https://jurnal.pustakagalerimandiri.co.id/index.php/pustakadata
DOI: https://doi.org/10.55382/jurnalpustakadata.v5i2.1550

Jurnal Pustaka Data
Pusat Akses Kajian Database, Analisa Teknologi, dan Arsitektur Komputer
Vol. 5 No. 2 (2025) 481 – 487
E ISSN: 2809-7483

Prediksi Risiko Diabetes Tahap Awal Menggunakan *Machine Learning* dengan Algoritma *K-Nearest Neighbor*

Novi Trisna[1], Raja Ayu Mahessya[2]
[1]Sistem Informasi, Ilmu Komputer, Universiitas Putra Indonesia YPTK Padang
[2]Teknik Informatika, Ilmu Komputer, Universiitas Putra Indonesia YPTK Padang
[1]novi_trisna@upiyptk.ac.id, [2]ayumahessya@upiyptk.ac.id

Abstract

Diabetes mellitus is a chronic metabolic disease with an increasing global prevalence and often remains undetected until complications emerge. Early-stage risk prediction based on initial symptoms is therefore critical to support preventive healthcare interventions. This study proposes a baseline classification model for early-stage diabetes risk using the K-Nearest Neighbor (KNN) algorithm on the publicly available Early Stage Diabetes Risk Prediction dataset. The dataset consists of 520 patient records with 17 attributes, including age, gender, and early clinical symptoms. Categorical attributes were transformed into numerical representations to enable distance-based computation. A systematic evaluation was conducted by analyzing the sensitivity of the KNN model to parameter selection under four data-splitting scenarios (60:40, 70:30, 80:20, and 90:10) and varying K values from 2 to 9 using Euclidean distance. Experimental results indicate that K = 3 consistently yields the highest classification performance, achieving a maximum accuracy of 97% under the 60:40 split scenario. Precision, recall, and F1-score analysis further demonstrate strong performance, particularly in identifying positive diabetes cases. This study provides a systematic baseline evaluation of KNN parameter sensitivity under multiple data-splitting scenarios for early diabetes risk detection. While the model demonstrates competitive performance, it remains limited by its sensitivity to feature scaling, class imbalance, and the absence of cross-validation. Therefore, the proposed model is intended as a benchmark reference rather than a state-of-the-art solution, offering a foundation for future studies incorporating normalization, imbalance handling, and advanced classifiers.

Keywords: Diabetes mellitus, Early Risk Prediction, K-Nearest Neighbor, Classification, K-Value, Baseline.

Abstrak

Diabetes mellitus merupakan penyakit metabolik kronis yang prevalensinya terus meningkat dan berkontribusi signifikan terhadap beban kesehatan masyarakat global akibat komplikasinya yang bersifat progresif dan sering terdeteksi pada tahap lanjut. Oleh karena itu, deteksi dini risiko diabetes berdasarkan gejala awal dan riwayat kesehatan pasien menjadi penting untuk mendukung intervensi preventif yang tepat waktu. Perkembangan *machine learning* memungkinkan pemanfaatan data kesehatan untuk membangun model prediksi yang lebih cepat, konsisten, dan objektif dibandingkan pendekatan manual konvensional. Penelitian ini bertujuan untuk mengembangkan model klasifikasi risiko diabetes menggunakan algoritma *K-Nearest Neighbor* (KNN) berdasarkan dataset *Early Stage Diabetes Risk Prediction*. Dataset tersebut terdiri dari 520 data dengan 17 atribut, termasuk fitur numerik seperti usia dan fitur kategorikal yang merepresentasikan gejala awal diabetes, seperti *polyuria*, *polydipsia*, *weakness*, dan *polyphagia*, dengan label kelas positif dan negatif diabetes. Tahap pra-pemrosesan data dilakukan dengan mentransformasikan atribut kategorikal menjadi nilai numerik (Yes=1, No=0; *male*=1, *female*=0) agar mendukung perhitungan jarak pada algoritma KNN. Model klasifikasi diimplementasikan menggunakan bahasa pemrograman Python pada Google Colab dan dievaluasi melalui empat skenario *percentage split*, yaitu 60:40, 70:30, 80:20, dan 90:10. Nilai K diuji pada rentang K=2 hingga K=9 menggunakan metrik jarak Euclidean untuk menentukan parameter optimal. Hasil eksperimen menunjukkan

Submitted : 16-12-2025 | Reviewed : 21-12-2025 | Accepted : 31-12-2025

bahwa K=3 secara konsisten memberikan performa terbaik dengan akurasi tertinggi sebesar 97% pada skenario pembagian data 60:40. Temuan ini menunjukkan bahwa model KNN yang diusulkan efektif untuk mendeteksi risiko diabetes pada tahap awal.

Kata kunci: Diabetes mellitus, Prediksi Risiko, K-Nearest Neighbor, Klasifikasi, Nilai K, *Percentage* S*plit*.

1. Pendahuluan

Diabetes mellitus merupakan masalah kesehatan global karena memicu gangguan metabolik jangka panjang yang berkontribusi pada komplikasi serius, seperti penyakit kardiovaskular, nefropati, neuropati, dan gangguan penglihatan. Kompleksitas diabetes tidak hanya terletak pada kondisi klinisnya, tetapi juga pada faktor risiko dan variasi gejala awal yang muncul pada tiap individu. Di praktik layanan, banyak kasus baru teridentifikasi ketika kondisi sudah berkembang dan menimbulkan komorbid. Kondisi ini menegaskan urgensi penguatan mekanisme skrining dan deteksi dini berbasis data. Seiring meningkatnya digitalisasi data kesehatan, pendekatan analitik berbasis *machine learning* menjadi relevan untuk membantu tenaga kesehatan melakukan klasifikasi risiko secara lebih cepat dan terstandar.

Dalam kajian *machine learning* semakin dominan digunakan untuk prediksi diabetes karena kemampuannya memanfaatkan pola multivariat dari data klinis dan demografis. Studi komunitas berbasis *big data mining* menunjukkan bahwa model prediksi dapat membantu skrining risiko secara masif di tingkat komunitas, sekaligus mendukung *early warning* untuk intervensi pencegahan yang lebih tepat sasaran[1]. Pada ranah yang lebih luas, penelitian komparatif dan studi *ensemble* juga berkembang pesat, baik pada data rekam medis elektronik maupun dataset publik, untuk meningkatkan akurasi serta robustitas model[2].

Namun, tantangan utama pada prediksi diabetes adalah heterogenitas dataset: perbedaan populasi, variasi fitur, kualitas data (missing, noise), ketidakseimbangan kelas, dan perbedaan standar pengukuran. Karena itu, banyak penelitian 2023–2025 menekankan pentingnya pra-pemrosesan, seleksi fitur, serta pemilihan algoritma dan skema evaluasi yang tepat agar model tidak sekadar mencapai akurasi tinggi, tetapi juga memiliki kinerja yang stabil dan dapat digeneralisasi[3].

Dalam konteks algoritma klasifikasi, K-Nearest Neighbor (KNN) merupakan metode yang cukup populer karena kesederhanaan, transparansi mekanisme prediksi, dan kemampuannya bekerja baik pada data nonlinier tertentu. KNN mengklasifikasikan data baru berdasarkan kedekatan jarak terhadap K tetangga terdekat dari data latih. Meski demikian, performa KNN sangat sensitif terhadap pemilihan nilai K, skala fitur, distribusi data, serta representasi numerik dari variabel kategorikal. Oleh sebab itu, beberapa studi menyoroti bahwa *tuning* nilai K dan normalisasi dapat mengubah performa secara signifikan dan menjadi faktor krusial dalam implementasi praktis [4].

Penelitian ini menempatkan diri pada arus riset prediksi diabetes 2023–2025 dengan fokus pada klasifikasi *early-stage diabetes risk* menggunakan dataset publik yang telah banyak digunakan dan relevan untuk skenario deteksi dini. Dataset yang dipakai adalah *Early Stage Diabetes Risk Prediction* yang berisi 520 data pasien dan 17 atribut, mencakup usia, gender, serta berbagai gejala awal (polyuria, polydipsia, sudden weight loss, weakness, dan lainnya) hingga label kelas positif/negatif diabetes. Pemilihan dataset ini menguatkan aspek replikasi dan pembandingan hasil karena dataset yang sama juga digunakan oleh berbagai studi terkini untuk menguji metode klasifikasi, termasuk KNN, Random Forest, AdaBoost, Bagging, dan pendekatan *ensemble* lainnya[5].

Tujuan utama penelitian ini adalah: (1) melakukan transformasi data agar sesuai untuk komputasi jarak pada KNN; (2) menguji pengaruh skema pembagian data latih-uji melalui *percentage split* (60:40, 70:30, 80:20, 90:10); (3) menemukan nilai K optimal pada rentang K=2 hingga K=9; serta (4) mengevaluasi performa model dengan metrik *accuracy*, *precision*, *recall*, dan *F1-score* menggunakan *confusion matrix*. Hasil yang diperoleh diharapkan dapat memberikan kontribusi praktis berupa model baseline yang kuat untuk prediksi diabetes berbasis gejala awal, sekaligus memberi bukti empiris bahwa konfigurasi parameter sederhana (nilai K dan skema split) dapat menghasilkan kinerja yang sangat kompetitif pada dataset ini[6].

Berdasarkan tujuan dan kerangka penelitian, hipotesis yang diajukan dalam penelitian ini adalah sebagai berikut:

H1: Perbedaan nilai parameter K pada algoritma KNN berpengaruh terhadap kinerja klasifikasi risiko diabetes tahap awal.

Submitted : 16-12-2025 | Reviewed : 21-12-2025 | Accepted : 31-12-2025

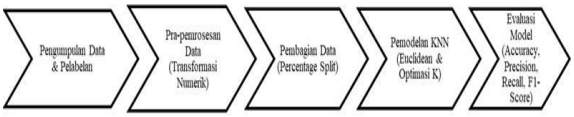
H2: Perbedaan skema pembagian data latih dan data uji (*percentage split*) menghasilkan variasi performa model KNN.

H3: Model KNN dengan parameter optimal mampu memberikan performa klasifikasi yang stabil pada kelas positif dan negatif diabetes.

Kontribusi ilmiah penelitian ini terletak pada penyajian eksperimen yang sistematis: transformasi fitur kategorikal yang konsisten, eksplorasi parameter K secara terukur, evaluasi empat skenario *percentage split*, dan interpretasi kinerja yang menyoroti kelas positif/negatif secara terpisah. Pada penelitian-penelitian 2023–2025, aspek interpretasi metrik per kelas sering disorot karena akurasi global dapat menutupi kelemahan model pada salah satu kelas—terutama ketika dataset tidak seimbang[7]. Dengan demikian, studi ini menekankan pembacaan presisi/recall per kelas agar hasil lebih bermakna untuk implementasi skrining[8].

2. Metode Penelitian

Tahapan penelitian ini dimulai dengan pengumpulan data dan penentuan label kelas berdasarkan kondisi diabetes positif dan negatif. Selanjutnya, dilakukan pra-pemrosesan data melalui transformasi atribut kategorikal menjadi numerik agar sesuai dengan kebutuhan algoritma K-Nearest Neighbor (KNN). Data yang telah diproses kemudian dibagi menjadi data latih dan data uji menggunakan metode *percentage split*. Pada tahap berikutnya, proses klasifikasi dilakukan menggunakan algoritma KNN dengan perhitungan jarak Euclidean dan pengujian beberapa nilai K untuk memperoleh parameter optimal. Model yang dihasilkan selanjutnya dievaluasi menggunakan *confusion matrix* dengan metrik akurasi, presisi, recall, dan F1-score. Seluruh rangkaian tahapan penelitian ini ditunjukkan pada Gambar 1.



Gambar 1. Tahap Penelitian

2.1 Pengumpulan Data & Pelabelan

Dataset penelitian berasal dari *UCI Machine Learning Repository* untuk prediksi risiko diabetes tahap awal (*Early Stage Diabetes Risk Prediction*). Dataset terdiri dari 520 record dan 17 atribut, yang mencakup usia (numerik), gender, serta 15 atribut gejala klinis (kategorikal) dan 1 atribut kelas target (positif/negatif) jurnal diabetes. Atribut-atribut gejala seperti polyuria, polydipsia, sudden weight loss, weakness, polyphagia, dan lainnya merupakan indikator yang lazim muncul pada fase awal gangguan regulasi glukosa. Dalam data Anda, distribusi kelas adalah 320 positif dan 200 negatif jurnal diabetes. Distribusi ini menunjukkan adanya kecenderungan ketidakseimbangan moderat yang berpotensi memengaruhi metrik tertentu misalnya presisi/recall pada kelas minoritas sehingga interpretasi per kelas menjadi penting.

2.2 Pra-pemrosesan Data (Transformasi Numerik)

Pra-pemrosesan data dilakukamentransformasikan atribut kategorikal ke dalam bentuk numerik agar dapat diproses oleh algoritma K-Nearest Neighbor (KNN). Seluruh atribut gejala dengan nilai *Yes/No* dikonversi menjadi 1/0, sedangkan atribut *Gender* dikonversi menjadi *male* = 1 dan *female* = 0. Atribut *Age* dipertahankan dalam bentuk numerik asli karena telah berupa data kontinu. Pada penelitian ini, normalisasi fitur numerik *Age* tidak diterapkan. Keputusan ini diambil untuk mempertahankan konfigurasi model KNN dalam bentuk baseline dan untuk mengevaluasi performa algoritma pada kondisi prapemrosesan minimal. Namun demikian, perlu dicatat bahwa KNN bersifat sensitif terhadap skala fitur, sehingga perbedaan rentang nilai antara atribut *Age* dan fitur biner berpotensi memengaruhi perhitungan jarak[9]. Transformasi biner dipilih karena sesuai dengan struktur data gejala (Yes/No) dan mempertahankan interpretabilitas[10].

2.3 Pembagian Data (Percentage Split)

Evaluasi model dilakukan dengan empat skenario *percentage split*: 60:40, 70:30, 80:20, dan 90:10 (latih : uji). Rinciannya: split 60:40 menghasilkan 312 data latih dan 208 data uji; split 70:30 menghasilkan 364 latih dan 156 uji; split 80:20 menghasilkan 416 latih dan 104 uji; split 90:10 menghasilkan 468 latih dan 52 uji jurnal diabetes. Pada penelitian ini, cross-validation tidak diterapkan. Meskipun *k-fold cross-validation* umumnya memberikan estimasi performa yang lebih stabil dan representatif, pendekatan tersebut tidak digunakan karena fokus penelitian diarahkan pada evaluasi pengaruh skema pembagian data terhadap performa model KNN dalam konfigurasi baseline. Dengan kata lain, penggunaan *percentage split* dipilih secara sadar sebagai bagian dari desain eksperimen, bukan sebagai keterbatasan yang tidak disengaja.[11].

2.4 Pemodelan KKN (Euclidean & Optimasi K)

KNN mengklasifikasikan sampel uji dengan melihat K sampel latih terdekat di ruang fitur. Ukuran kedekatan dihitung menggunakan Euclidean Distance[12]:

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

Pada data biner, Euclidean Distance tetap dapat digunakan untuk mengukur perbedaan pola gejala antar pasien. Meskipun beberapa literatur juga menggunakan *Hamming distance* untuk data biner,

Euclidean sering dipakai karena implementasinya umum pada pustaka *machine learning*.

2.5 Evaluasi Model (Accuracy, Precision, Recall, F1-Score)

Evaluasi model dilakukan menggunakan confusion matrix dan metrik, Accuracy: proporsi prediksi benar terhadap total data uji. Precision: ketepatan prediksi kelas tertentu (positif/negatif). Recall: kemampuan model menangkap seluruh kasus pada kelas tertentu. F1-score: rata-rata harmonik precision dan recall[13]. Penilaian per kelas penting karena dalam skrining kesehatan, kesalahan yang paling kritis sering kali adalah false negative (pasien sebenarnya positif namun diprediksi negatif). Oleh karena itu, interpretasi recall untuk kelas positif menjadi salah satu indikator kinerja yang esensial[14].

3. Hasil dan Pembahasan

Hasil penelitiaan menggunakan data penyakit diabetes diambil dari dataset UCI machine learning Repository yang diperoleh dari platform archive Resiko prediksi diabetes (https://archive.ics.uci.edu/dataset/529/early+stage+diabetes+risk+prediction+dataset) Dataset ini terdiri dari 520 record data dan 17 atribut. Berikut adalah deskripsi dari masing-masing atribut yang dapat dilihat tabel 1.

Tabel 1. Deskripsi Atribut

No	Atrribut	Deskripsi	Singkatan
1	Age	Umur	Age
2	Gender	Jenis Kelamin	Gd
3	Polyuria	Sering Buang Air Kecil	Pl
4	Polydipsia	Sering Haus	Pd
5	Sudden Weight Loss	Penurunan Berat Badan	Swl
6	Weakness	Rasa Lelah	W
7	Polyphagia	Sering Lapar	Pg
8	Genital Thrush	Infeksi Jamur Genital	Gt
9	Visual Blurring	Penglihatan Kabur	Vb
10	Itching	Gatal	it
11	Irritability	Mudah Tersinggung	ir
12	Delayed Healing	Penyembuhan Luka Yang Lambat	Dh
13	Partial Paresis	Kelemahan Otot	Pp
14	Muscle Stiffness	Kekakuan Otot	Ms
15	Alopecia	Kerontokan Rambut	Al
16	Obesity	Kelebihan Lemak Tubuh	Ob
17	Class	Kelas Diabetes termasuk Positif atau Negatif	Class

Data penyakit riwayat diabetes terdapat 520 data dapat dilihat pada tabel 2.

Tabel 2. Data Penyakit Database

No	Age	Gd	Pl	Pd	Swl	W	Pg	Gt	Vb	It	Ir	Dh	Pp	Ms	Al	Ob	class
1	40	Male	No	Yes	No	Yes	No	No	No	Yes	No	Yes	No	Yes	Yes	Yes	Positive
2	58	Male	No	No	No	Yes	No	No	Yes	No	No	No	Yes	No	Yes	No	Positive
3	41	Male	Yes	No	No	Yes	Yes	No	No	Yes	No	Yes	No	Yes	Yes	No	Positive
4	45	Male	No	No	Yes	Yes	Yes	Yes	No	Yes	No	Yes	No	No	No	No	Positive
5	60	Male	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Positive
...																	
516	39	Female	Yes	Yes	Yes	No	Yes	No	No	Yes	No	Yes	Yes	No	No	No	Positive
517	48	Female	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	No	No	No	Positive
518	58	Female	Yes	Yes	Yes	Yes	Yes	No	Yes	No	No	No	Yes	Yes	No	Yes	Positive
519	32	Male	No	No	No	Yes	No	No	No	Yes	No	Yes	No	No	Yes	No	Negative
520	42	Male	No	No	No	No	No	No	No	No	No	No	No	No	No	No	Negative

Data pada tabel 2 perlu ditansformasi data yang dimana proses mengubah data dari satu format atau keadaan ke format atau keadaan yang berbeda agar lebih berguna untuk analisis, pemodelan, atau pengambilan keputusan. Sehingga data diatas diubah menjadi numerik selain atribut age yang sudah tipe data numerik dan class karena merupakan target menentukan positif diabetes dan negatif diabetes yang dapat dilihat pada tabel 3.

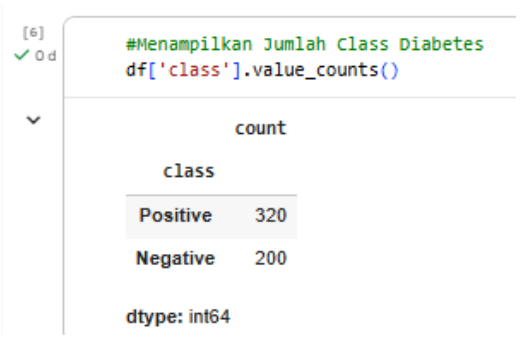
Tabel 3. Transformasi Data

No	Age	Gd	Pl	Pd	Swl	W	Pg	Gt	Vb	It	Ir	Dh	Pp	Ms	Al	Ob	class
1	40	1	0	1	0	1	0	0	0	1	0	1	0	1	1	1	Positive
2	58	1	0	0	0	1	0	0	1	0	0	0	1	0	1	0	Positive
3	41	1	1	0	0	1	1	0	0	1	0	1	0	1	1	0	Positive
4	45	1	0	0	1	1	1	1	0	1	0	1	0	0	0	0	Positive
5	60	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	Positive
...																	

516	39	0	1	1	1	0	1	0	0	1	0	1	1	0	0	0	Positive
517	48	0	1	1	1	1	1	0	1	1	1	1	1	0	0	0	Positive
518	58	0	1	1	1	1	1	0	1	0	0	0	1	1	0	1	Positive
519	32	1	0	0	0	1	0	0	0	1	0	1	0	0	1	0	Negative
520	42	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Negative

Tabel 3. Transformasi data dimana data diubah menjadi numerik yang data Yes menjadi angka 1 sedangkan data No menjadi angka 0. Kemudian untuk gender yang data male menjadi angka 1 sedangkan data female menjadi angka 0.

Data diproses menggunakan bahasa pemograman python di google colab yang dimana dari 520 data terdapat 320 data positif diabetes dan 200 data negatif diabetes yang dapat dilihat pada gambar 2.



Gambar 2. Jumlah Class Diabetes

Setelah melewati tahap dan *transformation data*, proses berikutnya adalah tahap data mining. Pada tahap ini dilakukan proses klasifikasi penyakit diabetes menggunakan algoritma K-Nearest Neighbor (KNN), di mana data dibagi menjadi data latih (*training*) dan data uji (*testing*). Pembagian data menerapkan metode *percentage split* dengan empat skenario proporsi, yaitu 90:10, 80:20, 70:30, dan 60:40. Pembagian masing-masing skenario ditampilkan pada Tabel 4.

Tabel 4. Pembagian Data

Split	Data Latih	Data Uji
60:40	312	208
70:30	364	156
80:20	416	104
90:10	468	52

Langkah ini diperlukan agar pelatihan dan pengujian model machine learning dilakukan pada data terpisah, sehingga hasil evaluasi model lebih objektif dan bekerja lebih baik terhadap baru.

Kemudian melakukan pencarian nilai K terbaik di mulai dari k = 2 sampai K = 9 dengan menggunakan rumus Eucludian Distance untuk melakukan pencarian nilai k tetangga terdekat, yang mana bertujuan agar mendapatkan nilai K tetangga terdekat dengan nilai optimal yang dapat dilihat pada tabel 5.

Tabel 5. pencarian nilai K terbaik

K	Akurasi			
	60:40	70:30	80:20	90:10
K2	0.95	0.94	0.94	0.94
K3	0.97	0.96	0.95	0.96
K4	0.92	0.92	0.93	0.90
K5	0.93	0.93	0.94	0.92
K6	0.91	0.90	0.93	0.92
K7	0.92	0.92	0.93	0.94
K8	0.90	0.90	0.92	0.92
K9	0.92	0.90	0.93	0.94

Pada tabel 5. Nilai K terbaik pada empat skenario proporsi, yaitu 90:10, 80:20, 70:30, dan 60:40 teletak pada nilai K3 yang paling optimal.

Setelah data diolah menggunakan Algoritma KNN, tahap selanjutnya adalah melakukan pengujian algoritma menggunakan confusion matrix dengan nilai K=3 menggunakan bahasa pemrograman python digoogle colab yang hasilnya dapat dilihat pada tabel 6.

Tabel 6. Hasil Klasifikasi

Split	K	Accuracy	Precision	Recall	F1-Score
60:40	3	97%	Negatif : 92% Positif : 100%	Negatif : 100% Positif : 94%	Negatif : 96% Positif : 97%
70:30	3	96%	Negatif : 90% Positif : 100%	Negatif : 100% Positif : 92%	Negatif : 95% Positif : 96%
80:20	3	95%	Negatif : 90% Positif : 100%	Negatif : 100% Positif : 92%	Negatif : 95% Positif : 96%
90:10	3	96%	Negatif : 90% Positif : 100%	Negatif : 100% Positif : 94%	Negatif : 95% Positif : 97%

Berdasarkan tabel di atas, penerapan metode K-NN dengan nilai k = 3 terdapat nilai klasifikasi tinggi pada skenario proporsi 60:40 dimana data latih ada 312 dan data uji 208 menghasilkan nilai akurasi sebesar 97%. Hasil laporan klasifikasi menggunakan bahasa pemrograman python digoogle colab pada rasio 60:40 dapat dilihat pada gambar 3.

Classification Report:

	precision	recall	f1-score	support
Negative	0.92	1.00	0.96	84
Positive	1.00	0.94	0.97	124
accuracy			0.97	208
macro avg	0.96	0.97	0.97	208
weighted avg	0.97	0.97	0.97	208

Gambar 3. Laporan Klasifikasi

Pada gambar 3 menghasilkan nilai akurasi 97%. Selain itu model ini menghasilkan presisi tertinggi sebesar 100% untuk pasien positif terkena diabetes dan 92% untuk pasien negatif terkena diabetes. Recall untuk pasien positif adalah 94%, sedangkan untuk pasien negatif adalah 100%. F1-Score untuk pasien positif adalah 97%, sedangkan untuk pasien negatif adalah 96%. Hasil ini memberikan gambaran tentang seberapa efektif model K-NN dalam mengidentifikasi pasien dengan diabetes. Hasil eksperimen menunjukkan bahwa algoritma K-Nearest Neighbor (KNN) dengan nilai parameter K = 3 secara konsisten memberikan performa terbaik pada seluruh skenario pembagian data. Akurasi tertinggi, yaitu sebesar 97%, diperoleh pada skenario *percentage split* 60:40.

Temuan ini mengindikasikan bahwa konfigurasi parameter K memiliki pengaruh signifikan terhadap kinerja model, terutama pada data dengan kombinasi fitur numerik dan biner. Jika dibandingkan secara konseptual dengan penelitian terdahulu, hasil yang diperoleh dalam penelitian ini berada pada kisaran yang kompetitif. Beberapa studi menggunakan algoritma Naïve Bayes, Random Forest, dan Support Vector Machine pada dataset serupa melaporkan akurasi yang relatif tinggi, terutama ketika dikombinasikan dengan teknik seleksi fitur, penanganan *imbalance*, atau pendekatan ensemble. Namun, metode-metode tersebut umumnya melibatkan kompleksitas komputasi dan konfigurasi yang lebih tinggi dibandingkan KNN. Dalam konteks ini, penelitian ini menempatkan KNN sebagai model baseline yang sederhana namun efektif, yang dapat menjadi titik awal sebelum penerapan algoritma yang lebih kompleks.

4. Kesimpulan

Penelitian ini menggunakan dataset diabetes dengan 17 atribut klinis dan demografis yang bertujuan untuk mengembangkan model prediksi diabetes menggunakan algoritma K-Nearest Neighbor (KNN). Hasil eksperimen menunjukkan bahwa nilai parameter terbaik diperoleh pada K = 3, dengan performa tertinggi pada skenario *percentage split* 60:40 (312 data latih dan 208 data uji) yang menghasilkan akurasi sebesar 97%. Pada skenario ini, model mencapai presisi sebesar 100% untuk kelas positif dan 92% untuk kelas negatif, dengan nilai *recall* masing-masing sebesar 94% dan 100%, serta F1-score sebesar 97% untuk kelas positif dan 96% untuk kelas negatif. Hasil ini menunjukkan bahwa pemilihan nilai K berpengaruh signifikan terhadap kinerja model dan bahwa KNN cukup efektif sebagai model baseline dalam klasifikasi risiko diabetes tahap awal. Performa model sangat dipengaruhi oleh kualitas dan karakteristik data, sehingga diperlukan pengembangan dan evaluasi lanjutan untuk meningkatkan keandalan model.

Daftar Rujukan

[1] F. Ruziq and M. R. Wayahdi, "Web-Based Diabetes Risk Prediction System Using K-NN on Kaggle Early Stage Diabetes Dataset," vol. 6, no. 5, pp. 3217–3229, 2025.

[2] O. M. Haq, A. Ridwan, and T. G. Pratama, "Analisis Perbandingan Kinerja Algoritma Naïve Bayes Dan KNN Untuk Memprediksi Penyakit Diabetes," pp. 193–201.

[3] N. P. Korina, B. Prasetiyo, A. A. Hakim, and M. R. A. Septian, "Journal of Information System The Influence of Determining the K-Value on Improving the Diabetes Classification Model using the K-NN Algorithm," vol. 2, no. 2, pp. 69–76, 2024.

[4] F. Rahman, S. Hossain, and J. Tiang, "Diabetes Prediction Using Feature Selection Algorithms and Boosting-Based Machine Learning Classifiers," pp. 1–24, 2025.

[5] S. Afolabi, N. Ajadi, A. Jimoh, and I. Adenekan, "Informatics and Health Predicting diabetes using supervised machine learning algorithms on E-health records," *Informatics Heal.*, vol. 2, no. 1, pp. 9–16, 2025, doi: 10.1016/j.infoh.2024.12.002.

[6] S. Yadu, R. Chandra, and V. K. Sinha, "Comparing Different Machine Learning Techniques in Predicting Diabetes on Early Stage †," pp. 1–9, 2024.

[7] L. Jiang *et al.*, "Diabetes risk prediction model based on community follow-up data using machine learning," *Prev. Med. Reports*, vol. 35, no. August, p. 102358, 2023, doi: 10.1016/j.pmedr.2023.102358.

[8] R. Pratama *et al.*, "IMPLEMENTATION OF DIABETES PREDICTION MODEL USING RANDOM IMPLEMENTASI MODEL PREDIKSI DIABETES MENGGUNAKAN ALGORITMA," vol. 5, no. 4, pp. 1165–1174, 2024.

[9] A. Oktaviana, D. P. Wijaya, A. Pramuntadi, and D. Heksaputra, "Prediction of Type 2 Diabetes Mellitus Using The K-Nearest Neighbor (K-NN) Algorithm Prediksi Penyakit Diabetes Melitus Tipe 2 Menggunakan Algoritma K-Nearest Neighbor (K-NN)," vol. 4, no. July, pp. 812–818, 2024.

[10] N. Agustus, R. N. Fitria, W. Sugianto, and A. Cemara, "Prediksi Penyakit Diabetes Mellitus Tipe I dan Tipe II Menggunakan Metode KNN di Klinik Dharma Husada Universitas PGRI Yogyakarta , Indonesia Diabetes Mellitus Diabetes Mellitus adalah gangguan metabolic yang ditandai peningkatan kadar glukosa darah (Hiperglikemia) akibat kerusakan pada sekresi insulin dan kerja insulin , kadar glukosa," vol. 2, no. 3, 2024.

[11] L. Nuril, Z. Fatah, and I. Yunita, "KLASIFIKASI PENYAKIT DIABETES MENGGUNAKAN METODE K-NEAREST NEIGHBORS," vol. 1, no. 4, pp. 41–46, 2024.

[12] V. No, Z. Amri, M. Rodi, M. N. Wathani, A. Bagja, and V. No, "Infotek : Jurnal Informatika dan Teknologi Prediksi Diabetes Menggunakan Algoritma K-Nearest (KNN) Teknik SMOTE-ENN Infotek :

[13] P. N. Sabrina and A. Komarudin, "PREDIKSI PENYAKIT DIABETES DENGAN METODE K-NEAREST NEIGHBOR (KNN) DAN SELEKSI FITUR INFORMATION GAIN," vol. 8, no. 6, pp. 11320–11326, 2024.

[14] R. Ajeng, M. Farhan, and A. Cahyani, "Prediksi Terkena Diabetes menggunakan Metode K- Nearest Neighbor (KNN) pada Dataset UCI Machine Learning Diabetes Abstract : Abstrak :," vol. 3, no. 2, pp. 15–19, 2023, doi: 10.35472/indojam.v3i2.1577.

Submitted : 16-12-2025 | Reviewed : 21-12-2025 | Accepted : 31-12-2025