APPLICATION OF KNN VOTING CLASSIFICATION AND NAIVE BAYES FOR CLASSIFICATION OF TYPE II DIABETES MELLITUS

PENERAPAN KLASIFIKASI VOTING KNN DAN NAIVE BAYES UNTUK KLASIFIKASI DIABETES MELITUS TIPE II

I Gusti Agung Made Suparta Yasa[1], Eddy Muntina Dharma[2], Nengah Widya Utami[3]
Universitas Primakara, Jalan Tukad Badung No. 135, Renon, Denpasar Selatan, Denpasar, Bali.
Email: *igstagungmadesupartayasa@gmail.com[1], eddy@primakara.ac.id[2], widya@primakara.ac.id[3]*

Abstract *- Type II diabetes mellitus (Type II DM) is a public health burden that requires fast and accurate early detection, particularly in primary care settings. Single machine-learning classifiers such as K-Nearest Neighbor (KNN) and Naive Bayes (NB) are widely used but have limitations, including the computational cost of KNN and the strong feature-independence assumption of NB. This study applies an ensemble Voting Classifier (VC) that combines KNN and NB to classify Type II DM using clinical data from 2,390 patients at Mengwi 1 Health Center. Following the CRISP-DM process, we evaluate the models under 80:20 and 70:30 train–test splits using accuracy, precision, recall, F1-score, and ROC/AUC. Compared with the KNN baseline, soft voting consistently improves performance: on the 80:20 split, accuracy increases from 80.33% to 81.59% (+1.26 percentage points) and F1-score from 79.52% to 80.91% (+1.39 %); on the 70:30 split, accuracy increases from 80.47% to 82.01% (+1.54 %) and F1-score from 79.65% to 81.24% (+1.59 %). The soft-voting ensemble also yields higher AUCs, reaching 0.8138 (80:20) and 0.8213 (70:30), and outperforms both single models and hard voting. The novelty of this work lies in demonstrating that a lightweight KNN–NB soft-voting ensemble, designed for the computational constraints of a primary health center and evaluated with repeated cross-validation, can provide small but consistent gains over single classifiers on real DM data. These findings indicate that such an ensemble is a promising building block for clinical decision support in resource-limited primary care, although further calibration, external validation, and prospective testing are still required.*

Keywords *- Type II Diabetes Mellitus, Voting Classifier, K-Nearest Neighbor, Naive Bayes, Classification.*

Abstrak - Diabetes melitus tipe II (DM Tipe II) merupakan beban kesehatan masyarakat yang memerlukan deteksi dini yang cepat dan akurat, terutama pada layanan kesehatan primer. Metode klasifikasi berbasis machine learning seperti K-Nearest Neighbor (KNN) dan Naive Bayes (NB) banyak digunakan, namun memiliki keterbatasan penting, antara lain beban komputasi yang tinggi pada KNN serta asumsi independensi fitur yang kuat pada NB. Penelitian ini menerapkan metode ensemble Voting Classifier (VC) yang mengombinasikan KNN dan NB untuk mengklasifikasikan DM Tipe II menggunakan data klinis rutin dari 2.390 pasien di Puskesmas 1 Mengwi. Mengikuti alur CRISP-DM, model dievaluasi menggunakan pembagian data 80:20 dan 70:30 serta metrik akurasi, precision, recall, F1-score, dan ROC/AUC. Dibandingkan dengan baseline KNN, soft voting menunjukkan peningkatan yang konsisten: pada skema 80:20, akurasi naik dari 80,33% menjadi 81,59% (+1,26 poin persentase) dan F1-score meningkat dari 79,52% menjadi 80,91% (+1,39 %); pada skema 70:30, akurasi naik dari 80,47% menjadi 82,01% (+1,54 %) dan F1-score dari 79,65% menjadi 81,24% (+1,59 %). Ensemble soft voting juga menghasilkan AUC yang lebih tinggi, yaitu 0,8138 (80:20) dan 0,8213 (70:30), serta mengungguli model tunggal maupun hard voting. Kebaruan penelitian ini terletak pada perancangan ensemble KNN–NB yang ringan dan sesuai dengan keterbatasan komputasi di layanan primer, serta diuji menggunakan cross-validation berulang untuk memastikan stabilitas hasil. Temuan ini menunjukkan bahwa ensemble tersebut berpotensi menjadi komponen awal sistem pendukung keputusan klinis pada fasilitas kesehatan dengan sumber daya terbatas, meskipun kalibrasi lebih lanjut, validasi eksternal, dan uji prospektif tetap diperlukan sebelum implementasi nyata.

Kata Kunci - Diabetes Melitus Tipe II, Voting Classifier, K-Nearest Neighbor, Naive Bayes, Klasifikasi.

I. PENDAHULUAN

Diabetes melitus merupakan penyakit kronis yang menjadi masalah kesehatan global dan menimbulkan beban besar bagi sistem pelayanan kesehatan serta kualitas hidup masyarakat di berbagai negara, termasuk Indonesia. Menurut International Diabetes Federation, jumlah penderita diabetes di Indonesia mencapai sekitar 19,5 juta jiwa pada tahun 2021 dan diprediksi meningkat menjadi 28,6 juta jiwa pada tahun 2045. Khusus diabetes melitus tipe II, jumlah penderita diperkirakan meningkat dari 8,4 juta pada tahun 2000 menjadi sekitar 21,3 juta pada tahun 2030[1]. Di tingkat layanan primer, Puskesmas 1 Mengwi menghadapi beban kasus DM Tipe II yang terus meningkat. Dalam satu hari kerja, poli penyakit kronis dapat melayani [isi: rata-rata jumlah] pasien dengan keluhan terkait diabetes dan faktor risikonya, sementara sumber daya dokter dan tenaga laboratorium terbatas. Alur diagnosis konvensional mulai dari anamnesis, pemeriksaan fisik, hingga konfirmasi laboratorium menyebabkan waktu tunggu hasil dan keputusan klinis dapat mencapai 1–2 jam pada jam sibuk. Keterlambatan ini berpotensi menunda penatalaksanaan awal, sementara variasi pengalaman klinisi dan kualitas data rekam medis menimbulkan risiko ketidakkonsistenan dalam klasifikasi awal status diabetes pada pasien baru maupun pasien dengan kontrol teratur. Kondisi tersebut menegaskan perlunya alat bantu klasifikasi yang cepat, andal, dan realistis untuk diintegrasikan dalam alur kerja Puskesmas Kondisi ini menunjukkan bahwa upaya deteksi dan pencegahan dini sangat penting untuk mengendalikan dampak penyakit ini, termasuk di fasilitas pelayanan kesehatan tingkat pertama seperti Puskesmas 1 Mengwi[2].

Penelitian terdahulu yang berjudul "Prediksi Penyakit Diabetes Berdasarkan Perbandingan Klasifikasi Metode K-Nearest Neighbor, Naïve Bayes, dan Decision Tree Menggunakan RapidMiner" menunjukkan bahwa metode K-Nearest Neighbor (KNN) mampu mencapai akurasi sebesar 96,13%. Namun, KNN memiliki kelemahan pada beban komputasi yang tinggi, sensitivitas terhadap noise, serta kinerja yang menurun pada data yang tidak seimbang. Naive Bayes pada penelitian tersebut memperoleh akurasi 84,10%, tetapi mengandalkan asumsi independensi antar fitur yang dalam praktik data medis sering kali tidak sepenuhnya terpenuhi, sehingga dapat menurunkan kinerja model. Adapun Decision Tree memiliki akurasi 93,60%, namun rentan terhadap overfitting, memerlukan proses pruning, dan sensitif terhadap perubahan data. Temuan ini mengindikasikan bahwa penggunaan single classifier, meskipun menjanjikan, masih memiliki keterbatasan ketika dihadapkan pada karakteristik data medis yang kompleks.

Secara umum, penggunaan single classifier seperti Naive Bayes dan KNN dalam klasifikasi diabetes melitus tipe II masih menghadapi beberapa kendala. Naive Bayes, walaupun cepat dan sederhana, mengasumsikan bahwa setiap fitur bersifat independen, padahal dalam data klinis dan gaya hidup pasien sering terdapat korelasi antar variabel yang penting. Selain itu, Naive Bayes sensitif terhadap fitur yang kurang relevan dan berpotensi mengalami penurunan performa jika proses seleksi fitur tidak optimal [3]. Pada dataset yang berukuran besar, KNN juga membutuhkan komputasi yang cukup berat. Keterbatasan-keterbatasan tersebut menyebabkan model cenderung kurang stabil dan kurang optimal dalam menangani kompleksitas data nyata di lapangan. Pada data pasien Puskesmas 1 Mengwi, atribut yang digunakan tidak benar-benar saling bebas, misalnya hubungan antara umur, indeks massa tubuh (IMT), tekanan darah, dan kadar gula darah sewaktu/puasa yang cenderung saling berkorelasi. Kondisi ini berpotensi melanggar asumsi independensi fitur pada Naive Bayes dan menjelaskan mengapa kinerja metode tersebut dapat tereduksi meskipun algoritmanya sederhana dan cepat.

Untuk mengatasi kelemahan klasifikasi tunggal, pendekatan *ensemble learning* menjadi salah satu solusi yang banyak dikembangkan. Salah satu teknik ensemble yang sederhana namun efektif adalah Klasifikasi Voting, yaitu metode yang menggabungkan beberapa algoritma untuk menghasilkan keputusan akhir berdasarkan suara mayoritas (*hard voting*) atau rata-rata probabilitas (*soft voting*)[4]. Dengan mengombinasikan beberapa model, Klasifikasi Voting diharapkan mampu meningkatkan

stabilitas dan akurasi prediksi, mengurangi risiko kesalahan yang mungkin muncul pada satu model tunggal, serta memperbaiki kemampuan generalisasi terhadap data baru. Pendekatan ini dinilai relevan untuk diaplikasikan pada klasifikasi penyakit diabetes, yang memiliki banyak atribut klinis dan gaya hidup dengan pola hubungan yang kompleks[5]. Pemilihan kombinasi Naive Bayes dan KNN dalam Klasifikasi Voting didasarkan pada sifat komplementer keduanya. Naive Bayes unggul dalam kecepatan dan kemampuannya menangani data berdimensi tinggi lewat pendekatan probabilistik, sedangkan KNN efektif dalam mengenali pola lokal dan hubungan non-linear antar data[6].

Berdasarkan uraian di atas, terdapat celah penelitian berupa kebutuhan akan model klasifikasi yang lebih andal, stabil, dan akurat dibandingkan single classifier, khususnya dalam konteks data pasien diabetes melitus tipe II di fasilitas layanan primer seperti Puskesmas 1 Mengwi. Oleh karena itu, penelitian ini bertujuan untuk mengetahui sejauh mana Klasifikasi Voting berbasis *soft voting* antara KNN dan Naive Bayes dapat meningkatkan metrik akurasi, *recall*, *F1-score*, dan AUC dibandingkan masing-masing klasifikasi tunggal pada data DM Tipe II di Puskesmas 1 Mengwi.

II. SIGNIFIKANSI STUDI

A. Penelitian Terdahulu

Berbagai penelitian sebelumnya dilihat pada Tabel 1. menunjukkan bahwa metode klasifikasi berbasis machine learning telah banyak diterapkan untuk deteksi diabetes melitus tipe II.

Tabel 1. Penelitian Terdahulu

No	Judul	Metode	Hasil
1	Penerapan Naive Bayes Untuk Prediksi Penyakit Diabetes dengan Menggunakan Rapid Miner. [7]	Naïve Bayes	Penelitian ini bertujuan mengimplementasikan algoritma Naive Bayes untuk memprediksi kemungkinan seseorang menderita diabetes menggunakan dataset penyakit diabetes dari Kaggle. Data terdiri dari sejumlah atribut klinis seperti jumlah kehamilan, kadar glukosa, tekanan darah, ketebalan kulit, kadar insulin, indeks massa tubuh, diabetes pedigree function, dan usia. Hasil pengujian menunjukkan bahwa algoritma Naive Bayes mampu memberikan performa yang baik dalam memprediksi status diabetes, ditunjukkan oleh nilai akurasi sebesar 80,52%
2	Perbandingan Algoritma Naïve Bayes dan (KNN) untuk Mengetahui Keakuratan Diagnosa Penyakit Diabetes. [8]	Naïve Bayes dan KNN	Penelitian ini membandingkan algoritma Naïve Bayes dan K-Nearest Neighbor (KNN) untuk diagnosis diabetes menggunakan dataset Kaggle berisi 768 data dengan 9 atribut, melalui tahapan KDD (seleksi, preprocessing, transformasi, dan pemodelan dengan Python dan RapidMiner). Hasilnya, Naïve Bayes menjadi model terbaik dengan akurasi 77%, precision 66%, recall 71%, dan AUC 0,83, sedangkan KNN dengan K=3, 5, dan 7 hanya mencapai akurasi 71%, 69%, dan 68% dengan AUC sekitar 0,75–0,76. Temuan ini menegaskan bahwa Naïve Bayes lebih unggul daripada KNN dalam mengenali pola dan mengklasifikasikan penyakit diabetes pada dataset tersebut.
3	Prediksi Diabetes Menggunakan Klasifikasi (KNN) pada Perempuan Indian Pima.[9]	KNN	Penelitian ini memprediksi penyakit diabetes pada perempuan Indian Pima menggunakan algoritma K-Nearest Neighbors (K-NN) dengan dataset resmi berisi 769 pasien dan 9 atribut klinis, seperti kehamilan, glukosa, tekanan darah, ketebalan kulit, insulin, BMI, riwayat diabetes keluarga, usia, dan status diabetes. Analisis dilakukan dengan RapidMiner 10.3 melalui tahapan transformasi data, pembagian data latih–uji (90:10), penerapan KNN, dan evaluasi kinerja. Dengan nilai k = 5, model K-NN

			menghasilkan akurasi 70,13%, dengan presisi 59,09% untuk pasien positif dan 74,55% untuk pasien negatif.
4	Prediksi Risiko Diabetes Dengan Metode Naive Bayes:Identifikasi Faktor Risiko Utama Dan Evaluasi Akurasi Model.[10]	Naïve Bayes	Penelitian ini mengevaluasi kemampuan algoritma Naive Bayes dalam memprediksi risiko diabetes menggunakan dataset dari Kaggle serta mengidentifikasi faktor-faktor risiko utama yang berkontribusi terhadap penyakit tersebut. Proses penelitian meliputi pengumpulan data, pra-proses berupa transformasi dan pemilihan atribut, pembagian data menjadi training dan testing, serta pembangunan model klasifikasi menggunakan RapidMiner. Hasil evaluasi menunjukkan bahwa model Naive Bayes mampu mencapai akurasi sebesar 76,05%, dengan recall 57,34% untuk pasien yang mengidap diabetes dan precision 68,05% pada kelas positif.

Keempat penelitian pada tabel menunjukkan bahwa model yang digunakan masih berupa single classifier (Naive Bayes atau KNN) dengan dataset publik (Kaggle/Pima/NIDDK), sehingga akurasi berkisar 70–80% namun belum menyentuh konteks data klinis riil di layanan primer dan belum mengeksplorasi kombinasi model. Penelitian [7], [8], dan [10] menegaskan bahwa Naive Bayes cenderung lebih unggul dibanding KNN, sementara [9] menunjukkan KNN saja menghasilkan akurasi yang lebih rendah, tetapi semuanya berhenti pada level perbandingan algoritma, bukan perancangan ensemble yang saling melengkapi. Dari celah tersebut, penelitian Anda mengusulkan penggunaan data rekam medis 2.390 pasien DM Tipe II di Puskesmas 1 Mengwi dan penerapan Voting Classifier (soft/hard) yang menggabungkan KNN dan Naive Bayes, lalu membandingkannya dengan model tunggal menggunakan metrik lengkap (akurasi, precision, recall, F1, AUC). Dengan cara ini, penelitian Anda tidak hanya menawarkan peningkatan kinerja yang konsisten dibanding single classifier, tetapi juga menghadirkan model ensemble yang ringan dan lebih relevan sebagai calon komponen awal sistem pendukung keputusan klinis di lingkungan Puskesmas.

A. Metode penelitian

Penelitian ini mengikuti alur kerangka berpikir *Cross-Industry Standard Process for Data Mining*, karena pada umumnya dalam mencari pengetahuan di dalam sebuah basis data diperlukan beberapa proses pengelolaan [11]Adapun proses pengelolaan dapat dilihat pada Gambar 1. sebagai berikut:

Gambar 1. Metode Knowledge Discovery

1. Business Understanding

Penyakit Diabetes Melitus Tipe II merupakan masalah kesehatan yang meningkat di Indonesia, termasuk di Puskesmas 1 Mengwi, dengan karakteristik data medis yang heterogen (biologis, kebiasaan, IMT, tekanan vaskular, glukosa sewaktu dan puasa). Ketepatan klasifikasi sangat krusial untuk mencegah komplikasi akibat keterlambatan diagnosis. Metode KNN dan Naive Bayes banyak digunakan, namun masing-masing memiliki keterbatasan: KNN sensitif terhadap ketidakseimbangan data dan beban komputasi, sedangkan Naive Bayes dibatasi asumsi independensi fitur pada data medis yang kompleks. Oleh karena itu, penelitian ini berfokus mengembangkan model ensemble

learning berbasis voting classifier yang menggabungkan KNN dan Naive Bayes guna meningkatkan akurasi, stabilitas klasifikasi, dan dukungan pengambilan keputusan klinis.

2. Data Understanding

Data penelitian diperoleh dari Puskesmas 1 Mengwi dan mencakup informasi demografis serta hasil pemeriksaan medis pasien dari sisi etika dan tata kelola data, seluruh data pasien diperoleh dari rekam medis Puskesmas 1 Mengwi dengan izin tertulis dari pihak pengelola Puskesmas. Identitas pribadi (nama, alamat, nomor rekam medis) dihilangkan dan digantikan dengan kode anonim sebelum dilakukan pengolahan data. Penggunaan data telah mendapatkan persetujuan etik dari dr. I Made Ariyoga Budiana dengan nomor surat 1340/SKP/DPMPTSP/VII/2025, serta mengikuti kebijakan retensi data klinis yang berlaku di Puskesmas, di mana dataset penelitian disimpan dalam media terproteksi dan hanya diakses oleh tim peneliti untuk keperluan analisis. Dapat dilihat Pada Tabel 2. Atribut. Seluruh atribut tersebut dipilih karena relevan sebagai indicator klinis terkait Diabetes Melitus Tipe II.

Tabel 2. Atribut Dalam Dataset

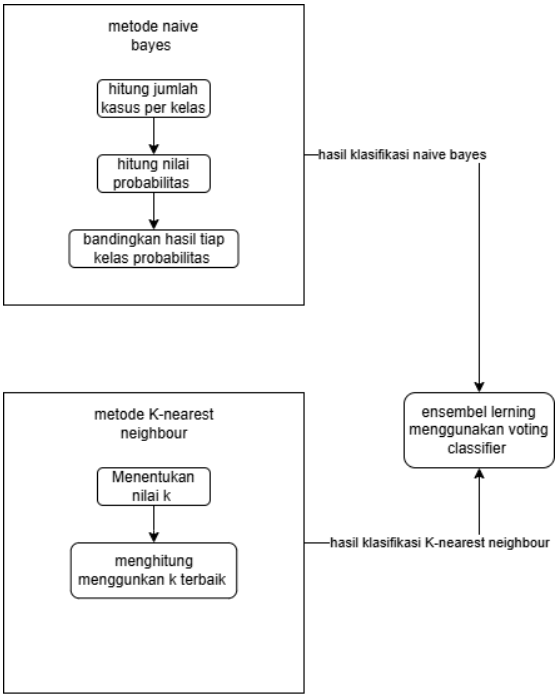
No	Atribut	Jenis Variabel	Nilai
1	Jenis kelamin	Nominal	Laki-Laki, Permpuan
2	Umur	Interval	Dewasa awal: 26-35, Dewasa akhir: 36-45, Lansia awal: 46-55, Lansia akhir: 56-65, Manula: 65-100
3	Konsumsi alkohol	Nominal	Ya, Tidak
4	Merokok	Nominal	Ya, Tidak
5	Hasil IMT	Interval	Kurus: 17,0-18,4, Normal/Ideal: 18,5-25,0, Gemuk: 25,1-27,0, Sangat Gemuk: ≥27,5
6	Sistole	Interval	Normal: 120, Pra-Hipertensi: 120-139, Hipertensi 1: 140-159, Hipertensi 2: 150
7	Diastole	Interval	Normal: 80, Pra-Hipertensi: 80-89, Hipertensi 1: 90-99, Hipertensi 2: 100
8	Gula darah	Interval	Bukan: ‹100, Belum pasti: 100-199, Pasti : ≥200
9	Gula darah puasa	Interval	Bukan: ‹100, Belum pasti: 100-125, Pasti: ≥ 126
10	Gula darah 2 jam	Interval	Bukan: 110-144, Belum pasti: 145-179, Pasti: ≥ 180

3. Data Preparation

Tahap persiapan data meliputi penanganan nilai hilang dengan menghapus baris yang mengandung missing value serta proses pembersihan dan standarisasi format data. Selanjutnya dilakukan label encoding untuk mengubah data kategorik menjadi representasi numerik dan mengelompokkan beberapa atribut ke dalam interval tertentu (misalnya kategori umur, IMT, tekanan darah, dan kadar gula darah). Proses ini bertujuan menyederhanakan struktur data, mengurangi inkonsistensi penulisan, dan mempermudah proses pemodelan klasifikasi[12].

4. Modeling

Pemodelan dilakukan menggunakan dua algoritma utama, yaitu KNN dan Naive Bayes, yang kemudian digabungkan dalam kerangka ensemble learning melalui klasifikasi voting dapat dilihat pada Gambar 2. Pada hard voting, kelas akhir ditentukan berdasarkan suara terbanyak dari kedua model, sedangkan soft voting menggunakan rata-rata probabilitas prediksi untuk menentukan kelas dengan probabilitas tertinggi. Kombinasi ini dirancang untuk meminimalkan kelemahan masing-masing algoritma dan memaksimalkan keunggulannya[12], sehingga diharapkan menghasilkan model klasifikasi Diabetes Melitus Tipe II yang lebih akurat, stabil, dan robust dibandingkan penggunaan model tunggal.



Gambar 2. Alur Penerapan Klasifikasi Voting

A. Naïve Bayes

Naive Bayes adalah algoritma pengelompokan data yang mengandalkan pendekatan probabilistik yang didasarkan pada pengalaman atau data sebelumnya. Metode ini memiliki karakteristik utama berupa asumsi independensi yang kuat antar kondisi atau kejadian yang dianalisis[13]. Beberapa keunggulan dari Naive Bayes meliputi kemudahan pemahaman, kecepatan dalam proses perhitungan, serta kemampuannya dalam menangani data kuantitatif. Selain itu, algoritma ini efektif digunakan meskipun jumlah data pelatihan terbatas, serta efisien dalam penggunaan memori[14]. Berikut ini adalah rumus dasar yang digunakan dalam Naive Bayes:

B. K-Nearest Neighbour (KNN)

Metode KNN dalam pembelajaran mesin berfungsi untuk mengkategorikan data dengan menilai tingkat kemiripan jarak antara data baru dan data yang sudah ada Kelebihan KNN adalah algoritma ini sederhana dan mudah diimplementasikan karena tidak memerlukan proses pelatihan yang rumit. Selain itu, KNN juga efektif untuk digunakan pada dataset skala besar karena dapat bekerja dengan optimal dalam kondisi tersebut [15]. Algoritma ini juga mampu menangani masalah data yang hilang dengan memanfaatkan nilai dari tetangga terdekat sebagai pengganti Kekurangan KNN antara lain adalah kebutuhan untuk penentuan nilai K yang tepat sangat krusial karena nilai yang kecil membuat model rentan terhadap noise, sedangkan nilai yang besar dapat mengurangi ketepatan dalam pengambilan keputusan klasifikasi. Selain itu, KNN memiliki performa yang lambat karena harus menghitung jarak dari semua data latih untuk setiap prediksi, sehingga memerlukan waktu yang lebih lama. Kelemahan lain adalah tidak efisienan ketika terdapat banyak fitur yang tidak relevan, karena semua fitur dianggap penting, yang pada akhirnya dapat mempengaruhi kinerja KNN secara keseluruhan[16]. Adapun rumus KNN:

$$\mathrm{d}(x1, x2) = \sqrt{\sum_{i=1}^{n}(x1i - x2i)^2} \quad (1)$$

Keterangan :
X1i : data training
X2i : data testing

C. *Voting Classifier*

Ensemble Learning Classifier adalah sebuah pendekatan dalam machine learning yang menggabungkan beberapa model untuk membentuk model yang lebih kuat dan akurat. peningkatan akurasi ini bisa terjadi karena voting *classifier* akan menangkap pola data yang lebih kompleks kemudian akan menggabungkan pola tersebut[17]. Selain itu voting classifier dapat mengurangi risiko kesalahan prediksi yang mungkin terjadi di dalam metode tunggal yaitu dengan cara kesalahan individual model dapat dikompensasi oleh model lain sehingga prediksi akhir menjadi lebih akurat dan dapat diandalkan[18].

5. Evaluation

Kinerja model dievaluasi menggunakan metrik akurasi, precision, recall, dan F1-score untuk memberikan gambaran komprehensif terhadap kemampuan klasifikasi, khususnya dalam mendeteksi pasien positif diabetes. Precision menilai ketepatan prediksi positif, recall mengukur kemampuan menemukan seluruh kasus positif, sedangkan F1-score mengombinasikan keduanya dalam satu nilai keseimbangan. Selain itu, digunakan teknik cross-validation untuk membagi data ke beberapa fold sehingga hasil evaluasi lebih stabil, mengurangi bias terhadap satu subset data, dan meningkatkan keandalan performa model[19].

III. HASIL DAN PEMBAHASAN

1. Penerapan Ensemble Learning Voting Classifier pada Naive Bayes dan K-Nearest Neighbor

Pada tahap ini dilakukan penerapan ensemble learning dengan voting classifier yang menggabungkan dua algoritma, yaitu K-Nearest Neighbor (KNN) dan Naive Bayes, untuk klasifikasi Diabetes Melitus Tipe II di Puskesmas 1 Mengwi. Sebelum pemodelan, data pasien melalui proses data preparation yang mencakup pembersihan data dan label encoding terhadap atribut-atribut seperti umur, kebiasaan merokok, konsumsi alkohol, hasil IMT, tekanan darah, serta kadar gula darah sewaktu, puasa, dan 2 jam. Dataset kemudian dibagi menjadi data latih dan data uji dilihat pada Tabel 3. dengan dua skema, yaitu 80:20 dan 70:30.

Tabel 3. Pembagian Data

Total data	Data traning	Data testing	Perbandingan
2390	1.912	478	80:20
	1.673	717	70:30

Menguji konsistensi performa model dan mengurangi risiko overfitting. Model dasar KNN dan Naive Bayes terlebih dahulu dibangun, kemudian dikombinasikan menggunakan voting classifier dengan dua skema, yaitu *hard voting* dan *soft voting*. Pada *hard voting*, kelas akhir ditentukan berdasarkan mayoritas suara dari kedua model[20], sedangkan pada *soft voting* keputusan kelas diambil berdasarkan agregasi probabilitas prediksi dari masing-masing model [21]. Pendekatan ensemble ini dirancang untuk meminimalkan kelemahan masing-masing algoritma (sensitivitas KNN terhadap ketidakseimbangan kelas dan asumsi independensi fitur pada Naive Bayes) sekaligus memaksimalkan keunggulannya dalam mengenali pola data medis yang kompleks.

2. Perbandingan Kinerja Sebelum dan Sesudah Ensemble Learning dengan Split Data

Pengujian dilakukan untuk membandingkan kinerja model tunggal (KNN dan Naive Bayes) dengan model ensemble voting classifier menggunakan metrik akurasi, precision, recall, dan F1-score.

Tabel 4. Hasil Akurasi Dengan 70:30

Metode	Accuracy (%)	Precision	Recall	F1 Score
Naïve Bayes	81.73	82.19	81.73	81.03
KNN	80.47	80.94	80.47	79.65
Soft Voting	82.01	82.70	82.01	81.24
Hard Voting	81.03	80.99	81.03	80.58

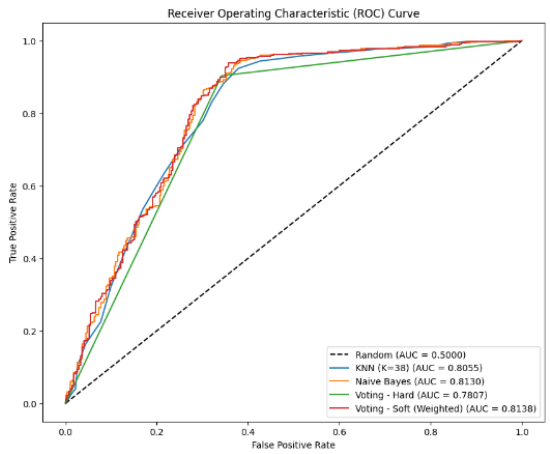
Tabel 5. Hasil Akurasi Dengan 80:20

Metode	Accuracy (%)	Precision	Recall	F1 Score
Naïve Bayes	80.33	80.75	80.33	79.52
KNN	80.33	80.75	80.33	79.52
Soft Voting	81.59	81.97	81.59	80.91
Hard Voting	81.38	81.25	81.38	81.04

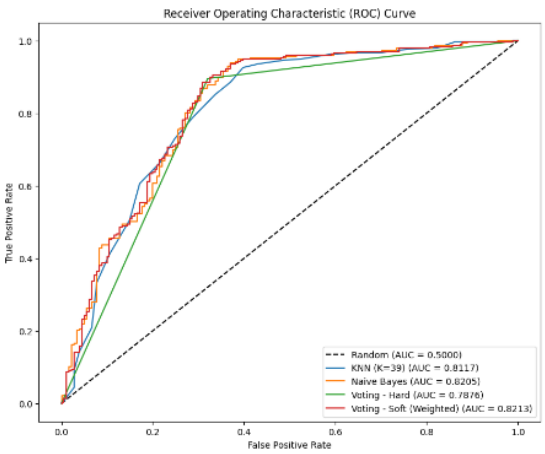
Jika dibandingkan dengan model KNN tunggal, ensemble soft voting memberikan peningkatan performa yang konsisten pada kedua skema pembagian data. Pada skema 70:30 (Tabel 4), akurasi naik dari 80,47% menjadi 82,01% (+1,54 %) dan F1-score dari 79,65% menjadi 81,24% (+1,59%). Pada skema 80:20 (Tabel 5), akurasi meningkat dari 80,33% menjadi 81,59% (+1,26%) dan F1-score dari 79,52% menjadi 80,91% (+1,39%). Dibandingkan Naive Bayes, peningkatan akurasi dan F1-score soft voting sedikit lebih kecil tetapi tetap positif pada kedua skema. Secara praktis, meskipun besaran peningkatan ini tergolong moderat (1–2 poin persentase), pola peningkatan yang konsisten di dua skema pembagian data dan di seluruh metrik utama menunjukkan bahwa ensembel soft voting mampu memberikan perbaikan yang stabil dibandingkan menggunakan model tunggal saja.

3. *Analisis Hasil ROC*

Gambar 2. ROC Dari Pembagian Data 80:20 dan Gambar 3. ROC Dari Pembagian Data 70:30. menunjukkan bahwa model ensemble *soft voting* memiliki performa terbaik dibandingkan model tunggal maupun *hard voting*, dengan area di bawah kurva (AUC) yang lebih tinggi. Hal ini mengindikasikan kemampuan diskriminasi yang lebih baik antara kelas positif dan negatif. Temuan ini sejalan dengan literatur yang menyatakan bahwa *soft voting* umumnya memberikan performa lebih baik karena memanfaatkan informasi probabilistik dari masing-masing model, bukan sekadar suara mayoritas. Secara keseluruhan, hasil pengujian menunjukkan bahwa penerapan ensemble learning dengan voting classifier (khususnya *soft voting* KNN–Naive Bayes) mampu meningkatkan akurasi, stabilitas, dan kemampuan generalisasi model dalam klasifikasi Diabetes Melitus Tipe II dibandingkan penggunaan model tunggal.



Gambar 2. ROC Dari Pembagian Data 80:20



Gambar 3. ROC Dari Pembagian Data 70:30

IV. KESIMPULAN

Penelitian ini mengevaluasi penggabungan algoritma K-Nearest Neighbor (KNN) dan Naive Bayes ke dalam kerangka Voting Classifier untuk klasifikasi Diabetes Melitus Tipe II pada data pasien Puskesmas 1 Mengwi. Hasil eksperimen pada dua skema pembagian data (80:20 dan 70:30) menunjukkan bahwa ensemble soft voting secara konsisten menghasilkan akurasi, precision, recall, F1-score, dan AUC yang sedikit lebih tinggi dibandingkan masing-masing model tunggal dan hard voting. Temuan ini menunjukkan bahwa ensemble KNN–Naive Bayes yang relatif ringan dan komplementer layak dipertimbangkan sebagai kandidat komponen awal dalam sistem pendukung keputusan klinis (clinical decision support) untuk deteksi dini DM Tipe II di layanan primer, khususnya pada lingkungan dengan sumber daya komputasi terbatas. Namun, penelitian ini belum membahas kalibrasi probabilitas prediksi, analisis biaya klinis atas kesalahan klasifikasi (false positive dan false negative), maupun validasi prospektif di setting klinis nyata. Oleh karena itu, hasil yang diperoleh sebaiknya dipandang sebagai bukti awal (proof-of-concept) yang masih memerlukan serangkaian studi lanjutan sebelum diintegrasikan ke dalam praktik klinis rutin. Berdasarkan keterbatasan yang telah diidentifikasi, beberapa arah pengembangan dapat dilakukan. Pertama, dari sisi data dan metrik, disarankan menerapkan teknik pembelajaran yang peka terhadap ketidakseimbangan kelas (imbalance-aware learning), seperti SMOTE atau ADASYN, class weighting, maupun cost-sensitive learning, serta menambahkan metrik PR-AUC untuk mengevaluasi kemampuan model dalam mendeteksi kelas positif yang relatif lebih jarang. Kedua, dari sisi pemodelan, soft voting KNN–Naive Bayes dapat dibandingkan secara sistematis dengan arsitektur ensemble lain yang lebih kuat, seperti bagging (Random Forest), boosting (XGBoost, LightGBM, CatBoost), dan stacking, disertai ablation study untuk mengidentifikasi fitur yang paling berpengaruh.

REFERENSI

[1] S. Ahmed, H. Adnan, M. A. Khawaja, and A. E. Butler, "Novel Micro-Ribonucleic Acid Biomarkers for Early Detection of Type 2 Diabetes Mellitus and Associated Complications—A Literature Review," *Int J Mol Sci*, vol. 26, no. 2, p. 753, Jan. 2025, doi: 10.3390/ijms26020753.

[2] R. Farnoosh, K. Abnoosian, and R. A. Isewid, "Two Machine-learning Hybrid Models for Predicting Type 2 Diabetes Mellitus," *J Med Signals Sens*, vol. 15, no. 4, Apr. 2025, doi: 10.4103/jmss.jmss_29_24.

[3] L. Jiang, L. Zhang, C. Li, and J. Wu, "A Correlation-Based Feature Weighting Filter for Naive Bayes," *IEEE Trans Knowl Data Eng*, vol. 31, no. 2, pp. 201–213, Feb. 2019, doi: 10.1109/TKDE.2018.2836440.

[4] D. Riccio, F. Maturo, and E. Romano, "Supervised learning via ensembles of diverse functional representations: the functional voting classifier," *Stat Comput*, vol. 34, no. 6, p. 191, Dec. 2024, doi: 10.1007/s11222-024-10503-8.

[5] S. Kumari, D. Kumar, and M. Mittal, "An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier," *International Journal of Cognitive Computing in Engineering*, vol. 2, pp. 40–46, Jun. 2021, doi: 10.1016/j.ijcce.2021.01.001.

[6] S. Jindal, M. Sachdeva, and A. K. S. Kushwaha, "Performance evaluation of machine learning based voting classifier system for human activity recognition," *Kuwait Journal of Science*, Jun. 2022, doi: 10.48129/kjs.splml.19189.

[7] I. Reza Pahlevi, "Penerapan Naive Bayes Untuk Prediksi Penyakit Diabetes dengan Menggunakan Rapid Miner," *Jurnal Cendekia Ilmiah*, vol. 4, no. 4, 2025.

[8] Q. A. Puteri, T. Sagirani, and J. Lemantara, "Perbandingan Algoritma Naïve Bayes dan K-Nearest Neighbor (KNN) untuk Mengetahui Keakuratan Diagnosa Penyakit Diabetes," *Jurnal Nasional Teknologi dan Sistem Informasi*, vol. 9, no. 3, pp. 247–254, Dec. 2023, doi: 10.25077/teknosi.v9i3.2023.247-254.

[9] S. Arrohman and Z. Fatah, "Prediksi Diabetes Menggunakan Algoritma Klasifikasi K-Nearest Neighbors (K-NN) pada Perempuan Indian Pima," *Oktober*, pp. 220–226, 2024, doi: 10.59435/gjmi.v2i10.986.

[10] A. Davinka Sembiring Depari, C. Cha Kirana, C. Nissa Oktariana, and F. Akbar, "PREDIKSI RISIKO DIABETES DENGAN METODE NAIVE BAYES: IDENTIFIKASI FAKTOR RISIKO UTAMA DAN EVALUASI AKURASI MODEL," 2025.

[11] F. Martinez-Plumed *et al.*, "CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories," *IEEE Trans Knowl Data Eng*, vol. 33, no. 8, pp. 3048–3061, Aug. 2021, doi: 10.1109/TKDE.2019.2962680.

[12] H. Wu, Y. Mao, J. Weng, Y. Yu, and J. Wang, "Fractional light gradient boosting machine ensemble learning model: A non-causal fractional difference descent approach," *Information Fusion*, vol. 118, p. 102947, Jun. 2025, doi: 10.1016/j.inffus.2025.102947.

[13] I. G. A. P. Mahendra, I. M. A. Wirawan, and I. G. A. Gunadi, "Enhancement performance of the Naïve Bayes method using AdaBoost for classification of diabetes mellitus dataset type II," *International Journal of Advances in Applied Sciences*, vol. 13, no. 3, p. 733, Sep. 2024, doi: 10.11591/ijaas.v13.i3.pp733-742.

[14] O. Peretz, M. Koren, and O. Koren, "Naive Bayes classifier – An ensemble procedure for recall and precision enrichment," *Eng Appl Artif Intell*, vol. 136, p. 108972, Oct. 2024, doi: 10.1016/j.engappai.2024.108972.

[15] R. K. Halder, M. N. Uddin, Md. A. Uddin, S. Aryal, and A. Khraisat, "Enhancing K-nearest neighbor algorithm: a comprehensive review and performance analysis of modifications," *J Big Data*, vol. 11, no. 1, p. 113, Aug. 2024, doi: 10.1186/s40537-024-00973-y.

[16] J. Lu and H. Gweon, "Random k conditional nearest neighbor for high-dimensional data," *PeerJ Comput Sci*, vol. 11, p. e2497, Jan. 2025, doi: 10.7717/peerj-cs.2497.

[17] K. Pham, D. Kim, S. Park, and H. Choi, "Ensemble learning-based classification models for slope stability analysis," *Catena (Amst)*, vol. 196, p. 104886, Jan. 2021, doi: 10.1016/j.catena.2020.104886.

[18] S. Liu, P. Reviriego, J. A. Hernandez, and F. Lombardi, "Voting Margin: A Scheme for Error-Tolerant *k* Nearest Neighbors Classifiers for Machine Learning," *IEEE Trans Emerg Top Comput*, vol. 9, no. 4, pp. 2089–2098, Oct. 2021, doi: 10.1109/TETC.2019.2963268.

[19] K. Abnoosian, R. Farnoosh, and M. H. Behzadi, "Prediction of diabetes disease using an ensemble of machine learning multi-classifier models," *BMC Bioinformatics*, vol. 24, no. 1, p. 337, Sep. 2023, doi: 10.1186/s12859-023-05465-z.

[20] L. AMALIANA, A. B. ASTUTI, R. S. GADIS, N. A. RABBANI, and N. A. SOVIA, "HARD-VOTING DAN SOFT-VOTING CLASSIFIER: MODEL KLASIFIKASI RISIKO KEMATIAN PADA PASIEN GAGAL GINJAL KRONIK," *E-Jurnal Matematika*, vol. 13, no. 4, p. 210, Nov. 2024, doi: 10.24843/MTK.2024.v13.i04.p464.

[21] B. P. Lohani, A. Dagur, and D. K. Shukla, "Synergistic ensemble classification framework: utilizing a soft voting algorithm for enhanced prediction and diagnosis of diabetes mellitus," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 37, no. 3, p. 1945, Mar. 2025, doi: 10.11591/ijeecs.v37.i3.pp1945-1953.