Penyakit Diabetes Melitus bagi Masyarakat,” Abdimas Polsaka, pp. 7–12, 2023, doi: 10.35816/abdimaspolsaka.v2i1.25.

[3] M. A. Sembiring, H. Saputra, R. A. Yusda, S. Sutarman, and E. B. Nababan, “Performance of Robust Support Vector Machine Classification Model on Balanced, Imbalanced and Outliers Datasets,” JITK (Jurnal Ilmu Pengetah. dan Teknol.

Journal of Information System Research (JOSH)
Volume 6, No. 4, Juli 2025, pp 2238−2247
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i4.7237

Komputer), vol. 10, no. 1, pp. 208–215, 2024, doi: 10.33480/jitk.v10i1.5272.
[4] 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," Indones. J. Softw. Eng., vol. 7, no. 1, pp. 63–68, 2021, doi: 10.31294/ijse.v7i1.10279.
[5] G. Abdurrahman, "Klasifikasi Penyakit Diabetes Melitus Menggunakan Adaboost Classifier," JUSTINDO (Jurnal Sist. Teknol. Inf. Indones., vol. 7, no. 1, pp. 59–66, 2022.
[6] L. Safitri and Z. Fatah, "Implementasi Prediksi Penyakit Diabetes Menggunakan Metode Decision Tree," JUSIFOR J. Sist. Inf. dan Inform., vol. 2, no. 2, pp. 125–132, 2023, doi: 10.70609/jusifor.v3i2.5788 .
[7] S. P. Nainggolan and A. Sinaga, "Comparative Analysis of Accuracy of Random Forest and Gradient Boosting Classifier Algorithm for Diabetes Classification," Sebatik, vol. 27, no. 1, pp. 97–102, 2023, doi: 10.46984/sebatik.v27i1.2157.
[8] P. R. Putri and R. Alit, "Klasifikasi Penyakit Diabetes menggunakan Metode Support Vector Machine," J. Pengemb. Teknol. Inf. dan Ilmu Komput., vol. 6, no. 3, pp. 740–746, 2024.
[9] A. Yogianto, A. Homaidi, and Z. Fatah, "Implementasi Metode K-Nearest Neighbors (KNN) untuk Klasifikasi Penyakit Jantung," G-Tech J. Teknol. Terap., vol. 8, no. 3, pp. 1720–1728, 2024, doi: 10.33379/gtech.v8i3.4495.
[10] S. G. Barus, "Klasifikasi Sentimen Data Tidak Seimbang Menggunakan Algoritma SMOTE dan K-Nearest Neighbor Pada Ulasan Pengguna Aplikasi Pedulilindungi," in Seminar Nasional Mahasiswa Ilmu Komputer dan Aplikasinya (SENAMIKA), 2022, pp. 162–173.
[11] M. T. T. B. Sirait, N. S. Fathonah, and M. N. Fauzan, "Pemanfaatan Algoritma ADASYN dan Support Vector Machine Dalam Meningkatkan Akurasi Prediksi Kanker Paru-Paru," JATI (Jurnal Mhs. Tek. Inform., vol. 8, no. 5, pp. 8773–8778, 2024.
[12] M. I. Anugrah, J. Zeniarja, and D. S. Setiawan, "Peningkatan Performa Model Hard Voting Classifier dengan Teknik Oversampling ADASYN pada Penyakit Diabetes," Edumatic J. Pendidik. Inform., vol. 8, no. 1, pp. 290–299, 2024, doi: 10.29408/edumatic.v8i1.25838.
[13] R. I. Borman, R. Napianto, N. Nugroho, D. Pasha, Y. Rahmanto, and Y. E. P. Yudoutomo, "Implementation of PCA and KNN Algorithms in the Classification of Indonesian Medicinal Plants," in International Conference on Computer Science, Information Technology and Electrical Engineering (ICOMITEE), IEEE, 2021, pp. 46–50.
[14] J. Dasilva, "Diabetes Dataset," Kaggle. [Online]. Available: https://www.kaggle.com/datasets/johndasilva/diabetes
[15] I. Ahmad, A. Tri Prastowo, R. Indra Borman, M. Tonggiroh, and Y. Jusman, "Lung Cancer Classification from CT Scan Images Using the LVQ Algorithm and GLCM Feature Extraction with Spatial Filters," in International Conference on Information Technology and Computing (ICITCOM), IEEE, 2024, pp. 99–104.
[16] I. O. Muraina, "Ideal Dataset Splitting Ratios in Machine Learning Algorithms: General Concerns for Data Scientists and Data Analysts," in International Mardin Artuklu Scientific Researches Conference, 2022, pp. 496–505.
[17] L. Muflikhah, F. A. Bachtiar, D. E. Ratnawati, and R. Darmawan, "Improving Performance for Diabetic Nephropathy Detection Using Adaptive Synthetic Sampling Data in Ensemble Method of Machine Learning Algorithms," J. Ilm. Tek. Elektro Komput. dan Inform., vol. 10, no. 1, p. 123, 2024, doi: 10.26555/jiteki.v10i1.28107.
[18] I. P. Putri, "Analisis Performa Metode K- Nearest Neighbor (KNN) dan Crossvalidation pada Data Penyakit Cardiovascular," Indones. J. Data Sci., vol. 2, no. 1, pp. 21–28, 2021, doi: 10.33096/ijodas.v2i1.25.
[19] Z. Abidin, R. I. Borman, F. B. Ananda, P. Prasetyawan, F. Rossi, and Y. Jusman, "Classification of Indonesian Traditional Snacks Based on Image Using Convolutional Neural Network (CNN) Algorithm," in International Conference on Electronic and Electrical Engineering and Intelligent System (ICE3IS), IEEE, 2022, pp. 18–23.
[20] Y. Liu, Y. Li, and D. Xie, "Implications of imbalanced datasets for empirical ROC-AUC estimation in binary classification tasks," J. Stat. Comput. Simul., vol. 94, no. 1, pp. 183–203, Jan. 2024, doi: 10.1080/00949655.2023.2238235.
[21] I. Pratama, A. Y. Chandra, and P. T. Presetyaningrum, "Seleksi Fitur dan Penanganan Imbalanced Data menggunakan RFECV dan ADASYN," J. Eksplora Inform., vol. 11, no. 1, pp. 38–49, 2022, doi: 10.30864/eksplora.v11i1.578.
[22] R. A. Nurdian, Mujib Ridwan, and Ahmad Yusuf, "Komparasi Metode SMOTE dan ADASYN dalam Meningkatkan Performa Klasifikasi Herregistrasi Mahasiswa Baru," J. Tek. Inform. dan Sist. Inf., vol. 8, no. 1, pp. 24–32, 2022, doi: 10.28932/jutisi.v8i1.4004.
[23] S. Sumarlinda and W. Lestari, "Aplikasi K-Nearest Neighbor (KNN) untuk Klasifikasi Penyakit Kardiovaskuler," in Sumarlinda, Sri Lestari, Wiji, 2022, pp. 259–262.
[24] M. N. Maskuri, K. Sukerti, and R. M. Herdian Bhakti, "Penerapan Algoritma K-Nearest Neighbor (KNN) untuk Memprediksi Penyakit Stroke Stroke Desease Predict Using KNN Algorithm," J. Ilm. Intech Inf. Technol. J. UMUS, vol. 4, no. 1, pp. 130–140, 2022.
[25] A. Razaki, Y. H. Chrisnanto, and M. Melina, "Penanganan Outlier Pada Metode Algoritma K- Nearest Neighbors (KNN) Dengan Metode Kernel Density Estimation Pada Kasus Penyakit Diabetes," INTECOMS J. Inf. Technol. Comput. Sci., vol. 7, no. 4, pp. 1177–1188, 2024, doi: 10.31539/intecoms.v7i4.10866.