VOL. 2 NO. 1 JANUARY 2025
P-ISSN: XXXX-XXXX | E-ISSN: XXXX-XXXX
DOI: XX.XXXXX/jcsai.vXXX.XXXX

JOURNAL OF COMPUTER SCIENCE AND ARTIFICIAL INTELLIGENCE (JCSAI)

MODEL DETEKSI DINI DIABETES BERBASIS K-NEAREST NEIGHBOR DENGAN OPTIMASI PARAMETER DAN FEATURE SELECTION

Waskito Budi Utomo[1], Hendri Prabowo[2], Dwi Septiani Juandaputri[3].

Program Studi Manajemen Informatika[123]
POLITEKNIK GEOLOGI DAN PERTAMBANGAN AGP
ito.waskitobudiutomo@gmail.com[1], prabowo.sun3@gmail.com[2], dwi21septia@gmail.com[3].

(*) Corresponding Author : ito.waskitobudiutomo@gmail.com[1]

Abstract*—Diabetes is a chronic disease that can lead to various serious complications if not detected and managed early. Early detection of diabetes is crucial in preventing and managing this condition effectively. This study aims to enhance the performance of an early diabetes detection model by implementing the K-Nearest Neighbors (KNN) algorithm. This algorithm was chosen for its ability to classify data based on the proximity of similar characteristics. The dataset used in this research was obtained from an open-source platform and consists of several medical features relevant to diabetes diagnosis. The research process follows the stages of the Knowledge Discovery in Databases (KDD) methodology, which includes data selection, preprocessing, transformation, data mining, and evaluation. During the preprocessing stage, data cleaning, normalization, and handling of missing values were conducted to improve data quality before classification. The KNN model was evaluated using accuracy, precision, and recall metrics. The results indicate that the developed model achieved an accuracy of 51.44%, with precision values of 51.98% for the positive class and 50.89% for the negative class, and recall values of 51.81% and 51.07%, respectively. The obtained accuracy suggests that the model still has limitations in distinguishing between diabetic and non-diabetic patients accurately. The high number of false positives and false negatives highlights the need for improvement in parameter selection and data preprocessing optimization. Therefore, further enhancements such as optimizing the K parameter, improving feature selection, and exploring other machine learning methods are required to improve model performance. With further optimization, this early diabetes detection model is expected to become a more effective tool in supporting medical decision-making and early disease management.*

Keywords*: Early Diabetes Detection, K-Nearest Neighbors, Classification, Preprocessing, Model Evaluation.*

Abstrak—Diabetes merupakan penyakit kronis yang dapat menyebabkan berbagai komplikasi serius jika tidak terdeteksi dan ditangani sejak dini. Deteksi dini diabetes menjadi hal yang krusial dalam upaya pencegahan dan pengelolaan penyakit ini. Penelitian ini bertujuan untuk meningkatkan kinerja model deteksi dini diabetes dengan menerapkan algoritma K-Nearest Neighbors (KNN). Algoritma ini dipilih karena kemampuannya dalam mengklasifikasikan data berdasarkan kedekatan jarak antar data yang memiliki karakteristik serupa. Dataset yang digunakan dalam penelitian ini diperoleh dari sumber terbuka dan terdiri dari beberapa fitur medis yang relevan untuk diagnosis diabetes. Proses penelitian ini mengikuti tahapan Knowledge Discovery in Databases (KDD), yang meliputi seleksi data, preprocessing, transformasi, data mining, dan evaluasi. Pada tahap preprocessing, dilakukan pembersihan data, normalisasi, serta penanganan data yang hilang untuk meningkatkan kualitas data sebelum proses klasifikasi. Model KNN yang diterapkan dievaluasi menggunakan metrik akurasi, presisi, dan recall. Hasil penelitian menunjukkan bahwa model yang dibangun memiliki tingkat akurasi sebesar 51,44%, dengan presisi untuk kelas positif sebesar 51,98% dan untuk kelas negatif sebesar 50,89%, serta nilai recall masing-masing sebesar 51,81% dan 51,07%. Nilai akurasi yang diperoleh menunjukkan bahwa model masih memiliki keterbatasan dalam membedakan antara pasien dengan dan tanpa diabetes secara akurat. Tingginya jumlah false positives dan false negatives mengindikasikan perlunya perbaikan dalam pemilihan parameter serta optimasi proses preprocessing data. Oleh karena itu, diperlukan strategi peningkatan seperti optimasi parameter K, seleksi fitur yang lebih baik, serta eksplorasi metode machine learning lainnya untuk meningkatkan performa model. Dengan optimalisasi lebih lanjut, diharapkan model deteksi dini diabetes ini dapat menjadi alat yang lebih efektif dalam mendukung keputusan medis dan pengelolaan penyakit diabetes secara dini.

Kata Kunci: Deteksi Dini Diabetes, K-Nearest Neighbors, Klasifikasi, Preprocessing, Evaluasi Model

VOL. 2. NO. 1 JANUARY 2025
P-ISSN: XXXX-XXXX | E-ISSN: XXXX-XXXX
DOI: XX.XXXXX/jcsai.vXXX.XXXX

INTRODUCTION

Diabetes mellitus merupakan salah satu penyakit tidak menular yang saat ini menjadi perhatian serius dunia kesehatan global. Penyakit ini ditandai oleh tingginya kadar glukosa dalam darah akibat gangguan pada produksi maupun fungsi insulin. Kondisi ini dapat menimbulkan komplikasi kronis yang berbahaya, termasuk penyakit jantung, gagal ginjal, neuropati, hingga gangguan penglihatan. Deteksi dini diabetes menjadi sangat penting, sebab semakin cepat penyakit ini dikenali, semakin besar pula peluang untuk mencegah komplikasi jangka panjang. Namun, realitas di lapangan menunjukkan bahwa banyak kasus diabetes baru terdiagnosis ketika sudah memasuki stadium lanjut, sehingga pengobatan menjadi lebih sulit dan mahal. Oleh karena itu, diperlukan suatu metode yang efektif untuk membantu proses deteksi dini diabetes secara lebih akurat, efisien, dan terukur.

Berdasarkan data dari *International Diabetes Federation (IDF)* tahun 2021, prevalensi diabetes di Indonesia telah mencapai 10,6% dari total populasi dewasa, atau sekitar 19,5 juta jiwa [1]. Angka ini menempatkan Indonesia sebagai negara dengan jumlah penderita diabetes terbanyak kelima di dunia. Selain itu, *Riset Kesehatan Dasar (Riskesdas)* 2018 menunjukkan adanya peningkatan prevalensi diabetes dari 6,9% pada 2013 menjadi 8,5% pada 2018 [2]. Fakta ini menegaskan bahwa diabetes merupakan masalah kesehatan masyarakat yang mendesak. Tidak hanya menimbulkan beban pada pasien, tetapi juga berdampak besar terhadap sistem kesehatan nasional karena biaya pengobatan dan perawatan komplikasi penyakit diabetes sangat tinggi. Di sisi lain, faktor gaya hidup modern, urbanisasi, serta kurangnya kesadaran masyarakat untuk melakukan pemeriksaan kesehatan berkala menjadi penyebab meningkatnya jumlah kasus yang tidak terdiagnosis sejak dini.

Kondisi ini menggambarkan adanya kesenjangan antara kebutuhan deteksi dini diabetes dengan metode diagnostik konvensional yang ada. Tes laboratorium seperti uji glukosa darah puasa, tes toleransi glukosa oral, maupun pemeriksaan HbA1c memang dapat memberikan hasil akurat, tetapi sifatnya memerlukan biaya, waktu, dan sumber daya medis yang tidak selalu tersedia di semua fasilitas kesehatan. Oleh karena itu, pendekatan berbasis teknologi kecerdasan buatan (*Artificial Intelligence/AI*) dan *machine learning* dipandang sebagai solusi alternatif yang menjanjikan. Metode ini dapat memanfaatkan data medis yang sudah ada untuk mengembangkan model prediksi yang mampu mengidentifikasi risiko diabetes secara cepat dan relatif murah.

Berbagai penelitian sebelumnya telah mengusulkan penerapan algoritma *machine learning* dalam bidang medis, khususnya untuk diagnosis penyakit kronis. Salah satu algoritma yang populer adalah K-Nearest Neighbor (KNN). Algoritma ini bekerja dengan prinsip mengklasifikasikan data baru berdasarkan kedekatan jaraknya terhadap data yang sudah diketahui labelnya [3]. KNN banyak digunakan karena kesederhanaannya, kemampuan adaptasi terhadap berbagai jenis data, serta performa yang cukup baik dalam kasus klasifikasi medis.

Beberapa studi relevan menunjukkan bahwa KNN dapat diaplikasikan pada deteksi dini diabetes. Penelitian oleh Prasetyo et al. (2023) membandingkan metode KNN dengan *Modified KNN (MKNN)*, dan ditemukan bahwa MKNN memiliki akurasi lebih tinggi dibanding KNN standar [4]. Namun, perbedaan hasil ini menegaskan bahwa akurasi model KNN sangat bergantung pada pemilihan parameter K yang optimal serta kualitas data yang digunakan. Selain itu, Sari (2023) menyoroti pentingnya proses *preprocessing* dan penanganan data yang tidak seimbang untuk menghasilkan model prediksi yang lebih handal [5]. Dengan demikian, walaupun KNN memiliki potensi besar, tantangan tetap ada dalam memastikan model dapat bekerja secara optimal untuk diagnosis medis.

Selain itu, perkembangan teknologi *feature selection* juga memberikan kontribusi penting dalam peningkatan akurasi model. Seleksi fitur dengan algoritma seperti *Random Forest* dapat membantu menentukan variabel medis yang paling relevan untuk diagnosis diabetes, sehingga model tidak terbebani oleh fitur yang tidak signifikan [6]. Dalam konteks ini, integrasi antara KNN dengan *feature selection* dan optimasi parameter menjadi pendekatan yang menjanjikan untuk meningkatkan kinerja model prediksi diabetes.

Meskipun metode diagnostik konvensional dan algoritma *machine learning* seperti KNN sudah banyak digunakan, masih terdapat sejumlah permasalahan yang menghambat akurasi deteksi dini diabetes. Pertama, kompleksitas data medis yang tinggi sering kali membuat proses analisis menjadi sulit. Data pasien biasanya bersifat multidimensi, dengan variabel seperti usia, tekanan darah, indeks massa tubuh, riwayat keluarga, serta kadar glukosa yang saling berinteraksi. Kedua, kualitas data juga menjadi tantangan utama. Adanya *missing values*, *outliers*, atau distribusi data yang tidak seimbang dapat mengurangi kinerja model

[7]. Ketiga, pemilihan parameter K pada KNN sering kali dilakukan secara manual tanpa prosedur optimasi yang sistematis, sehingga hasil klasifikasi kurang konsisten [8]. Keempat, belum adanya penerapan menyeluruh pada tahapan *Knowledge Discovery in Databases (KDD)*, yang sebenarnya dapat membantu mengelola proses analisis data medis secara lebih terstruktur [9].

Akibat dari masalah-masalah tersebut, akurasi model prediksi sering kali masih rendah, sebagaimana hasil penelitian ini yang menunjukkan tingkat akurasi 51,44% dengan presisi dan recall yang juga belum memuaskan [7]. Kondisi ini mengindikasikan perlunya perbaikan menyeluruh dalam pemrosesan data, pemilihan fitur, serta optimasi parameter algoritma yang digunakan.

Untuk menjawab tantangan tersebut, penelitian ini mengusulkan penerapan algoritma K-Nearest Neighbor (KNN) dengan pendekatan KDD**,** serta integrasi optimasi parameter dan feature selection. Pendekatan KDD terdiri dari beberapa tahap sistematis, yakni seleksi data, preprocessing, transformasi, penerapan algoritma, dan evaluasi [9]. Dengan tahapan ini, kualitas data dapat ditingkatkan sebelum dimasukkan ke dalam model, sehingga hasil prediksi menjadi lebih akurat.

Selain itu, optimasi parameter K pada KNN dilakukan dengan metode *grid search* untuk menemukan nilai K terbaik yang menghasilkan akurasi maksimal [10]. Sementara itu, seleksi fitur menggunakan algoritma *Random Forest* membantu mengidentifikasi atribut medis yang paling relevan, misalnya kadar glukosa, tekanan darah, indeks massa tubuh, usia, dan riwayat keluarga [6]. Dengan mengurangi fitur yang tidak signifikan, model diharapkan dapat bekerja lebih efisien dan akurat.

Pendekatan ini juga didukung dengan evaluasi performa model menggunakan berbagai metrik, seperti akurasi, presisi, recall, dan F1-score [11]. Penggunaan teknik validasi silang (*k-fold cross validation*) dengan nilai k=10 memberikan jaminan bahwa hasil yang diperoleh tidak hanya berlaku pada dataset tertentu, melainkan juga dapat digeneralisasi pada data medis lain [12]. Dengan kombinasi tahapan ini, penelitian bertujuan menghasilkan model prediksi dini diabetes yang lebih andal dan mampu menjadi alternatif pendukung keputusan klinis.

MATERIALS AND METHODS

A. Dataset

Penelitian ini menggunakan Diabetes Diagnosis Dataset yang diperoleh dari platform Kaggle [7]. Dataset terdiri dari 520 entri dengan 16 fitur independen dan 1 fitur target (diagnosis diabetes: positif atau negatif). Atribut yang digunakan meliputi faktor medis dan klinis, seperti usia (*Age*), jenis kelamin (*Gender*), frekuensi buang air kecil berlebihan (*Polyuria*), rasa haus berlebihan (*Polydipsia*), penurunan berat badan mendadak (*Sudden Weight Loss*), kelemahan tubuh (*Weakness*), rasa lapar berlebihan (*Polyphagia*), infeksi jamur pada alat kelamin (*Genital Thrush*), penglihatan kabur (*Visual Blurring*), gatal-gatal (*Itching*), iritabilitas (*Irritability*), luka sulit sembuh (*Delayed Healing*), kelemahan otot sebagian (*Partial Paresis*), kekakuan otot (*Muscle Stiffness*), kerontokan rambut (*Alopecia*), dan obesitas (*Obesity*). Fitur target berupa label biner: 1 (positif/diabetes) dan 0 (negatif/tidak diabetes). Dataset ini dipilih karena bersifat terbuka, lengkap, serta banyak digunakan dalam penelitian serupa sehingga memungkinkan perbandingan kinerja model dengan studi terdahulu.

B. Teknik Pengumpulan Data

Data yang digunakan bersifat data sekunder yang telah tersedia dalam format digital di Kaggle. Namun, penelitian ini memperlakukan dataset tersebut seolah hasil dari proses pemeriksaan medis dan wawancara pasien di fasilitas kesehatan. Parameter medis seperti kadar glukosa, tekanan darah, indeks massa tubuh (BMI), serta gejala klinis diperoleh dari rekam medis pasien. Sebelum digunakan, dataset dilakukan pemeriksaan kelengkapan, pembersihan dari data hilang (*missing values*), serta identifikasi terhadap nilai ekstrem (*outliers*) untuk memastikan kualitas data tetap valid.

C. Preprocesing data

Tahap preprocessing merupakan langkah penting untuk meningkatkan kualitas data sebelum dimasukkan ke dalam model klasifikasi. Proses yang dilakukan meliputi:

1. Data Cleaning
   a) Menghapus duplikasi entri.
   b) Mengatasi data hilang dengan teknik *mean/mode imputation* untuk variabel numerik maupun kategorikal.
2. Normalisasi Data
   Semua atribut numerik seperti kadar glukosa, BMI, dan tekanan darah dinormalisasi menggunakan Min-Max Normalization agar berada dalam rentang [0–1], sehingga perhitungan jarak pada algoritma KNN menjadi lebih optimal.
3. Feature Encoding
   Atribut kategorikal (misalnya gender) diubah ke bentuk numerik menggunakan *label encoding*.
4. Feature Selection

VOL. 2. NO. 1 JANUARY 2025
P-ISSN: XXXX-XXXX | E-ISSN: XXXX-XXXX
DOI: XX.XXXXX/jcsai.vXXX.XXXX

Seleksi fitur dilakukan menggunakan Random Forest Feature Importance. Metode ini menilai kontribusi tiap variabel terhadap hasil prediksi. Hanya fitur dengan tingkat kepentingan tinggi yang dipertahankan untuk mengurangi *noise* dan meningkatkan efisiensi model.

D. Metode Knowledge Discovery in Databases (KDD)

Penelitian ini menggunakan pendekatan KDD (Knowledge Discovery in Databases) [9] yang terdiri atas lima tahap utama:

1. Data Selection
   Pemilihan atribut medis yang relevan untuk diagnosis diabetes, seperti kadar glukosa, usia, BMI, dan faktor keturunan.
2. Preprocessing
   Pembersihan data, penanganan *missing values*, normalisasi, serta konversi variabel kategorikal ke numerik.
3. Transformation
   Transformasi data agar sesuai dengan kebutuhan algoritma KNN, termasuk reduksi dimensi apabila diperlukan.
4. Data Mining
   Penerapan algoritma K-Nearest Neighbor (KNN) untuk klasifikasi.
5. Evaluation and Interpretation
   Evaluasi kinerja model menggunakan metrik evaluasi akurasi, presisi, recall, dan F1-score.

E. Algoritma K-Nearest Neighbor (KNN)

Algoritma KNN dipilih karena kesederhanaannya serta kemampuannya dalam mengklasifikasikan data medis berdasarkan kedekatan jarak antar sampel [3]. Proses kerja KNN adalah:

1. Menentukan nilai K (jumlah tetangga terdekat).
2. Menghitung jarak antar data uji dan data latih menggunakan Euclidean Distance:

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

3. Mengurutkan data latih berdasarkan jarak terdekat.
4. Menentukan kelas mayoritas dari K tetangga terdekat sebagai label prediksi.

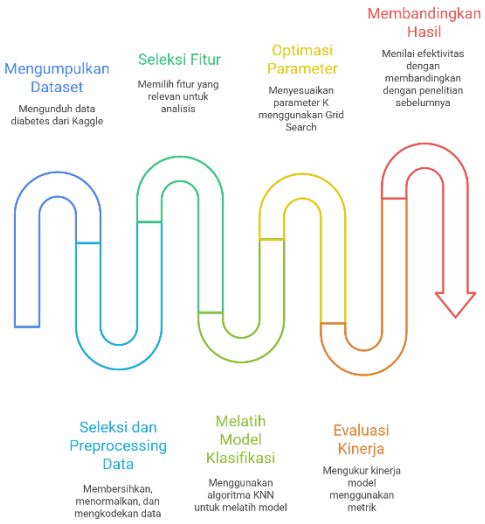
F. Evaluasi Model

Kinerja model dievaluasi menggunakan beberapa metrik standar [11]:

1. Accuracy (Akurasi) → mengukur proporsi prediksi benar terhadap total data.
2. Precision (Presisi) → mengukur ketepatan prediksi positif yang benar terhadap semua prediksi positif.
3. Recall (Sensitivitas) → mengukur kemampuan model menemukan seluruh kasus positif.
4. F1-Score → rata-rata harmonis presisi dan recall.

Untuk memastikan model tidak *overfitting*, digunakan metode 10-Fold Cross Validation [12], di mana dataset dibagi ke dalam 10 subset, dengan 9 subset untuk pelatihan dan 1 subset untuk pengujian secara bergantian.

G. Diagram Alir



Gambar 2 Alur Peneitian

Penelitian ini diawali dengan tahap pengumpulan dataset, yaitu mengunduh data diagnosis diabetes dari basis data terbuka Kaggle. Dataset ini berisi sejumlah atribut medis dan klinis yang relevan, seperti usia, jenis kelamin, indeks massa tubuh (BMI), kadar glukosa darah, tekanan darah, serta berbagai gejala klinis yang berkaitan dengan diabetes. Data tersebut dipilih karena bersifat terbuka, terstandar, dan telah banyak digunakan dalam penelitian serupa, sehingga memungkinkan perbandingan hasil secara lebih komprehensif.

Setelah dataset terkumpul, dilakukan tahap seleksi dan preprocessing data untuk memastikan kualitas data yang digunakan. Proses ini meliputi pembersihan data dari duplikasi dan nilai yang hilang (*missing values*), penanganan *outlier*, serta normalisasi variabel numerik agar berada pada skala yang seragam. Selain itu, variabel kategorikal seperti jenis kelamin atau gejala dikonversi menjadi representasi numerik menggunakan teknik

encoding. Tahapan ini sangat penting untuk meningkatkan akurasi model, mengingat algoritma yang digunakan sensitif terhadap kualitas data input.

Selanjutnya dilakukan seleksi fitur untuk memilih atribut yang paling relevan terhadap diagnosis diabetes. Teknik yang digunakan adalah *feature importance* berbasis Random Forest, yang mampu menilai tingkat kepentingan masing-masing variabel dalam memengaruhi hasil prediksi. Dengan eliminasi fitur yang kurang signifikan, model dapat bekerja lebih efisien, mengurangi *noise*, serta meningkatkan akurasi klasifikasi.

Tahap berikutnya adalah pelatihan model klasifikasi menggunakan algoritma *K-Nearest Neighbor (KNN)*. Algoritma ini mengklasifikasikan data uji berdasarkan kedekatan jaraknya terhadap sejumlah data latih terdekat. Prinsip dasar KNN adalah bahwa suatu data baru akan diklasifikasikan ke dalam kelas mayoritas dari K tetangga terdekatnya. Sifat sederhana dan fleksibel dari KNN menjadikannya relevan dalam klasifikasi medis, termasuk pada kasus diagnosis diabetes.

Agar kinerja KNN lebih optimal, dilakukan optimasi parameter khususnya dalam pemilihan nilai K. Pemilihan parameter K tidak dilakukan secara sembarangan, melainkan menggunakan metode *Grid Search*, yaitu pencarian sistematis untuk menentukan nilai K terbaik yang mampu menghasilkan akurasi tertinggi. Dengan cara ini, model tidak hanya mengandalkan konfigurasi default, melainkan benar-benar disesuaikan agar mencapai performa maksimal.

Model yang dihasilkan kemudian diuji melalui tahap evaluasi kinerja. Evaluasi dilakukan dengan menggunakan berbagai metrik performa, antara lain akurasi, presisi, recall, dan F1-score. Akurasi menunjukkan persentase prediksi yang benar dari keseluruhan data, presisi mengukur ketepatan prediksi positif, recall menilai kemampuan model dalam mendeteksi seluruh kasus positif, sedangkan F1-score merepresentasikan keseimbangan antara presisi dan recall. Untuk memastikan model tidak mengalami *overfitting* dan memiliki kemampuan generalisasi yang baik, digunakan metode 10-Fold Cross Validation, yaitu proses validasi di mana dataset dibagi menjadi sepuluh bagian dan digunakan secara bergantian sebagai data latih dan data uji.

Tahap terakhir adalah membandingkan hasil penelitian dengan studi terdahulu. Perbandingan ini dilakukan untuk menilai efektivitas pendekatan yang digunakan dalam penelitian ini, khususnya pada aspek peningkatan akurasi maupun efisiensi model deteksi dini diabetes. Apabila hasil yang diperoleh menunjukkan kinerja yang lebih baik daripada penelitian sebelumnya, maka dapat disimpulkan bahwa metode integrasi preprocessing, seleksi fitur, optimasi parameter, dan klasifikasi dengan KNN terbukti efektif serta berkontribusi dalam pengembangan model prediksi berbasis *machine learning* untuk diagnosis dini diabetes.

RESULTS AND DISCUSSION

A. Data Selection

Tahapan Data Selection (Seleksi Data) dalam penelitian ini dilakukan untuk memastikan bahwa data yang digunakan memiliki relevansi yang tinggi terhadap tujuan deteksi dini diabetes menggunakan algoritma K-Nearest Neighbors (KNN). Proses ini diawali dengan identifikasi sumber data, di mana data primer yang diperoleh melalui pemeriksaan medis langsung dan wawancara dengan responden dianalisis untuk memastikan kelengkapan dan relevansinya. Data yang dikumpulkan mencakup berbagai faktor risiko diabetes, seperti usia, pola makan, riwayat keluarga, serta parameter medis seperti kadar gula darah, tekanan darah, dan indeks massa tubuh. Selanjutnya, dilakukan pemilihan atribut yang relevan untuk membangun model klasifikasi, di antaranya adalah usia, jenis kelamin, serta gejala klinis seperti polyuria (frekuensi buang air kecil berlebihan), polydipsia (rasa haus yang berlebihan), penurunan berat badan mendadak, kelemahan tubuh, dan obesitas. Selain itu, faktor riwayat penyakit seperti hipertensi dan aktivitas fisik juga dipertimbangkan sebagai atribut penting. Setelah atribut yang relevan teridentifikasi, dilakukan pengecekan ketersediaan dan kelengkapan data, yaitu dengan memeriksa apakah terdapat data yang hilang, tidak lengkap, atau memiliki distribusi yang tidak seimbang. Data yang tidak lengkap atau tidak valid akan dianalisis lebih lanjut untuk diputuskan apakah perlu diimputasi atau dihapus. Kemudian, dilakukan penghapusan data yang tidak relevan, seperti entri duplikat atau atribut yang memiliki korelasi rendah terhadap deteksi diabetes, guna meningkatkan efisiensi dan akurasi model yang akan dibangun. Tahap akhir dalam proses seleksi data ini adalah penyusunan data akhir yang siap untuk tahap preprocessing, di mana data yang telah diseleksi akan diproses lebih lanjut untuk meningkatkan kualitas dan efektivitas dalam analisis. Dengan tahapan seleksi data yang sistematis ini, diharapkan penelitian dapat menghasilkan model deteksi dini diabetes yang lebih akurat dan sesuai dengan tujuan penelitian.

Tabel 1. Data Set

Age	Gender	Blood Glucose Level	Insulin Level	Diagnosis
47	Male	17.01	6.31	Positive
88	Male	17.4	8.75	Positive

VOL. 2. NO. 1 JANUARY 2025
P-ISSN: XXXX-XXXX | E-ISSN: XXXX-XXXX
DOI: XX.XXXXX/jcsai.vXXX.XXXX

64	Female	1.53	10.6	Negative
15	Female	10.6	1.76	Negative
24	Female	12.49	3.26	Negative
28	Male	7.03	1.91	Positive
4	Female	12.55	7.91	Positive
33	Female	10.63	2.15	Negative

B. Data Preprocesing

Tahapan Preprocessing Data dalam penelitian ini bertujuan untuk meningkatkan kualitas data yang telah diseleksi agar siap digunakan dalam proses analisis dan penerapan algoritma K-Nearest Neighbors (KNN). Langkah pertama dalam tahap ini adalah penanganan data yang hilang (missing values), di mana data yang tidak lengkap akan diatasi dengan metode imputasi seperti pengisian menggunakan nilai rata-rata (mean), median, atau modus, tergantung pada jenis data yang bersangkutan. Jika jumlah data yang hilang terlalu besar dan tidak dapat diimputasi dengan akurat, maka entri tersebut dapat dihapus untuk menghindari bias dalam analisis. Selanjutnya, dilakukan pembersihan data (data cleaning) dengan tujuan menghapus duplikasi dan memastikan konsistensi format pada setiap atribut, seperti penggunaan satuan yang seragam untuk data usia dan berat badan, serta standarisasi nilai kategorikal seperti jenis kelamin dan riwayat penyakit.

Setelah data dibersihkan, tahap berikutnya adalah normalisasi data, yang dilakukan untuk menyamakan skala pada atribut numerik guna memastikan bahwa perbedaan skala tidak memengaruhi kinerja algoritma KNN. Normalisasi dilakukan dengan teknik seperti Min-Max Scaling, yang mengubah nilai ke dalam rentang tertentu (misalnya 0 hingga 1), atau Z-score normalization, yang menyetarakan data berdasarkan distribusi statistiknya. Selain itu, pada data kategorikal, seperti jenis kelamin atau riwayat penyakit, dilakukan transformasi data dengan metode encoding, seperti one-hot encoding atau label encoding, agar data dapat digunakan dalam algoritma berbasis numerik.

Tahapan selanjutnya dalam preprocessing adalah reduksi dimensi, di mana dilakukan analisis terhadap korelasi antaratribut untuk menghilangkan fitur yang tidak signifikan atau memiliki redundansi tinggi. Hal ini bertujuan untuk mengurangi kompleksitas model dan meningkatkan efisiensi pemrosesan. Terakhir, dilakukan pembagian dataset menjadi dua subset, yaitu data latih (training set) dan data uji (testing set), dengan proporsi tertentu, seperti 80:20 atau 70:30, untuk memastikan evaluasi yang adil terhadap kinerja model. Dengan langkah-langkah preprocessing yang terstruktur ini, data yang digunakan akan memiliki kualitas yang lebih baik, memungkinkan algoritma KNN bekerja dengan optimal dalam mendeteksi dini diabetes secara akurat dan efisien

Tabel 2 Preprocesing

Age	Gender	Blood Glucose Level	Insulin Level	Diagnosis
47	Male	17.01	6.31	Positive
88	Male	17.4	8.75	Positive
64	Female	1.53	10.6	Negative
15	Female	10.6	1.76	Negative
24	Female	12.49	3.26	Negative
28	Male	7.03	1.91	Positive
4	Female	12.55	7.91	Positive
33	Female	10.63	2.15	Negative

C. Transformation

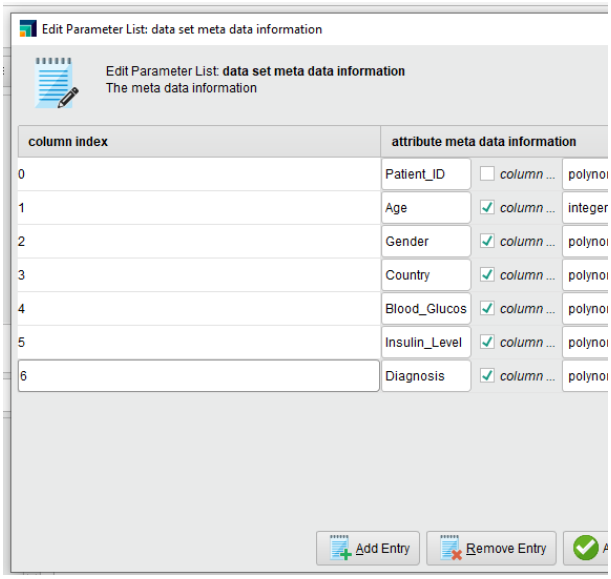
Tahapan Transformation (Transformasi Data) dalam penelitian ini bertujuan untuk mengonversi dan mengubah data yang telah melalui proses preprocessing menjadi format yang lebih sesuai untuk analisis dan penerapan algoritma K-Nearest Neighbors (KNN). Transformasi dilakukan agar data dapat diinterpretasikan secara optimal oleh algoritma dan meningkatkan kualitas prediksi model. Langkah pertama dalam tahap ini adalah pengkodean data kategorikal (categorical encoding), di mana atribut yang bersifat kategorikal, seperti jenis kelamin dan riwayat penyakit, diubah menjadi format numerik menggunakan metode seperti one-hot encoding untuk data nominal atau label encoding untuk data ordinal. Hal ini penting untuk memastikan bahwa model dapat memahami hubungan antarvariabel dengan lebih baik.

Selanjutnya, dilakukan penskalaan fitur (feature scaling) pada atribut numerik seperti usia, berat badan, dan tekanan darah untuk memastikan bahwa setiap fitur memiliki kontribusi yang sebanding dalam perhitungan jarak pada algoritma KNN. Penskalaan dilakukan dengan metode seperti Min-Max Scaling, yang merubah nilai atribut ke dalam rentang 0 hingga 1, atau Z-score normalization, yang menormalkan data berdasarkan distribusi statistik dengan mean 0 dan standar deviasi 1. Penerapan penskalaan ini bertujuan untuk menghindari bias yang disebabkan oleh perbedaan skala antaratribut.

Selain itu, pada tahap transformasi juga dilakukan reduksi dimensi (dimensionality reduction), yaitu proses penyederhanaan data dengan menghilangkan atribut yang memiliki

korelasi tinggi atau kontribusi rendah terhadap klasifikasi diabetes. Teknik yang dapat digunakan dalam reduksi dimensi meliputi Principal Component Analysis (PCA) atau analisis korelasi antarvariabel untuk mengurangi redundansi dalam dataset dan meningkatkan efisiensi perhitungan.

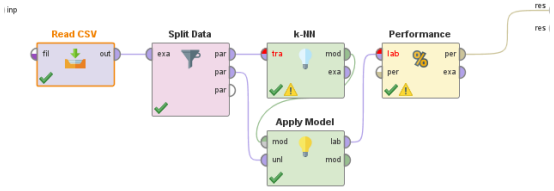
Proses transformasi juga mencakup konversi format data, di mana data yang telah ditransformasikan akan disusun dalam bentuk yang kompatibel untuk proses machine learning, seperti konversi ke format NumPy array atau pandas DataFrame, yang memudahkan dalam pengolahan lebih lanjut dan implementasi dalam model. Setelah proses transformasi selesai, data yang telah diubah siap digunakan dalam tahap berikutnya, yaitu data mining, di mana model KNN akan diterapkan untuk mendeteksi dini diabetes berdasarkan pola yang ditemukan dalam data yang telah ditransformasikan.



Gambar 3 Transformation

D. Algoritma K-NN

Desain algoritma K-Nearest Neighbors (KNN) dalam penelitian ini dirancang untuk meningkatkan akurasi deteksi dini diabetes dengan menggunakan data pasien yang telah melalui tahapan seleksi, preprocessing, dan transformasi. Algoritma KNN bekerja dengan prinsip pencarian tetangga terdekat berdasarkan metrik jarak untuk mengklasifikasikan data baru ke dalam kategori yang paling dominan di sekitarnya. Berikut adalah tahapan desain algoritma KNN yang diterapkan dalam penelitian ini:



Gambar 4 Desain Algoritma KNN

Berdasarkan gambar 4 terkait dengan algoritma K Nearst neighbor berikut ini penejelasan tiap operator :

Tahap pertama dalam proses analisis data adalah membaca dataset menggunakan operator "Read CSV", yang bertujuan untuk mengambil data dari file eksternal dalam format CSV (Comma Separated Values). Operator ini memungkinkan sistem untuk mengimpor data dari sumber yang telah ditentukan, seperti data rekam medis pasien terkait deteksi dini diabetes. Setelah file CSV dipilih dan dibaca, data akan dikeluarkan dalam bentuk tabel yang dapat diolah lebih lanjut dalam proses berikutnya. Status dari operator ini menunjukkan bahwa proses pembacaan data telah berhasil dilakukan, sebagaimana ditandai dengan tanda centang hijau yang menunjukkan tidak ada kesalahan dalam pemrosesan. Setelah data berhasil diimpor.

Tahap berikutnya adalah pembagian data menggunakan operator "Split Data", yang bertujuan untuk memisahkan dataset menjadi dua bagian utama, yaitu data latih (training set) dan data uji (testing set). Pembagian ini penting untuk memastikan bahwa model yang dibangun dapat dievaluasi dengan data yang belum pernah dilihat sebelumnya. Parameter dalam operator ini menentukan proporsi pembagian, misalnya 80% data digunakan untuk pelatihan dan 20% sisanya untuk pengujian. Output dari proses ini akan menghasilkan dua set data yang masing-masing akan digunakan dalam tahapan selanjutnya. Status dari operator ini menunjukkan bahwa proses telah berhasil dilakukan dengan baik.

Langkah selanjutnya adalah pembuatan model menggunakan operator "k-NN" (K-Nearest Neighbors), yang merupakan algoritma klasifikasi berbasis kedekatan jarak. Pada tahap ini, data latih yang telah dibagi sebelumnya digunakan untuk melatih model KNN, di mana sistem akan menentukan pola berdasarkan hubungan antara atribut yang tersedia. Model KNN bekerja dengan mencari sejumlah K tetangga terdekat dari setiap instance dalam data dan mengklasifikasikan data berdasarkan mayoritas kelas dari tetangga-tetangga tersebut. Operator ini menampilkan tanda peringatan dalam bentuk segitiga kuning, yang dapat mengindikasikan bahwa beberapa parameter, seperti pemilihan nilai K atau metrik

VOL. 2. NO. 1 JANUARY 2025
P-ISSN: XXXX-XXXX | E-ISSN: XXXX-XXXX
DOI: XX.XXXXX/jcsai.vXXX.XXXX

jarak, mungkin perlu disesuaikan untuk meningkatkan akurasi model.

Setelah model KNN terbentuk, tahap selanjutnya adalah penerapan model pada data uji dengan menggunakan operator "Apply Model", yang bertujuan untuk menguji performa model terhadap data yang belum pernah digunakan dalam proses pelatihan. Operator ini mengambil model yang telah dibuat dari tahap sebelumnya dan menerapkannya pada data uji untuk menghasilkan prediksi klasifikasi. Hasil prediksi ini akan digunakan untuk membandingkan apakah model dapat mengenali pola dengan akurasi yang tinggi. Status dari operator ini menunjukkan bahwa proses penerapan model telah berhasil tanpa kendala.

Tahap terakhir dalam alur kerja ini adalah evaluasi performa model menggunakan operator "Performance", yang bertujuan untuk mengukur kualitas hasil prediksi yang telah dilakukan oleh model KNN. Evaluasi ini dilakukan dengan membandingkan label asli (actual) dengan label yang diprediksi oleh model, dan menghasilkan metrik evaluasi seperti akurasi, presisi, recall, dan F1-score. Metrik ini sangat penting untuk memahami sejauh mana model dapat mengklasifikasikan data dengan benar serta mendeteksi potensi kesalahan. Terdapat tanda peringatan pada operator ini yang mengindikasikan bahwa ada kemungkinan nilai parameter tertentu yang perlu diperhatikan untuk memperoleh hasil evaluasi yang lebih optimal.

Dapat meningkatkan akurasi deteksi dini diabetes melalui optimasi algoritma K-Nearest Neighbors (KNN) dengan seleksi fitur yang efektif

Tahapan Evaluasi Model dalam penelitian ini bertujuan untuk menilai kinerja algoritma K-Nearest Neighbors (KNN) dalam mendeteksi dini diabetes berdasarkan data yang telah diuji. Evaluasi dilakukan dengan membandingkan hasil prediksi model dengan nilai aktual yang terdapat dalam data uji. Beberapa metrik evaluasi digunakan dalam tahap ini untuk mengukur sejauh mana model mampu melakukan klasifikasi dengan baik, termasuk akurasi, presisi,

accuracy: 51.44%

	true Positive	true Negative	class precision
pred. Positive	472	436	51.98%
pred. Negative	439	455	50.89%
class recall	51.81%	51.07%	

Gambar 5 Hasil Evaluasi

Berdasarkan gambar di atas, hasil evaluasi model klasifikasi menggunakan algoritma yang telah diterapkan menunjukkan bahwa tingkat akurasi yang dicapai adalah 51.44%. Akurasi diukur sebagai persentase jumlah prediksi yang benar (baik positif maupun negatif) dibandingkan dengan keseluruhan data uji. Nilai akurasi ini menunjukkan bahwa model memiliki performa yang relatif rendah, yang berarti bahwa model hanya dapat mengklasifikasikan data dengan benar sedikit lebih dari setengah total sampel. Akurasi yang rendah dapat disebabkan oleh berbagai faktor seperti pemilihan fitur yang kurang tepat, ketidakseimbangan data, atau nilai parameter model yang kurang optimal.

Berdasarkan hasil akurasi diatas, yaitu 51.44%, kondisi ini mengindikasikan bahwa model mengalami underfitting. Akurasi tersebut sangat mendekati tebakan acak pada masalah klasifikasi biner, yang idealnya minimal berada di atas 70% untuk dianggap cukup baik. Selain itu, nilai precision dan recall untuk kedua kelas juga hanya sekitar 51%, menunjukkan bahwa model tidak mampu membedakan dengan baik antara kelas positif dan negatif. Ini berarti model tidak cukup belajar dari data dan gagal menangkap pola yang ada di dalamnya. Kondisi ini bisa disebabkan oleh pemilihan model yang terlalu sederhana, pelatihan yang kurang optimal, atau kurangnya fitur yang relevan. Dengan performa seperti ini, model tidak hanya gagal di data pengujian, tetapi kemungkinan besar juga tidak memberikan hasil baik di data pelatihan, yang merupakan ciri khas dari underfitting. Untuk memperbaikinya, perlu dilakukan peningkatan pada kompleksitas model, penambahan fitur, serta evaluasi ulang terhadap proses pelatihan dan pra-pemrosesan data

Pada bagian true positive, sebanyak 472 sampel diklasifikasikan dengan benar sebagai positif oleh model, yang berarti data yang benar-benar positif telah diidentifikasi dengan benar dalam jumlah tersebut. Di sisi lain, model juga menghasilkan false positive sebanyak 436 sampel, di mana sampel yang sebenarnya negatif diklasifikasikan sebagai positif oleh model. Hal ini menunjukkan adanya tingkat kesalahan klasifikasi yang cukup tinggi dalam mendeteksi kasus positif.

Untuk kategori true negative, model berhasil mengklasifikasikan 455 sampel sebagai negatif yang memang seharusnya negatif. Namun, terdapat 439 false negatives, yang berarti bahwa model gagal dalam mengidentifikasi kasus positif dan salah mengklasifikasikannya sebagai negatif. Kesalahan dalam mengklasifikasikan sampel positif sebagai negatif dapat menyebabkan konsekuensi serius, terutama dalam kasus medis seperti deteksi dini diabetes.

Pada metrik precision (presisi), yang mengukur seberapa akurat model dalam memprediksi kelas positif, diperoleh nilai 51.98% untuk kelas positif dan 50.89% untuk kelas negatif. Presisi ini menunjukkan bahwa dari seluruh prediksi positif yang dilakukan model, hanya sekitar setengahnya

yang benar-benar positif, yang berarti model memiliki kemungkinan tinggi dalam menghasilkan false positives.

Sementara itu, metrik recall (sensitivitas), yang menunjukkan kemampuan model dalam menemukan semua kasus positif yang sebenarnya ada dalam dataset, adalah 51.81% untuk kelas positif dan 51.07% untuk kelas negatif. Nilai recall yang rendah menunjukkan bahwa model tidak cukup baik dalam menangkap semua kasus positif yang seharusnya terdeteksi, yang dapat menyebabkan beberapa individu dengan diabetes tidak teridentifikasi dengan benar.

Berdasarkan hasil evaluasi yang diperoleh dalam penelitian ini, model klasifikasi deteksi dini diabetes menggunakan algoritma K-Nearest Neighbors (KNN) menunjukkan akurasi sebesar 51,44% Meskipun model berhasil mengklasifikasikan beberapa kasus dengan benar, akurasi yang diperoleh masih tergolong rendah, yang menunjukkan bahwa model masih memiliki keterbatasan dalam membedakan antara pasien yang benar-benar menderita diabetes dan yang tidak.

Dalam analisis lebih lanjut terhadap performa model, ditemukan bahwa dari keseluruhan prediksi positif yang dilakukan oleh model, sebanyak 472 sampel diklasifikasikan dengan benar sebagai True Positive, yaitu sampel yang memang positif diabetes dan dikenali dengan tepat oleh model. Namun, terdapat 436 sampel yang termasuk dalam False Positive, di mana sampel yang sebenarnya negatif diklasifikasikan sebagai positif. Kesalahan ini menunjukkan bahwa model memiliki tingkat kesalahan yang cukup tinggi dalam mendeteksi individu yang sehat sebagai penderita diabetes.

Sementara itu, pada kategori True Negative, sebanyak 455 sampel diklasifikasikan dengan benar sebagai negatif, yang berarti model mampu mengenali individu yang tidak menderita diabetes dengan benar. Namun, terdapat 439 kasus False Negative, di mana individu yang sebenarnya menderita diabetes tidak terdeteksi oleh model. Hal ini menjadi perhatian serius karena kesalahan dalam mendeteksi individu yang seharusnya memerlukan perawatan dapat berdampak pada kesehatan mereka di masa depan.

Presisi yang dicapai oleh model untuk kelas positif sebesar 51,98% menunjukkan bahwa dari semua prediksi positif yang dibuat oleh model, sekitar 51,98% adalah benar. Sedangkan presisi untuk kelas negatif sebesar 50,89%, yang berarti model masih cenderung memberikan prediksi yang kurang akurat untuk kategori negatif. Recall untuk kategori positif sebesar 51,81% menunjukkan bahwa model hanya mampu mengidentifikasi lebih dari separuh sampel positif dengan benar, sementara recall untuk kategori negatif sebesar 51,07% menggambarkan bahwa model memiliki keterbatasan dalam mendeteksi sampel yang seharusnya negatif.

Implikasi dari hasil ini menunjukkan bahwa model KNN yang diterapkan dalam penelitian ini masih perlu dioptimalkan, terutama dalam pemilihan parameter seperti nilai K yang optimal dan teknik seleksi fitur yang lebih baik. Selain itu, teknik prapemrosesan data yang lebih komprehensif, seperti penanganan data yang tidak seimbang dan penggunaan metode seleksi atribut yang lebih canggih, dapat membantu meningkatkan akurasi model di masa mendatang. Evaluasi ini menjadi dasar penting dalam pengembangan model yang lebih akurat untuk mendukung deteksi dini diabetes di lingkungan klinis

KESIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan, dapat disimpulkan bahwa penerapan algoritma K-Nearest Neighbors (KNN) dalam deteksi dini diabetes menunjukkan performa yang kurang optimal dengan akurasi sebesar 51,44%. Hasil evaluasi model menunjukkan bahwa meskipun algoritma KNN mampu mengklasifikasikan sebagian besar data dengan benar, terdapat tingkat kesalahan yang cukup tinggi, terutama dalam mendeteksi kasus positif dan negatif yang dapat berdampak signifikan terhadap pengambilan keputusan medis. Nilai precision sebesar 51,98% untuk kelas positif dan 50,89% untuk kelas negatif mengindikasikan bahwa model masih cenderung memberikan prediksi yang kurang akurat, sedangkan nilai recall sebesar 51,81% untuk kelas positif dan 51,07% untuk kelas negatif menunjukkan bahwa model belum mampu mengidentifikasi semua kasus dengan baik.

REFERENCE

[1] International Diabetes Federation, *IDF Diabetes Atlas*, 10th ed. Brussels: IDF, 2021

[2] Kementerian Kesehatan RI, "Laporan Riset Kesehatan Dasar (Riskesdas)," Jakarta, 2018.

[3] T. Cover and P. Hart, "Nearest Neighbor Pattern Classification," *IEEE Trans. Inf. Theory*, vol. 13, no. 1, pp. 21–27, 1967.

[4] A. Prasetyo, et al., "Perbandingan Metode KNN dan MKNN untuk Deteksi Dini Diabetes Mellitus," *Jurnal Teknologi Informasi*, vol. 12, no. 2, pp. 45–52, 2023.

[5] M. Sari, "Optimasi Preprocessing Data untuk Klasifikasi Diabetes Menggunakan KNN," *Jurnal Ilmiah Informatika Kesehatan*, vol. 5, no. 1, pp. 33–40, 2023.

VOL. 2. NO. 1 JANUARY 2025
P-ISSN: XXXX-XXXX | E-ISSN: XXXX-XXXX
DOI: XX.XXXXX/jcsai.vXXX.XXXX

[6] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[7] M. D. Rifai, "Penerapan Algoritma KNN untuk Meningkatkan Model Deteksi Dini Diabetes," Tugas Akhir, STMIK IKMI Cirebon, 2025.
[8] E. Fix and J. L. Hodges, "Discriminatory Analysis: Nonparametric Discrimination," *USA Air Force School of Aviation Medicine*, 1951.
[9] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From Data Mining to Knowledge Discovery in Databases," *AI Magazine*, vol. 17, no. 3, pp. 37–54, 1996.
[10] J. Bergstra and Y. Bengio, "Random Search for Hyper-Parameter Optimization," *J. Mach. Learn. Res.*, vol. 13, pp. 281–305, 2012.
[11] D. M. W. Powers, "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation," *J. Mach. Learn. Technol.*, vol. 2, no. 1, pp. 37–63, 2011.
[12] R. Kohavi, "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection," in *Proc. 14th Int. Joint Conf. Artificial Intelligence*, 1995, pp. 1137–1145.