Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 357−368
DOI : 10.47065/bulletinds.v5i3.10391
https://ejurnal.seminar-id.com/index.php/bulletinds

Hybrid Weighted K-Nearest Neighbor dan Naive Bayes untuk Klasifikasi Diabetes Melitus

Aklima Laduna Ramadya[1,*], Solikhun[2], Timbo Faritcan P. Siallagan[3]

[1] Information System, STIKOM Tunas Bangsa, Pematangsiantar, Indonesia
[2] Informatics Engineering, STIKOM Tunas Bangsa, Pematangsiantar, Indonesia
[3] Faculty of Engineering, Computer and Network Engineering, Universitas Mandiri, Jawa Barat, Indonesia
Email: [1]1adunaaklima67gmail.com, [2]solikhun@amiktunasbangsa.ac.id, [3]timbofaritcansiallagan@gmail.com
Email Penulis Korespondensi: ladunaaklima67gmail.com

Abstrak−Diabetes melitus merupakan penyakit metabolik kronis dengan prevalensi yang terus meningkat dan berpotensi menimbulkan berbagai komplikasi serius apabila tidak dideteksi sejak dini. Pemanfaatan machine learning menjadi salah satu pendekatan yang efektif untuk mendukung proses klasifikasi pasien berdasarkan data klinis secara lebih cepat dan akurat berdasarkan data pasien. Penelitian ini mengusulkan model Hybrid Weighted K-Nearest Neighbor (WKNN)–Naive Bayes yang mengombinasikan probabilitas keluaran Weighted K-Nearest Neighbor dan Gaussian Naive Bayes melalui kombinasi berbobot menggunakan parameter α untuk menghasilkan keputusan klasifikasi akhir. Pendekatan ini memanfaatkan keunggulan metode berbasis kedekatan jarak dan probabilistik dalam proses klasifikasi diabetes melitus. Dataset yang digunakan adalah Pima Indians Diabetes Dataset yang terdiri atas 768 data pasien dengan delapan atribut prediktor dan satu variabel target. Tahap pra-pemrosesan meliputi penanganan *missing value* menggunakan imputasi, normalisasi data menggunakan RobustScaler, *feature engineering*, serta penyeimbangan kelas menggunakan Synthetic Minority Over-sampling Technique (SMOTE). Evaluasi model dilakukan menggunakan Stratified 10-Fold Cross Validation dan dibandingkan dengan model K-Nearest Neighbor (KNN), Weighted K-Nearest Neighbor (WKNN), dan Naive Bayes. Hasil penelitian menunjukkan bahwa model Hybrid WKNN–Naive Bayes memperoleh Accuracy sebesar 88,50%, Precision sebesar 85,32%, Recall sebesar 93,00%, F1-Score sebesar 88,99%, dan AUC-ROC sebesar 91,09%, sehingga mampu memberikan performa klasifikasi yang kompetitif. Meskipun nilai Accuracy yang diperoleh setara dengan WKNN, pendekatan hybrid mampu mengintegrasikan informasi berbasis kedekatan jarak dan probabilistik sehingga tetap menghasilkan model klasifikasi yang stabil. Model Weighted K-Nearest Neighbor (WKNN) juga memperoleh Accuracy sebesar 88,50%, dengan Recall sebesar 94,00%, F1-Score sebesar 89,10%, dan AUC-ROC sebesar 91,73%, sedangkan model Naive Bayes menghasilkan Accuracy sebesar 71,00%. Hasil tersebut menunjukkan bahwa pendekatan Hybrid WKNN–Naive Bayes mampu menghasilkan performa klasifikasi yang kompetitif melalui kombinasi metode berbasis kedekatan jarak dan probabilistik, sehingga berpotensi digunakan sebagai sistem pendukung keputusan untuk klasifikasi risiko diabetes melitus serta mendukung deteksi dini berdasarkan data pasien.

Keywords: *Diabetes Melitus; Klasifikasi Hybrid; Weighted K-Nearest Neighbor; Naive Bayes; Machine Learning.*

Abstract−Diabetes mellitus is a chronic metabolic disease with a continuously increasing prevalence and the potential to cause various serious complications if not detected early. The use of machine learning has become one of the effective approaches to support faster and more accurate classification of patients based on clinical data. This study proposes a Hybrid Weighted K-Nearest Neighbor (WKNN)–Naive Bayes model that combines the output probabilities of Weighted K-Nearest Neighbor and Gaussian Naive Bayes through a weighted combination using parameter α to produce the final classification decision. This approach leverages the strengths of both distance-based and probabilistic methods in the classification of diabetes mellitus. The dataset used is the Pima Indians Diabetes Dataset, consisting of 768 patient records with eight predictor attributes and one target variable. The preprocessing stage includes handling missing values through imputation, data normalization using RobustScaler, feature engineering, and class balancing using the Synthetic Minority Over-sampling Technique (SMOTE). Model evaluation was conducted using Stratified 10-Fold Cross Validation and compared against K-Nearest Neighbor (KNN), Weighted K-Nearest Neighbor (WKNN), and Naive Bayes models. The results show that the Hybrid WKNN–Naive Bayes model achieved an Accuracy of 88.50%, Precision of 85.32%, Recall of 93.00%, F1-Score of 88.99%, and AUC-ROC of 91.09%, demonstrating competitive classification performance. Although the Accuracy obtained is equal to that of WKNN, the hybrid approach is able to integrate distance-based and probabilistic information, thereby still producing a stable classification model. The Weighted K-Nearest Neighbor (WKNN) model also achieved an Accuracy of 88.50%, with a Recall of 94.00%, F1-Score of 89.10%, and AUC-ROC of 91.73%, while the Naive Bayes model achieved an Accuracy of 71.00%. These results indicate that the Hybrid WKNN–Naive Bayes approach is capable of producing competitive classification performance through the combination of distance-based and probabilistic methods, making it potentially useful as a decision support system for diabetes mellitus risk classification and for supporting early detection based on patient data.

Keywords: Diabetes Mellitus; Hybrid Classification; Weighted K-Nearest Neighbor; Naive Bayes; Machine Learning.

1. PENDAHULUAN

Diabetes melitus (DM) merupakan salah satu penyakit kronis tidak menular yang menjadi permasalahan kesehatan global paling serius pada abad ke-21. Menurut laporan International Diabetes Federation (IDF), pada tahun 2021 tercatat sebanyak 537 juta orang dewasa di seluruh dunia hidup dengan diabetes, dan angka tersebut diproyeksikan akan meningkat menjadi 643 juta pada tahun 2030 serta 783 juta pada tahun 2045[1], [2]. Di Indonesia, situasi ini tidak kalah mengkhawatirkan. Berdasarkan data IDF Diabetes Atlas, Indonesia menempati posisi kelima dunia dengan jumlah penderita diabetes terbesar, yakni sekitar 19,47 juta jiwa pada tahun 2021. Tingginya prevalensi ini berdampak besar tidak hanya pada kualitas hidup penderita, tetapi juga pada beban sistem layanan kesehatan nasional[3].

Diabetes melitus ditandai dengan kondisi hiperglikemia kronis yang terjadi akibat defisiensi sekresi insulin, gangguan kerja insulin, atau keduanya. Secara klinis, penyakit ini dibagi menjadi Diabetes Melitus Tipe 1, Tipe 2, dan diabetes gestasional[4]-[6]. Diabetes Tipe 2 merupakan jenis yang paling dominan, mencakup lebih dari 90% kasus secara

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 357–368
DOI : 10.47065/bulletinds.v5i3.10391
https://ejurnal.seminar-id.com/index.php/bulletinds

global, dan erat kaitannya dengan gaya hidup tidak sehat, obesitas, serta faktor genetik. Jika tidak terdeteksi dan ditangani sejak dini, diabetes dapat memicu komplikasi serius seperti neuropati, nefropati, retinopati, penyakit jantung koroner, hingga amputasi anggota tubuh. Oleh karena itu, deteksi dini dan diagnosis yang akurat menjadi faktor kritis dalam manajemen penyakit ini[7].

Perkembangan pesat di bidang teknologi informasi, khususnya machine learning dan data mining, membuka peluang baru dalam pengembangan sistem diagnosis berbasis data medis. Berbagai algoritma klasifikasi telah diterapkan untuk membantu klasifikasi diabetes, di antaranya K-Nearest Neighbor (KNN), Naive Bayes, Decision Tree, Support Vector Machine (SVM), dan Random Forest[8]-[15]. Setiap algoritma memiliki kelebihan dan keterbatasan tersendiri. Algoritma KNN dikenal sederhana dan efektif dalam menangani data non-linear, namun sensitif terhadap pemilihan nilai K serta skala fitur [16]. Sementara itu, Naive Bayes unggul dalam kecepatan komputasi dan kinerjanya pada dataset berukuran kecil hingga menengah, namun asumsi independensi antar fitur yang menjadi fondasinya seringkali tidak terpenuhi dalam data medis nyata[17].

Beberapa penelitian sebelumnya telah menerapkan berbagai algoritma *machine learning* untuk klasifikasi diabetes melitus. Penelitian oleh Bangun dan Rachmat [1] membandingkan algoritma K-Nearest Neighbor (KNN) dan Naive Bayes, yang menunjukkan bahwa KNN menghasilkan performa klasifikasi yang lebih baik dibandingkan Naive Bayes. Selanjutnya, Mulyani dkk. [2] membandingkan algoritma KNN dan Support Vector Machine (SVM) dengan menerapkan teknik Synthetic Minority Over-sampling Technique (SMOTE) untuk mengatasi ketidakseimbangan kelas, sehingga mampu meningkatkan performa klasifikasi pada dataset diabetes. Penelitian lain oleh Ibrahim [7] menunjukkan bahwa setiap algoritma *machine learning* memiliki karakteristik performa yang berbeda bergantung pada karakteristik dataset yang digunakan. Selain itu, Nasien dkk. [19] membandingkan implementasi K-Nearest Neighbor, Naive Bayes, dan Logistic Regression pada klasifikasi diabetes, sedangkan Sholekhah dkk. [16] menunjukkan bahwa Weighted K-Nearest Neighbor (WKNN) mampu meningkatkan performa dibandingkan KNN konvensional melalui mekanisme pembobotan berdasarkan jarak tetangga terdekat. Hasil penelitian-penelitian tersebut menunjukkan bahwa setiap algoritma memiliki keunggulan dan keterbatasan masing-masing dalam klasifikasi diabetes melitus, sehingga masih diperlukan pengembangan metode yang mampu memanfaatkan keunggulan lebih dari satu algoritma untuk meningkatkan performa klasifikasi.

Berdasarkan penelitian-penelitian terdahulu, berbagai algoritma *machine learning* seperti K-Nearest Neighbor (KNN), Weighted K-Nearest Neighbor (WKNN), dan Naive Bayes telah menunjukkan performa yang baik dalam klasifikasi diabetes melitus. Namun, sebagian besar penelitian masih menggunakan algoritma tersebut secara terpisah atau hanya membandingkan performanya, sehingga belum memanfaatkan keunggulan masing-masing algoritma dalam satu model klasifikasi. Selain itu, penelitian yang mengombinasikan Weighted K-Nearest Neighbor (WKNN) dan Gaussian Naive Bayes untuk klasifikasi diabetes melitus masih belum banyak dilaporkan. Di sisi lain, evaluasi pada beberapa penelitian sebelumnya lebih berfokus pada peningkatan nilai akurasi, sedangkan analisis performa model berdasarkan metrik Accuracy, Precision, Recall, F1-Score, dan AUC-ROC secara bersamaan masih belum banyak dibahas. Oleh karena itu, masih terdapat peluang untuk mengembangkan model Hybrid Weighted K-Nearest Neighbor dan Gaussian Naive Bayes yang mampu memanfaatkan keunggulan metode berbasis kedekatan jarak dan metode probabilistik guna meningkatkan performa klasifikasi diabetes melitus.

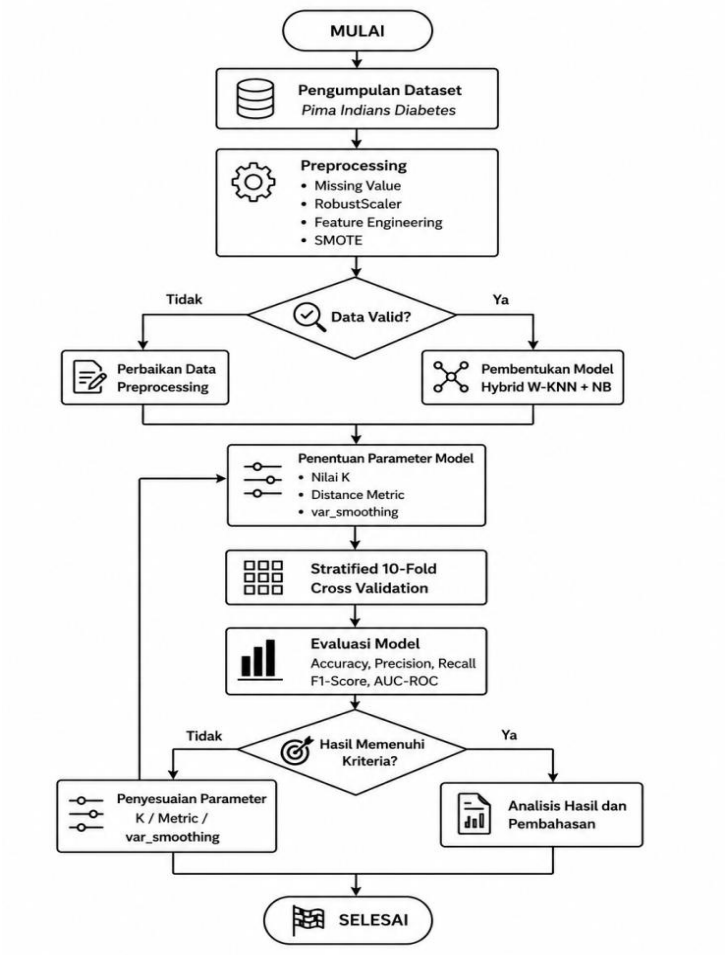
Dalam upaya mengatasi kelemahan masing-masing algoritma tunggal, pendekatan hybrid yang mengombinasikan dua atau lebih metode klasifikasi telah banyak mendapat perhatian dari para peneliti. Hybrid model KNN-Naive Bayes memanfaatkan kekuatan komplementer dari kedua algoritma, di mana KNN berperan dalam mengidentifikasi tetangga terdekat berdasarkan kemiripan data, sedangkan Naive Bayes memberikan probabilitas klasifikasi berbasis teorema Bayes[18]. Pengembangan lebih lanjut berupa Weighted KNN memungkinkan pemberian bobot berbeda pada setiap tetangga berdasarkan jaraknya terhadap data uji, sehingga tetangga yang lebih dekat memiliki pengaruh lebih besar dalam penentuan kelas akhir[19], [20]. Mekanisme pembobotan ini secara teoritis dapat meningkatkan akurasi klasifikasi secara signifikan.

Berdasarkan penelitian-penelitian terdahulu, berbagai algoritma *machine learning* seperti K-Nearest Neighbor (KNN), Weighted K-Nearest Neighbor (WKNN), Naive Bayes, dan algoritma lainnya telah menunjukkan performa yang cukup baik dalam klasifikasi diabetes melitus. Namun, sebagian besar penelitian masih berfokus pada penggunaan algoritma secara individual atau hanya melakukan perbandingan performa antaralgoritma tanpa mengombinasikan keunggulan masing-masing metode. Di sisi lain, penelitian yang mengembangkan model Hybrid Weighted K-Nearest Neighbor (WKNN) dan Gaussian Naive Bayes untuk klasifikasi diabetes melitus masih relatif terbatas. Padahal, WKNN memiliki keunggulan dalam memberikan bobot berdasarkan kedekatan jarak antar data sehingga mampu meningkatkan ketepatan klasifikasi, sedangkan Gaussian Naive Bayes memiliki kemampuan melakukan klasifikasi secara probabilistik dengan proses komputasi yang sederhana dan efisien. Oleh karena itu, masih terdapat peluang untuk mengembangkan model hybrid yang mampu memanfaatkan keunggulan kedua algoritma tersebut agar menghasilkan performa klasifikasi yang lebih optimal. Selain itu, sebagian penelitian sebelumnya lebih menitikberatkan pada peningkatan nilai akurasi, sedangkan evaluasi terhadap keseimbangan performa model berdasarkan metrik Accuracy, Precision, Recall, F1-Score, dan AUC-ROC secara bersamaan masih belum banyak dilakukan. Berdasarkan kondisi tersebut, penelitian ini mengusulkan model Hybrid Weighted K-Nearest Neighbor dan Gaussian Naive Bayes sebagai alternatif untuk meningkatkan performa klasifikasi diabetes melitus.

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 357–368
DOI : 10.47065/bulletinds.v5i3.10391
https://ejurnal.seminar-id.com/index.php/bulletinds

2. METODOLOGI PENELITIAN

Penelitian ini dirancang sebagai penelitian eksperimental berbasis komputasi menggunakan pendekatan machine learning. Tahapan penelitian meliputi: (1) preprocessing dataset; (2) formulasi model Hybrid Weighted KNN-NB; (3) pembentukan model Hybrid WKNN–Naive Bayes (4) pelatihan dan evaluasi menggunakan Stratified 10-Fold Cross Validation; serta (5) perbandingan performa terhadap model baseline. Seluruh proses diimplementasikan menggunakan Python 3.12 dengan pustaka scikit-learn, NumPy, dan Pandas.



Gambar 1. Alur Penelitian

Gambar 1 menunjukkan alur penelitian yang digunakan dalam membangun model Hybrid Weighted K-Nearest Neighbor (WKNN)–Gaussian Naive Bayes untuk klasifikasi diabetes melitus. Penelitian diawali dengan pengumpulan Pima Indians Diabetes Dataset. Selanjutnya dilakukan tahap preprocessing yang meliputi penanganan *missing value* menggunakan imputasi, normalisasi data menggunakan RobustScaler, *feature engineering*, dan penyeimbangan kelas menggunakan SMOTE. Data hasil preprocessing kemudian digunakan untuk membangun model Hybrid WKNN–Gaussian Naive Bayes melalui penentuan parameter α, nilai K, metrik Manhattan, dan parameter *var_smoothing*. Selanjutnya model dilatih dan diuji menggunakan metode Stratified 10-Fold Cross Validation. Tahap akhir penelitian adalah evaluasi performa model menggunakan metrik Accuracy, Precision, Recall, F1-Score, dan AUC-ROC untuk menilai kemampuan model dalam mengklasifikasikan diabetes melitus

2.1 Dataset dan Preprocessing

Dataset yang digunakan dalam penelitian ini adalah Pima Indians Diabetes Dataset (PIDD) yang terdiri atas 768 data pasien dengan delapan atribut prediktor dan satu atribut target berupa Outcome[27]. Tahap penelitian diawali dengan proses pengumpulan dataset, kemudian dilanjutkan dengan tahap preprocessing yang meliputi penanganan missing value menggunakan teknik imputasi, normalisasi data menggunakan RobustScaler, feature engineering, serta penyeimbangan kelas menggunakan Synthetic Minority Over-sampling Technique (SMOTE). Dataset yang telah diproses selanjutnya digunakan untuk membangun model Hybrid Weighted K-Nearest Neighbor (WKNN)–Naive Bayes. Konfigurasi model ditentukan melalui proses hyperparameter tuning yang meliputi parameter α, nilai K, metrik jarak, dan parameter var_smoothing. Selanjutnya model dievaluasi menggunakan Stratified 10-Fold Cross Validation dengan metrik Accuracy, Precision, Recall, F1-Score, dan AUC-ROC untuk mengetahui performa klasifikasi.

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 357–368
DOI : 10.47065/bulletinds.v5i3.10391
https://ejurnal.seminar-id.com/index.php/bulletinds

Proses preprocessing dilakukan melalui penanganan missing value menggunakan teknik imputasi sehingga seluruh nilai yang hilang berhasil ditangani, normalisasi menggunakan RobustScaler, feature engineering, serta penyeimbangan kelas menggunakan SMOTE[28], [29]. Dataset akhir terdiri atas 75 sampel kelas tidak diabetes dan 44 sampel kelas diabetes. Seluruh fitur telah dinormalisasi sehingga memiliki skala yang seragam sebelum digunakan dalam proses klasifikasi. Hasil preprocessing ini memastikan kualitas data yang lebih baik dan mengurangi potensi bias akibat nilai ekstrem maupun data yang tidak lengkap. Deskripsi fitur disajikan pada Tabel 1.

Tabel 1. Deskripsi Fitur Dataset Pima Indians Diabetes

Fitur	Deskripsi	Min	Maks
Pregnancies	Jumlah kehamilan	0	17
Glucose	Kadar glukosa plasma (mg/dL)	44	199
BloodPressure	Tekanan darah diastolik (mmHg)	24	122
SkinThickness	Ketebalan lipatan kulit (mm)	7	99
Insulin	Kadar insulin serum (µU/mL)	14	846
BMI	Indeks massa tubuh (kg/m²)	18.2	67.1
DiabetesPedigree Function	Fungsi silsilah diabetes	0.08	2.42
Age	Usia pasien (tahun)	21	81
Outcome	Label kelas (0=negatif, 1=positif)	0	1

Normalisasi dilakukan menggunakan RobustScaler, yaitu metode normalisasi yang memanfaatkan median dan interquartile range (IQR) sehingga lebih tahan terhadap keberadaan outlier. Pendekatan ini dipilih karena beberapa fitur pada dataset diabetes, seperti Insulin dan BMI, memiliki rentang nilai yang cukup besar sehingga RobustScaler mampu menghasilkan skala data yang lebih stabil sebelum proses klasifikasi, diformulasikan sebagai:

$$x' = \frac{(x - \mu)}{\sigma} \tag{1}$$

di mana x adalah nilai asli, μ adalah rata-rata, dan σ adalah standar deviasi fitur pada data latih. Dataset akhir terdiri dari 658 sampel (433 kelas 0, 225 kelas 1)

2.2 K-Nearest Neighbor Berbobot Jarak

KNN berbobot jarak mengklasifikasikan sampel uji berdasarkan k tetangga terdekat dengan bobot berbanding terbalik dengan kuadrat jaraknya[30]-[33]. Bobot tetangga ke-i dihitung sebagai:

$$w_i = \frac{1}{d(x, x_i)^2} \tag{2}$$

Jarak Euclidean antara sampel uji x dan tetangga ke-i adalah:

$$d(x, x_i) = \sqrt{[\,\Sigma_j\,(x_j - x_{ij})^2\,]} \tag{3}$$

Probabilitas kelas c pada sampel uji dihitung sebagai:

$$P_{knn}(c \mid x) = \frac{[\,\Sigma_i\, w_i \cdot \mathbb{1}(y_i = c)\,]}{\Sigma_i\, w_i} \tag{4}$$

2.3 Gaussian Naive Bayes

Gaussian Naive Bayes menghitung probabilitas posterior menggunakan teorema Bayes dengan asumsi distribusi Gaussian pada setiap fitur:

$$P(c \mid x) \propto P(c) \cdot \Pi_j\, P(x_j \mid c) \tag{5}$$

$$P(x_j \mid c) = [\frac{1}{(\sqrt{2\pi} \cdot \sigma\_cj)}] \cdot \frac{exp[-(x_j - \mu\_cj)^2}{(2\sigma^2\_cj)]} \tag{6}$$

2.4 Model Hybrid Weighted KNN-Naive Bayes

Model hybrid menggabungkan output probabilitas KNN dan Naive Bayes melalui pembobotan linier[12]. Prediksi probabilitas akhir untuk kelas c dinyatakan sebagai:

$$P_{hybrid}(c \mid x) = \frac{w_{knn} P_{KNN}(c \mid x) + w_{nb} \times P_{NB(c|x)}}{w_{knn} + w_{nb}} \tag{7}$$

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 357–368
DOI : 10.47065/bulletinds.v5i3.10391
https://ejurnal.seminar-id.com/index.php/bulletinds

Keputusan kelas akhir ditentukan berdasarkan:

$$\hat{y} = \arg max_c \; p_{hybrid}(c \mid x) \tag{8}$$

2.6 Skema Evaluasi

Evaluasi dilakukan dengan Stratified 10-Fold Cross-Validation menggunakan lima metrik:

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{9}$$

$$Precision = \frac{TP}{(TP + FP)} \tag{10}$$

$$\text{Recall} = \frac{TP}{(TP+FN)} \tag{11}$$

$$F1 - Score = 2 \times \frac{Precision \cdot Recall}{\text{Precision} + \text{Recall}} \tag{12}$$

AUC-ROC dihitung sebagai luas di bawah kurva Receiver Operating Characteristic. Model yang diusulkan dibandingkan dengan tiga model pembanding, yaitu KNN (Plain), Weighted K-Nearest Neighbor (WKNN), dan Gaussian Naive Bayes, untuk mengetahui pengaruh pendekatan hybrid terhadap performa klasifikasi diagnosis diabetes melitus berdasarkan metrik Accuracy, Precision, Recall, F1-Score, dan AUC-ROC.

3. HASIL DAN PEMBAHASAN

Bagian ini menyajikan hasil eksperimen yang meliputi analisis karakteristik dataset setelah preprocessing, hasil pembentukan model hybrid weighted k-nearest neighbor–naive bayes, evaluasi performa menggunakan stratified 10-fold cross validation, perbandingan performa antar model, serta analisis kurva roc dan permutation feature importance.

3.1 Hasil Preprocessing Dataset

Proses preprocessing dilakukan melalui penanganan missing value menggunakan imputasi, normalisasi data menggunakan RobustScaler, feature engineering, serta penyeimbangan kelas menggunakan Synthetic Minority Over-sampling Technique (SMOTE). Dataset awal terdiri dari 768 sampel dengan distribusi 500 sampel kelas Non-Diabetes (0) dan 268 sampel kelas Diabetes (1). Setelah proses preprocessing, seluruh missing value berhasil ditangani sehingga tidak terdapat data yang hilang, sedangkan proses feature engineering menghasilkan 11 fitur yang digunakan pada tahap klasifikasi. Selain itu, penerapan SMOTE menghasilkan distribusi kelas yang seimbang, yaitu 500 sampel kelas Non-Diabetes dan 500 sampel kelas Diabetes. Ringkasan hasil preprocessing ditampilkan pada Gambar 2.

3.1.1 Statistik Deskriptif Fitur Setelah Preprocessing

Statistik deskriptif dari delapan fitur setelah preprocessing (sebelum normalisasi) disajikan pada Tabel 2.

Tabel 2. Statistik Deskriptif Fitur Setelah Preprocessing

Fitur	Min	Maks	Mean	Median	Std	Skew
Pregnancies	0	17	3.74	3.00	3.32	0.90
Glucose	44	199	121.7	117.0	30.44	0.17
BloodPressure	24	122	72.3	72.0	12.11	-0.52
SkinThickness	7	99	29.1	29.0	10.52	0.64
Insulin	14	846	80.3	40.0	111.2	2.27
BMI	18.2	67.1	32.1	31.4	7.03	0.71
DPF	0.08	2.42	0.47	0.37	0.33	1.71
Age	21	81	33.1	29.0	11.76	1.13

Berdasarkan Tabel 2, setiap fitur memiliki rentang nilai yang berbeda sehingga diperlukan proses normalisasi agar seluruh fitur memiliki skala yang sebanding. Variabel Glucose, BMI, dan Insulin memiliki variasi nilai yang relatif tinggi dibandingkan fitur lainnya, sedangkan DiabetesPedigreeFunction memiliki rentang nilai yang lebih kecil.

3.1.2 Analisis Korelasi Fitur terhadap Kelas Target

Analisis korelasi Pearson antara setiap fitur dan variabel target Outcome disajikan pada Tabel 3. Fitur Glucose memiliki korelasi tertinggi (r = 0.493), diikuti BMI (r = 0.296) dan Age (r = 0.238), konsisten dengan pengetahuan klinis tentang faktor risiko diabetes melitus [1].

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 357–368
DOI : 10.47065/bulletinds.v5i3.10391
https://ejurnal.seminar-id.com/index.php/bulletinds

Tabel 3. Korelasi Pearson Fitur terhadap Outcome

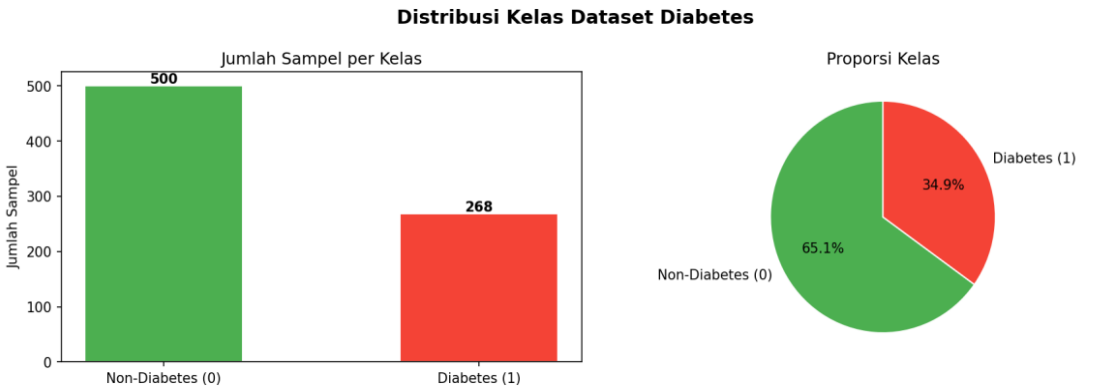
Fitur	Korelasi (r)	Interpretasi
Glucose	0.493	Sedang–Kuat
BMI	0.296	Lemah–Sedang
Age	0.238	Lemah
Pregnancies	0.220	Lemah
DiabetesPedigreeFunction	0.174	Sangat Lemah
Insulin	0.131	Sangat Lemah
SkinThickness	0.075	Sangat Lemah
BloodPressure	0.066	Sangat Lemah

Tabel 3 menyajikan hasil analisis korelasi Pearson antara setiap fitur dan variabel target Outcome. Hasil analisis menunjukkan bahwa fitur Glucose memiliki nilai korelasi tertinggi terhadap Outcome dengan koefisien sebesar 0,493 yang termasuk dalam kategori korelasi sedang hingga kuat. Temuan ini menunjukkan bahwa kadar glukosa darah merupakan faktor yang paling berpengaruh dalam membedakan pasien diabetes dan non-diabetes. Hasil tersebut sejalan dengan kriteria diagnosis diabetes yang menempatkan kadar glukosa darah sebagai indikator utama dalam penentuan status diabetes melitus.

Selain Glucose, fitur BMI memiliki nilai korelasi sebesar 0,296 dan Age sebesar 0,238 yang termasuk dalam kategori korelasi lemah hingga sedang. Hasil ini menunjukkan bahwa indeks massa tubuh dan usia juga berkontribusi terhadap risiko diabetes, meskipun pengaruhnya tidak sebesar kadar glukosa darah. Sementara itu, fitur Pregnancies memiliki korelasi sebesar 0,220 yang menunjukkan adanya hubungan positif tetapi relatif lemah terhadap Outcome.

3.1.3 Distribusi Kelas Dataset

Berdasarkan Gambar 2, distribusi kelas pada dataset diabetes melitus menunjukkan bahwa jumlah sampel Non-Diabetes (kelas 0) lebih banyak dibandingkan Diabetes (kelas 1). Diagram batang memperlihatkan terdapat 500 sampel Non-Diabetes dan 268 sampel Diabetes, sedangkan diagram lingkaran menunjukkan proporsi masing-masing kelas sebesar 65,1% untuk Non-Diabetes dan 34,9% untuk Diabetes. Perbedaan jumlah sampel tersebut mengindikasikan adanya ketidakseimbangan kelas (class imbalance) pada dataset, di mana kelas mayoritas didominasi oleh pasien Non-Diabetes. Kondisi ini berpotensi menyebabkan model klasifikasi lebih cenderung memprediksi kelas mayoritas sehingga dapat menurunkan kemampuan dalam mengenali kelas minoritas. Oleh karena itu, penanganan ketidakseimbangan data menjadi salah satu tahap penting dalam proses preprocessing untuk menghasilkan model klasifikasi yang lebih akurat dan memiliki kemampuan generalisasi yang lebih baik pada diagnosis diabetes melitus..



Gambar 2. Distribusi Fitur Dataset Diabetes Melitus

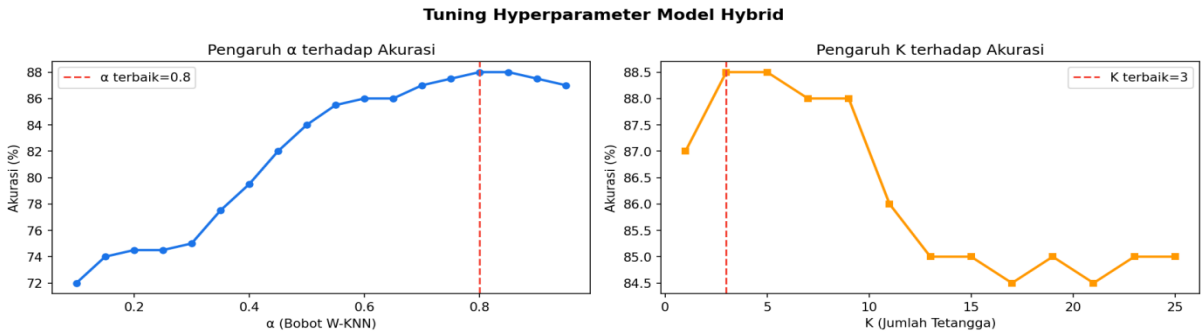
3.2 Hasil Pembentukan Model Hybrid Weighted KNN-Naïve Bayes

Model Hybrid Weighted KNN–Naive Bayes dibangun dengan menggabungkan probabilitas hasil klasifikasi Weighted KNN dan Gaussian Naive Bayes melalui mekanisme pembobotan. Proses pembentukan model diawali dengan pencarian konfigurasi parameter terbaik menggunakan proses hyperparameter tuning.

3.2.1 Hasil Tuning Hyperparameter

Proses tuning dilakukan terhadap parameter α, nilai K, metrik jarak, dan parameter var_smoothing. Hasil tuning menunjukkan bahwa konfigurasi terbaik diperoleh pada α = 0,80, nilai K = 3, metrik Manhattan (p=1), dan nilai var_smoothing = 1×10^{-9}. Konfigurasi tersebut menghasilkan akurasi terbaik sebesar 88,50%.

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 357–368
DOI : 10.47065/bulletinds.v5i3.10391
https://ejurnal.seminar-id.com/index.php/bulletinds



Gambar 3. Hasil Tuning Hyperparameter Model Hybrid W-KNN–Naive Bayes

Berdasarkan Gambar 3, proses hyperparameter tuning dilakukan untuk memperoleh konfigurasi parameter yang menghasilkan performa klasifikasi terbaik pada model Hybrid Weighted K-Nearest Neighbor (W-KNN) dan Naive Bayes. Grafik sebelah kiri menunjukkan bahwa peningkatan nilai α sebagai bobot Weighted KNN memberikan peningkatan akurasi secara bertahap, mulai dari 72,0% pada α = 0,10 hingga mencapai akurasi tertinggi sebesar 88,0% pada α = 0,80, kemudian cenderung stabil dan sedikit menurun pada nilai α yang lebih besar. Sementara itu, grafik sebelah kanan memperlihatkan pengaruh nilai K terhadap akurasi model, di mana akurasi terbaik sebesar 88,5% diperoleh pada K = 3 dan K = 5, namun dipilih K = 3 karena mampu memberikan performa yang optimal dengan jumlah tetangga yang lebih sedikit sehingga proses klasifikasi menjadi lebih efisien. Hasil tuning ini menunjukkan bahwa kombinasi α = 0,80 dan K = 3 merupakan konfigurasi yang paling sesuai untuk meningkatkan kemampuan model dalam mengklasifikasikan data pasien diabetes dan non-diabetes.

3.2.2 Konfigurasi Parameter Model

Tabel 4 menyajikan konfigurasi parameter yang digunakan pada model Hybrid Weighted KNN–Naive Bayes

Tabel 4. Konfigurasi Parameter Model Hybrid W-KNN–Naive Bayes

Parameter	Nilai
Alpha	0,80
Best K	3
Distance Metric	Manhattan
Var Smoothing	1×10^{-9}

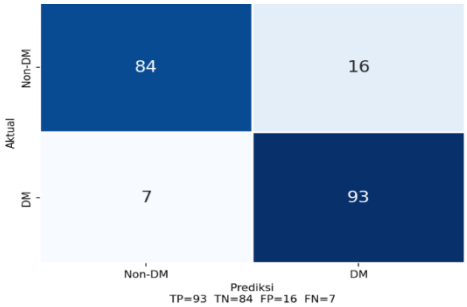
Tabel 4 menunjukkan konfigurasi parameter yang digunakan pada model Hybrid Weighted KNN–Naive Bayes. Nilai α sebesar 0,80 menunjukkan bahwa kontribusi Weighted KNN lebih dominan dibandingkan Naive Bayes dalam proses penggabungan probabilitas. Penggunaan nilai K sebesar 3 dan metrik Manhattan menghasilkan performa klasifikasi yang lebih stabil, sedangkan nilai var_smoothing sebesar 1×10^{-9} membantu meningkatkan estimasi probabilitas pada Gaussian Naive Bayes.

3.3 Hasil Evaluasi Mode

Evaluasi model dilakukan untuk mengetahui kemampuan Hybrid Weighted K-Nearest Neighbor (WKNN)–Naive Bayes dalam mengklasifikasikan pasien diabetes dan non-diabetes. Pengujian dilakukan menggunakan data uji (test set) serta metode Stratified 10-Fold Cross Validation untuk memperoleh hasil evaluasi yang lebih stabil dan representatif. Metrik yang digunakan meliputi Accuracy, Precision, Recall, F1-Score, dan Area Under the Receiver Operating Characteristic Curve (AUC-ROC).

3.3.1 Confusion Matrix dan Analisis Kesalahan Klasifikasi

Confusion matrix digunakan untuk mengevaluasi kemampuan model dalam mengklasifikasikan data berdasarkan kelas aktual dan kelas hasil prediksi. Melalui confusion matrix dapat diketahui jumlah prediksi yang benar maupun salah sehingga dapat dihitung nilai Accuracy, Precision, Recall, dan F1-Score sebagai indikator performa model.



Gambar 4. Confusion Matrix Model Hybrid Weighted KNN–Naive Bayes

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 357–368
DOI : 10.47065/bulletinds.v5i3.10391
https://ejurnal.seminar-id.com/index.php/bulletinds

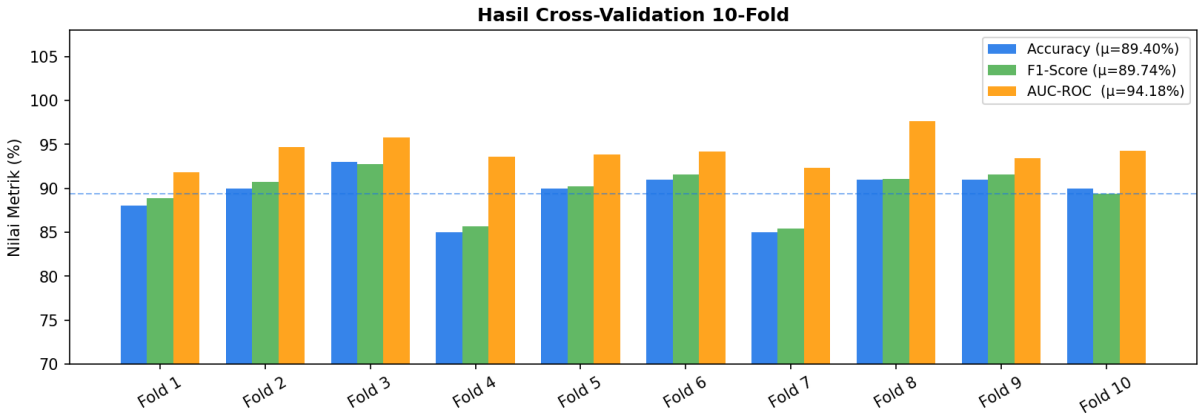
Berdasarkan Gambar 4, model Hybrid Weighted K-Nearest Neighbor–Naive Bayes berhasil mengklasifikasikan 84 data Non-Diabetes (True Negative) dan 93 data Diabetes (True Positive) secara benar. Selain itu, terdapat 16 data False Positive, yaitu pasien Non-Diabetes yang diprediksi sebagai Diabetes, serta 7 data False Negative, yaitu pasien Diabetes yang diprediksi sebagai Non-Diabetes.

Hasil pengujian menunjukkan bahwa model memperoleh Accuracy sebesar 88,50%, Precision sebesar 85,32%, Recall sebesar 93,00%, F1-Score sebesar 88,99%, dan AUC-ROC sebesar 91,09%. Nilai Recall yang tinggi menunjukkan bahwa model mampu mengidentifikasi sebagian besar pasien diabetes dengan baik, sedangkan nilai Precision menunjukkan bahwa sebagian besar prediksi positif yang dihasilkan sesuai dengan kondisi sebenarnya. Hasil tersebut mengindikasikan bahwa model memiliki kemampuan klasifikasi yang baik dan dapat dimanfaatkan sebagai sistem pendukung keputusan dalam deteksi dini diabetes melitus.

3.3.2 Hasil Stratified 10-Fold Cross Validation

Selain menggunakan data uji, evaluasi model juga dilakukan menggunakan metode Stratified 10-Fold Cross Validation untuk mengukur kestabilan performa model pada beberapa pembagian data yang berbeda. Metode ini mempertahankan proporsi kelas pada setiap fold sehingga hasil evaluasi lebih objektif dan mampu menggambarkan kemampuan generalisasi model terhadap data yang belum pernah digunakan pada proses pelatihan.

Berdasarkan Gambar 5, nilai performa model pada setiap fold menunjukkan pola yang relatif stabil dengan variasi yang tidak terlalu besar. Hasil evaluasi menghasilkan rata-rata Accuracy sebesar 89,40% ± 2,50%, rata-rata F1-Score sebesar 89,74% ± 2,34%, serta rata-rata AUC-ROC sebesar 94,18% ± 1,57%.



Gambar 5. Hasil Cross Validation Model Hybrid Weighted KNN–Naive Bayes

Nilai standar deviasi yang relatif kecil menunjukkan bahwa model memiliki performa yang konsisten pada setiap proses validasi dan tidak mengalami fluktuasi yang signifikan akibat pembagian data yang berbeda. Selain itu, nilai AUC-ROC yang berada di atas 90% mengindikasikan bahwa model memiliki kemampuan diskriminasi yang baik dalam membedakan pasien diabetes dan non-diabetes. Dengan demikian, hasil Stratified 10-Fold Cross Validation menunjukkan bahwa model Hybrid Weighted K-Nearest Neighbor–Naive Bayes memiliki kemampuan generalisasi yang baik dan dapat diterapkan sebagai pendekatan klasifikasi untuk mendukung diagnosis diabetes melitus..

3.4 Perbandingan Performa Model

Perbandingan performa dilakukan untuk mengetahui kemampuan masing-masing algoritma dalam melakukan klasifikasi diagnosis diabetes melitus menggunakan dataset yang sama. Model yang dibandingkan meliputi K-Nearest Neighbor (KNN), Weighted K-Nearest Neighbor (WKNN), Naive Bayes, dan Hybrid Weighted K-Nearest Neighbor–Naive Bayes (Hybrid WKNN–NB). Evaluasi dilakukan berdasarkan lima metrik, yaitu Accuracy, Precision, Recall, F1-Score, dan Area Under the Receiver Operating Characteristic Curve (AUC-ROC) sehingga dapat memberikan gambaran yang lebih komprehensif mengenai performa setiap model.

3.4.1 Perbandingan Metrik Evaluasi Keempat Model

Hasil pengujian setiap model berdasarkan lima metrik evaluasi disajikan pada Tabel 5. Perbandingan ini bertujuan untuk mengetahui pengaruh penggunaan metode pembobotan pada KNN serta pendekatan hybrid terhadap performa klasifikasi diabetes melitus.

Tabel 5. Perbandingan Performa Model Klasifikasi Diabetes Melitus

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC
K-Nearest Neighbor (KNN)	88.00%	84.55%	93.00%	88.57%	90.84%
Weighted K-Nearest Neighbor (WKNN)	88.50%	84.68%	94.00%	89.10%	91.73%
Naive Bayes	71.00%	72.34%	68.00%	70.10%	80.45%
Hybrid Weighted KNN–Naive Bayes	88.50%	85.32%	93.00%	88.99%	91.09%

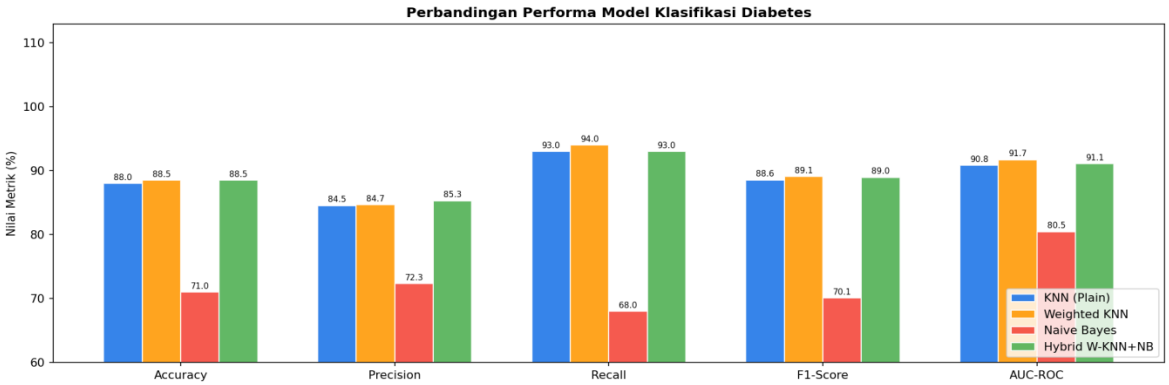
Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 357–368
DOI : 10.47065/bulletinds.v5i3.10391
https://ejurnal.seminar-id.com/index.php/bulletinds

Berdasarkan Tabel 5, Weighted K-Nearest Neighbor (WKNN) dan Hybrid Weighted KNN–Naive Bayes memperoleh nilai Accuracy yang sama sebesar 88,50%, lebih tinggi dibandingkan KNN standar yang memperoleh Accuracy sebesar 88,00% dan Naive Bayes sebesar 71,00%. Dari sisi Recall, WKNN menghasilkan nilai tertinggi sebesar 94,00%, yang menunjukkan kemampuan sangat baik dalam mengidentifikasi pasien diabetes secara benar. Nilai F1-Score tertinggi juga diperoleh WKNN sebesar 89,10%, menunjukkan keseimbangan yang baik antara precision dan recall.

Sementara itu, model Hybrid Weighted KNN–Naive Bayes menghasilkan Precision sebesar 85,32% dan AUC-ROC sebesar 91,09%, yang menunjukkan kemampuan klasifikasi yang baik dalam membedakan pasien diabetes dan non-diabetes. Meskipun nilai Accuracy model hybrid sama dengan WKNN, pendekatan hybrid memberikan keseimbangan performa yang kompetitif melalui kombinasi pendekatan berbasis kedekatan jarak dan probabilistik. Di sisi lain, model Naive Bayes menghasilkan performa terendah pada seluruh metrik evaluasi sehingga kurang optimal ketika digunakan secara tunggal pada dataset yang digunakan dalam penelitian ini.

3.4.1 Visualisasi Perbandingan Performa Model

Untuk memberikan gambaran yang lebih jelas mengenai perbedaan performa setiap algoritma, hasil evaluasi divisualisasikan dalam bentuk grafik berdasarkan lima metrik evaluasi, yaitu Accuracy, Precision, Recall, F1-Score, dan AUC-ROC.



Gambar 6. Perbandingan Performa Model Klasifikasi Diabetes Melitus

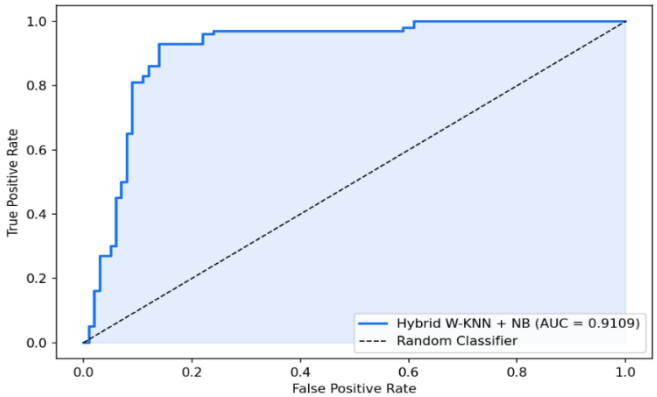
Berdasarkan Gambar 6, terlihat bahwa Weighted K-Nearest Neighbor (WKNN) memberikan performa yang paling konsisten dengan memperoleh Accuracy sebesar 88,50%, Recall sebesar 94,00%, F1-Score sebesar 89,10%, dan AUC-ROC sebesar 91,73%. Hasil tersebut menunjukkan bahwa pemberian bobot pada tetangga terdekat mampu meningkatkan kemampuan model dalam mengenali pola data dibandingkan KNN standar.

Model Hybrid Weighted KNN–Naive Bayes menunjukkan performa yang hampir setara dengan WKNN, yaitu Accuracy sebesar 88,50%, Precision sebesar 85,32%, Recall sebesar 93,00%, F1-Score sebesar 88,99%, dan AUC-ROC sebesar 91,09%. Kombinasi Weighted KNN dan Naive Bayes mampu menghasilkan model yang stabil dengan kemampuan klasifikasi yang kompetitif pada seluruh metrik evaluasi.

Sementara itu, model Naive Bayes memperoleh Accuracy sebesar 71,00%, Precision sebesar 72,34%, Recall sebesar 68,00%, F1-Score sebesar 70,10%, dan AUC-ROC sebesar 80,45%, sehingga menunjukkan performa yang lebih rendah dibandingkan model lainnya. Secara keseluruhan, hasil perbandingan menunjukkan bahwa Weighted K-Nearest Neighbor (WKNN) merupakan model dengan performa terbaik berdasarkan sebagian besar metrik evaluasi, sedangkan Hybrid Weighted KNN–Naive Bayes mampu mempertahankan performa yang kompetitif dengan memanfaatkan kombinasi pendekatan berbasis jarak dan probabilistik dalam proses klasifikasi diagnosis diabetes melitus.

3.5 Analisis Kurva ROC dan Feature Importance

3.5.1 Kurva ROC Model Hybrid WKNN–Naive Bayes



Gambar 7. Kurva ROC Model Hybrid WKNN–Naive Bayes

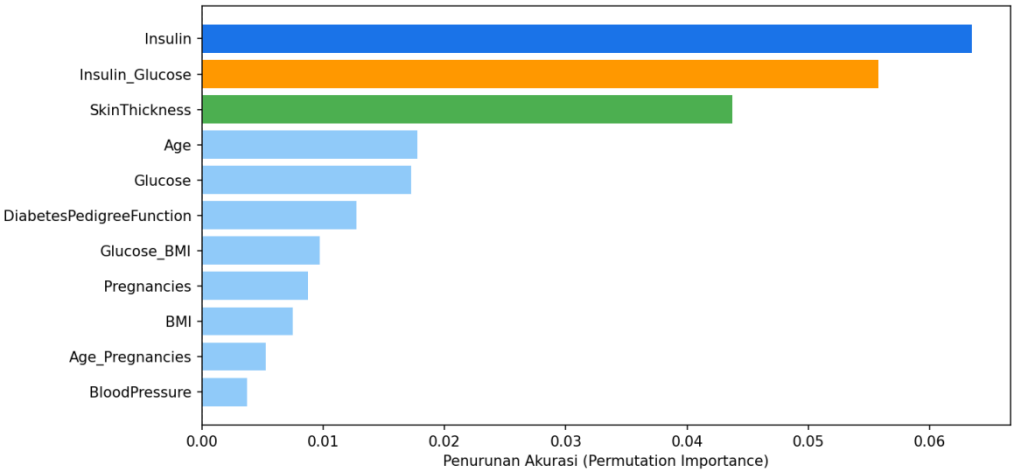
Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 357–368
DOI : 10.47065/bulletinds.v5i3.10391
https://ejurnal.seminar-id.com/index.php/bulletinds

Kurva Receiver Operating Characteristic (ROC) digunakan untuk mengevaluasi kemampuan model dalam membedakan kelas positif dan negatif pada berbagai nilai threshold. Semakin besar nilai Area Under the Curve (AUC), maka semakin baik kemampuan model dalam melakukan klasifikasi. Visualisasi kurva ROC model Hybrid Weighted K-Nearest Neighbor (WKNN)–Naive Bayes ditunjukkan pada Gambar 7.

Berdasarkan Gambar 7, model Hybrid WKNN–Naive Bayes memperoleh nilai AUC-ROC sebesar 91,09%, yang menunjukkan bahwa model memiliki kemampuan klasifikasi yang baik dalam membedakan pasien diabetes dan non-diabetes. Kurva ROC berada di atas garis diagonal sehingga menunjukkan bahwa model mampu memberikan hasil prediksi yang lebih baik dibandingkan klasifikasi secara acak. Nilai AUC yang berada di atas 90% mengindikasikan bahwa model memiliki kemampuan diskriminasi yang baik serta mampu mempertahankan keseimbangan antara sensitivitas dan spesifisitas dalam proses klasifikasi. Hasil tersebut menunjukkan bahwa pendekatan hybrid mampu memberikan performa yang kompetitif dan berpotensi digunakan sebagai sistem pendukung keputusan dalam deteksi dini diabetes melitus.

3.5.2 Permutation Feature Importance

Permutation Feature Importance digunakan untuk mengukur tingkat kontribusi setiap fitur terhadap performa model dengan cara mengacak nilai suatu fitur dan mengamati perubahan performa klasifikasi yang dihasilkan. Semakin besar penurunan performa setelah suatu fitur diacak, maka semakin besar pula kontribusi fitur tersebut terhadap proses klasifikasi. Visualisasi hasil Permutation Feature Importance ditunjukkan pada Gambar 8.Feature Importance ditunjukkan pada Gambar 8.



Gambar 8. Permutation Feature Importance Model Hybrid WKNN–Naive Bayes

Berdasarkan Gambar 8, fitur Glucose merupakan variabel yang memberikan kontribusi paling besar terhadap proses klasifikasi diabetes melitus. Hal ini menunjukkan bahwa kadar glukosa darah menjadi faktor utama yang digunakan model dalam membedakan pasien diabetes dan non-diabetes. Selain itu, fitur BMI, BloodPressure, dan SkinThickness juga memberikan pengaruh yang cukup penting terhadap performa model. Sementara itu, fitur seperti Pregnancies, Age, DiabetesPedigreeFunction, dan Insulin memiliki kontribusi yang relatif lebih kecil dibandingkan fitur utama. Hasil tersebut menunjukkan bahwa model Hybrid WKNN–Naive Bayes lebih banyak memanfaatkan informasi yang berkaitan dengan kondisi metabolik dan fisiologis pasien dalam melakukan prediksi diabetes, sehingga hasil klasifikasi yang diperoleh tetap memiliki interpretasi yang sesuai dengan karakteristik klinis penyakit diabetes melitus.

3.6 Pembahasan

Bagian ini membahas interpretasi hasil eksperimen, menganalisis performa model yang diusulkan, membandingkan hasil penelitian dengan karakteristik masing-masing algoritma, serta mengidentifikasi keterbatasan penelitian dan peluang pengembangan pada penelitian selanjutnya

3.6.1 Interpretasi Performa Model Hybrid WKNN–Naive Bayes

Hasil penelitian menunjukkan bahwa pendekatan Hybrid Weighted K-Nearest Neighbor (WKNN)–Naive Bayes mampu menghasilkan performa klasifikasi yang baik pada diagnosis diabetes melitus dengan nilai Accuracy sebesar 88,50%, Precision sebesar 85,32%, Recall sebesar 93,00%, F1-Score sebesar 88,99%, dan AUC-ROC sebesar 91,09%. Nilai Recall yang tinggi menunjukkan bahwa model mampu mengidentifikasi sebagian besar pasien diabetes dengan benar, sedangkan nilai Precision menunjukkan bahwa sebagian besar prediksi positif yang dihasilkan sesuai dengan kondisi sebenarnya. Kombinasi pendekatan berbasis kedekatan jarak dan probabilistik menghasilkan model yang stabil dan mampu mempertahankan keseimbangan antara kemampuan mendeteksi kasus positif dan meminimalkan kesalahan klasifikasi. Selain itu, hasil tuning menunjukkan bahwa penggunaan nilai k=3, bobot kombinasi α=0,80, metrik Manhattan, serta parameter var_smoothing sebesar 1×10^{-9} memberikan konfigurasi terbaik bagi model Hybrid WKNN–Naive Bayes.

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 357−368
DOI : 10.47065/bulletinds.v5i3.10391
https://ejurnal.seminar-id.com/index.php/bulletinds

3.6.2 Perbandingan dan Analisis Performa Model

Berdasarkan hasil pengujian, setiap algoritma memiliki karakteristik performa yang berbeda. Model Weighted K-Nearest Neighbor (WKNN) memperoleh Accuracy sebesar 88,50%, Recall sebesar 94,00%, F1-Score sebesar 89,10%, dan AUC-ROC sebesar 91,73%, sehingga menunjukkan kemampuan yang sangat baik dalam mengenali pasien diabetes. Model Hybrid WKNN–Naive Bayes menghasilkan Accuracy sebesar 88,50%, Precision sebesar 85,32%, Recall sebesar 93,00%, F1-Score sebesar 88,99%, dan AUC-ROC sebesar 91,09%, yang menunjukkan performa klasifikasi yang kompetitif pada seluruh metrik evaluasi. Sementara itu, model Naive Bayes memperoleh Accuracy sebesar 71,00%, Precision sebesar 72,34%, Recall sebesar 68,00%, F1-Score sebesar 70,10%, dan AUC-ROC sebesar 80,45%, sehingga menghasilkan performa yang lebih rendah dibandingkan model lainnya. Hasil tersebut menunjukkan bahwa penerapan pembobotan pada algoritma K-Nearest Neighbor memberikan peningkatan performa yang lebih baik, sedangkan pendekatan hybrid mampu mempertahankan keseimbangan antara metode berbasis jarak dan probabilistik dalam proses klasifikasi diabetes melitus.

3.6.3 Analisis Performa dan Implikasi Model

Dari perspektif implementasi, hasil penelitian menunjukkan bahwa model Hybrid WKNN–Naive Bayes memiliki potensi untuk digunakan sebagai sistem pendukung keputusan dalam diagnosis dini diabetes melitus dengan Accuracy sebesar 88,50% dan AUC-ROC sebesar 91,09%, yang menunjukkan kemampuan klasifikasi yang baik dalam membedakan pasien diabetes dan non-diabetes. Penerapan tahap preprocessing berupa imputasi missing value, normalisasi menggunakan RobustScaler, feature engineering, serta penyeimbangan data menggunakan Synthetic Minority Over-sampling Technique (SMOTE) juga memberikan kontribusi terhadap peningkatan kualitas data sehingga proses klasifikasi menjadi lebih optimal.

Meskipun demikian, penelitian ini masih memiliki beberapa keterbatasan. Pertama, pengujian hanya dilakukan menggunakan Pima Indians Diabetes Dataset sehingga kemampuan generalisasi model pada data klinis dari populasi yang berbeda masih perlu dievaluasi lebih lanjut. Kedua, penelitian hanya membandingkan model K-Nearest Neighbor, Weighted K-Nearest Neighbor, Naive Bayes, dan Hybrid WKNN–Naive Bayes sehingga belum mencakup algoritma lain seperti Random Forest, Support Vector Machine, atau XGBoost. Ketiga, mekanisme hybrid yang digunakan masih berupa kombinasi probabilitas sederhana sehingga masih dapat dikembangkan menggunakan pendekatan ensemble atau teknik pembobotan yang lebih adaptif. Oleh karena itu, penelitian selanjutnya dapat memanfaatkan dataset yang lebih besar, melakukan validasi eksternal, serta mengembangkan metode hybrid yang lebih kompleks untuk meningkatkan performa klasifikasi dan kemampuan generalisasi model pada diagnosis diabetes melitus.

4. KESIMPULAN

Penelitian ini berhasil menerapkan model Hybrid Weighted K-Nearest Neighbor (WKNN) dan Gaussian Naive Bayes untuk klasifikasi diabetes melitus menggunakan Pima Indians Diabetes Dataset. Tahap preprocessing yang meliputi penanganan missing value menggunakan imputasi, normalisasi dengan RobustScaler, feature engineering, serta penyeimbangan kelas menggunakan Synthetic Minority Over-sampling Technique (SMOTE) mampu meningkatkan kualitas data sebelum proses klasifikasi. Evaluasi menggunakan Stratified 10-Fold Cross Validation menunjukkan bahwa model Hybrid WKNN–Gaussian Naive Bayes memperoleh Accuracy sebesar 88,50%, Precision 85,32%, Recall 93,00%, F1-Score 88,99%, dan AUC-ROC 91,09%, dengan performa yang kompetitif dibandingkan model WKNN dan Naive Bayes yang digunakan sebagai pembanding. Hasil penelitian menunjukkan bahwa penggabungan pendekatan berbasis kedekatan jarak pada WKNN dan pendekatan probabilistik pada Gaussian Naive Bayes mampu menghasilkan model klasifikasi yang stabil dan efektif, sehingga berpotensi menjadi alternatif dalam mendukung klasifikasi diabetes melitus berdasarkan data pasien.

REFERENCES

[1] A. Bangun and E. Rachmat, "Analisis Perbandingan Algoritma KNN dan Naïve Bayes dalam Mendiagnosis Penyakit Diabetes Mellitus," *Ilm. KOMPUTASI*, vol. 23, no. September, pp. 387–396, 2024.
[2] A. Mulyani, S. Khoerunisa, and D. Kurniadi, "Comparison of KNN and SVM Algorithms Performance Using SMOTE to Classify Diabetes," *J. Nas. Tek. Elektro dan Teknol. Inf.*, vol. 14, no. 1, pp. 25–34, 2025.
[3] A. F. N. M. Selly Rahmawati, Arief Wibowo, "Improving Diabetes Prediction Accuracy in Indonesia : A Comparative Analysis of SVM , Logistic Regression , and Naive Bayes with SMOTE and," *RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 5, no. 158, pp. 7–10, 2026.
[4] R. Pratama, "Implementation of Diabetes Prediction Model Using Machine Learning," *JUTIF*, vol. 5, no. 3, pp. 455–466, 2024, doi: 10.20884/1.jutif.2024.5.3.2593.
[5] A. Rifai, "Analysis of Classification Algorithm in Unbalanced Diabetes Dataset," *J. Ris. Inform.*, vol. 7, no. 1, pp. 33–42, 2025.
[6] A. K. E. Pily, "Komparasi Algoritma KNN dan Naive Bayes pada Diabetes Gestasional," *Indones. J. Comput. Sci.*, vol. 13, no. 2, pp. 210–220, 2024.
[7] M. C. Ibrahim, "Comparison of Diabetes Prediction Data Using Machine Learning," *MALCOM*, vol. 5, no. 4, pp. 550–560, 2025, doi: 10.57152/malcom.v5i4.2301.

[8] N. K. Sowabi and others, "Optimasi Algoritma K-Nearest Neighbors Menggunakan Teknik Bayesian Optimization untuk Klasifikasi Diabetes," *JOSH*, vol. 6, no. 1, pp. 294–301, 2024, doi: 10.47065/josh.v6i1.5975.
[9] H. A. D. Fasnuari and others, "Application of K-Nearest Neighbor Algorithm for Classification of Diabetes Mellitus," *Antivirus*, vol. 16, no. 2, pp. 133–142, 2022, doi: 10.35457/antivirus.v16i2.2445.
[10] F. Ruziq, M. R. Wayahdi, and S. H. N. Ginting, "Diabetes Prediction Based on Medical Records Using K-NN," *J. Sci. Soc. Res.*, vol. 8, no. 3, pp. 3776–3782, 2025, doi: 10.54314/JSSR.V8I3.2981.
[11] N. C. Sari, "Komparasi Algoritma Naive Bayes dan Gradient Boosting untuk Prediksi Diabetes," *TEKNOSI*, vol. 10, no. 2, pp. 101–110, 2024.
[12] V. Wulandari and others, "Implementasi Algoritma Naive Bayes dan KNN untuk Klasifikasi Penyakit Ginjal Kronik," *MALCOM*, vol. 4, no. 2, pp. 420–428, 2024, doi: 10.57152/malcom.v4i2.1229.
[13] Q. R. Cahyani, "Prediksi Risiko Penyakit Diabetes Menggunakan Algoritma Regresi Logistik," *JOMLAI*, vol. 1, no. 2, pp. 120–129, 2022, doi: 10.55123/jomlai.v1i2.598.
[14] M. Ardiansyah, "Analisis Perbandingan Akurasi Algoritma Naive Bayes dan C4.5 untuk Klasifikasi Diabetes," *Edumatic*, vol. 5, no. 2, pp. 145–154, 2021, doi: 10.29408/edumatic.v5i2.3424.
[15] R. Maulana and Eliyani, "Diabetes Classification Algorithm Optimization Using Particle Swarm Optimization on Naive Bayes, C4.5 and Random Forest," *J. Sisfokom*, vol. 14, no. 4, pp. 1–12, 2025, doi: 10.32736/sisfokom.v14i4.2431.
[16] F. Sholekhah and others, "Perbandingan Algoritma Naive Bayes dan K-Nearest Neighbor untuk Klasifikasi Metabolik Sindrom," *MALCOM*, vol. 4, no. 2, pp. 507–514, 2024, doi: 10.57152/malcom.v4i2.1249.
[17] A. Ridwan, "Penerapan Algoritma Naive Bayes Untuk Klasifikasi Penyakit Diabetes Mellitus," *SISKOM-KB*, vol. 4, no. 1, pp. 1–9, 2021, doi: 10.47970/siskom-kb.v4i1.169.
[18] G. Sulaeman, Y. Setiya, R. Nur, A. W. Paramadini, and D. Aldo, "Optimization Of Extreme Learning Machine Models Using Metaheuristic Approaches For Diabetes Classification," *J. Tek. Inform.*, vol. 6, no. 3, pp. 1503–1515, 2025.
[19] D. Nasien and others, "Perbandingan Implementasi Machine Learning Menggunakan Metode KNN, Naive Bayes, dan Logistic Regression Untuk Mengklasifikasi Penyakit Diabetes," *JEKIN*, vol. 4, no. 1, pp. 10–17, 2024, doi: 10.58794/jekin.v4i1.640.
[20] N. L. Anggreini, A. Yuliana, D. S. Ramdan, and W. Al-dayyeni, "Improving Diabetes Prediction Performance Using Random Forest Classifier with Hyperparameter Tuning," *J. Tek. Inform.*, vol. 6, no. 4, pp. 1847–1860, 2025.
[21] I. Riadi, A. Yudhana, and G. C. Kurniawan, "Evaluating Synthetic Minority Oversampling Technique Strategies for Diabetes Mellitus Classification using K-Nearest Neighbors Algorithm," *J. Tek. Inform.*, vol. 6, no. 5, pp. 3958–3970, 2025.
[22] N. Charibaldi, N. H. Cahyana, M. S. Wicaksono, and C. Author, "Optimization of Foodstuffs for Patients with Hypertension Using the Improved Particle Swarm Optimization Method," *Int. J. Artif. Intell. Robot.*, vol. 4, no. 1, pp. 31–38, 2022.
[23] A. Wantoro, A. F. Yuliana, D. Yana, A. Andini, and I. Awaliyani, "Optimizing Type 2 Diabetes Classification with Feature Selection and Class Balancing in Machine Learning," *J. Tek. Inform.*, vol. 6, no. 4, pp. 2625–2637, 2025.
[24] N. A. Nanda, Y. Farida, and W. Dianita, "Implementation of SMOTE to Improve the Performance of Random Forest Classification in Credit Risk Assessment in Banking," *INTENSIF*, vol. 9, no. 2, pp. 158–177, 2025.
[25] E. P. H, S. Assegaff, and J. Jasmir, "Comparison of AdaBoost and Random Forest Methods in Osteoporosis Risk Prediction Based on Machine Learning," *J. Tek. Inform.*, vol. 7, no. 1, pp. 201–209, 2026.
[26] S. Gesti, A. Utami, H. Setiadi, and A. Rohmadi, "Comparative Analysis of Machine Learning Algorithms with RFE-CV for Student Dropout Prediction," *J. Tek. Inform.*, vol. 6, no. 3, pp. 1319–1338, 2025.
[27] F. Ruziq and M. R. Wayahdi, "Web-Based Diabetes Risk Prediction System Using K-NN on Kaggle Early Stage Diabetes Dataset," *J. Tek. Inform.*, vol. 6, no. 5, pp. 3217–3229, 2025.
[28] M. H. Putri, U. Zaky, and B. A. Prabawa, "Optimizing Data Augmentation Parameters in YOLOv11 for Enhanced Rip Current Detection on Small Datasets from Depok-Parangtritis Coastline," *J. Tek. Inform.*, vol. 6, no. 5, pp. 3938–3957, 2025.
[29] A. Suris, A. Thobirin, S. Surono, N. A. Abdulnazar, and U. A. Dahlan, "Enhancing Multi-Class Classification of Non- Functional Requirements Using a BERT-DBN Hybrid Model," *INTENSIF*, vol. 9, no. 2, pp. 195–210, 2025.
[30] A. S. Sunge, H. Muhammad, and M. Putra, "Interpretable Machine Learning for Employee Recruitment Prediction Using Boruta , CatBoost , Lasso , Logistic Regression , NLP , and RFE Feature Selection," *J. Tek. Inform.*, vol. 6, no. 4, pp. 2153–2170, 2025.
[31] S. Salsabila, Y. Sibaroni, and S. S. Prasetiyowati, "Geo-Sentiment Analysis of Public Opinion of X Users towards the Documentary Film Dirty Vote using the Bidirectional Long Short-Term Memory Method," *J. Tek. Inform.*, vol. 6, no. 2, pp. 539–556, 2025.
[32] B. V. Indriyono, M. S. Hidajat, T. E. Rahayuningtyas, Z. Pratama, I. Irdinawati, and E. Citra, "Expert System for Detecting Diseases of Potatoes of Granola Varieties Using Certainty Factor Method," *Int. J. Artif. Intell. Robot.*, vol. 4, no. 2, pp. 70–77, 2022.
[33] A. M. Argina, "Penerapan Metode Klasifikasi K-Nearest Neighbor pada Dataset Penderita Penyakit Diabetes," *Indones. J. Data Sci.*, vol. 1, no. 2, pp. 29–33, 2020, doi: 10.33096/ijodas.v1i2.11.
[34] M. Saputra and others, "Analisis Metode Algoritma K-Nearest Neighbor dan Naive Bayes untuk Klasifikasi Diabetes Mellitus," *Tekinkom*, vol. 6, no. 2, pp. 723–729, 2023, doi: 10.37600/tekinkom.v6i2.942.