JTIM : Jurnal Teknologi Informasi dan Multimedia

p-ISSN : 2715-2529
e-ISSN : 2684-9151
https://journal.sekawan-org.id/index.php/jtim

Perbandingan Metode Naïve Bayes dan Random Forest dalam Memprediksi Penyakit Diabetes Melitus pada Klinik Citra Sejati

Mohammad Radja Alyfa Amri[1], Egi Permana[2], Pramana Anwas Panchadria [3] dan Shafirah Fitri[4]

[1] Program Studi Teknik Informatika, Institut Teknologi dan Bisnis Bina Sarana Global, Indonesia.

* Korespondensi: radja.amri28@gmail.com

Sitasi: Amri, M. R. A.; Permana, E.; Panchadria, P. A.; dan Fitri, S. (**2025**). Perbandingan Metode Naïve Bayes dan Random Forest dalam Memprediksi Penyakit Diabetes Melitus pada Klinik Citra Sejati. JTIM: Jurnal Teknologi Informasi Dan Multimedia, 7(3), 847-858. https://doi.org/10.35746/jtim.v7i3.747

Diterima: 21-05-2025
Direvisi: 16-07-2025
Disetujui: 22-07-2025

Abstract: Diabetes mellitus is a chronic disease with a steadily increasing prevalence in Indonesia and is one of the leading causes of death, particularly in urban areas. Early detection of this disease is crucial to prevent serious complications such as heart disease, kidney failure, and vision impairment. In the era of digital transformation, machine learning techniques offer great potential to support early and automated diagnosis with higher accuracy. This study aims to develop a diabetes prediction system based on medical record data using two machine learning algorithms: Naïve Bayes and Random Forest. The dataset was obtained from Klinik Citra Sejati, consisting of 266 patient records with seven clinical features: age, gender, leukocytes, platelets, hematocrit, erythrocytes, and erythrocyte sedimentation rate (ESR). The models were implemented using Python programming language and the Scikit-learn library. Performance evaluation was carried out using the confusion matrix and classification metrics such as accuracy, precision, recall, and F1-score. Furthermore, ROC curve analysis and 95% confidence interval calculation were used to assess the stability and reliability of the predictions. The results showed that the Random Forest algorithm achieved an average accuracy of 89.97% with an AUC of 0.93, while Naïve Bayes achieved an accuracy of 85.97% with an AUC of 0.72. Based on these results, Random Forest is considered more effective for diabetes classification and is recommended as the primary algorithm for the development of clinical decision support systems based on local medical data.

Keywords: Prediction, Diabetes Mellitus, Naïve Bayes, Random Forest.

Abstrak: Diabetes melitus merupakan penyakit kronis yang prevalensinya terus meningkat di Indonesia dan menjadi salah satu penyebab kematian utama, khususnya di wilayah perkotaan. Deteksi dini terhadap penyakit ini sangat penting guna mencegah komplikasi serius seperti penyakit jantung, gagal ginjal, dan gangguan penglihatan. Dalam era digital, metode machine learning memberikan peluang besar untuk membantu proses diagnosis dini secara otomatis dan akurat. Penelitian ini bertujuan untuk mengembangkan sistem prediksi penyakit diabetes berbasis data rekam medis dengan menggunakan dua algoritma machine learning, yaitu Naïve Bayes dan Random Forest. Dataset yang digunakan berasal dari Klinik Citra Sejati, terdiri atas 266 data pasien yang memuat tujuh atribut klinis, yakni umur, jenis kelamin, leukosit, trombosit, hematokrit, eritrosit, dan laju endap darah (LED). Model dibangun menggunakan bahasa pemrograman Python dan library Scikit-learn. Evaluasi dilakukan dengan confusion matrix dan metrik akurasi, precision, recall, serta F1-score. Selain itu, dilakukan juga analisis kurva ROC dan perhitungan confidence interval 95% untuk menilai kestabilan dan keandalan prediksi. Hasil penelitian menunjukkan bahwa algoritma Random Forest memperoleh akurasi rata-rata 89,97% dengan AUC sebesar 0,93, sedangkan Naïve Bayes memperoleh akurasi 85,97% dengan AUC sebesar 0,72. Dengan hasil tersebut, Random Forest dinilai lebih unggul dan direkomendasikan sebagai algoritma utama dalam pengembangan sistem pendukung keputusan medis berbasis data klinis lokal.

Kata kunci: Prediksi, Diabetes Melitus, Naïve Bayes, Random Forest.

1. Pendahuluan

Diabetes melitus merupakan gangguan metabolisme yang terjadi akibat disfungsi pankreas dan ditandai dengan meningkatnya kadar gula darah [1]. Penyakit ini termasuk dalam kategori kronis dan menjadi isu serius dalam dunia kesehatan global. Di Indonesia sendiri, diabetes menempati urutan ketiga sebagai penyebab kematian tertinggi. Berdasarkan data dari *Institute for Health Metrics and Evaluation*, pada tahun 2019 tercatat sekitar 57,42 kematian per 100.000 penduduk disebabkan oleh penyakit ini [2]. Berdasarkan laporan Riskesdas 2018, prevalensi diabetes melitus di Indonesia mencapai 10,9% danberdasarkan data dari *International Diabetes Federation* total penderita penyakit diabetes pada tahun 2024 berjumlah 20,436,400 penduduk. Data Profil Kesehatan Indonesia 2022 dari Kementerian Kesehatan juga menyebutkan bahwa diabetes termasuk dalam lima besar penyebab kematian tidak menular yang paling umum, terutama di wilayah perkotaan.

Menurut *World Health Organization* (WHO), prevalensi diabetes terus meningkat setiap tahunnya, sehingga diperlukan upaya deteksi dini dan penanganan yang efektif guna mencegah komplikasi penyakit seperti gagal ginjal, penyakit jantung dan gangguan penglihatan [3].

Dalam era digitalisasi dan perkembangan teknologi kecerdasan buatan, metode machine learning semakin banyak digunakan dalam dunia medis, termasuk untuk prediksi diabetes. Algoritma seperti Naive Bayes dan Random Forest telah terbukti efektif untuk proses klasifikasi dan prediksi penyakit berdasarkan data pasien dalam mendiagnosis penyakit [4]. Naive Bayes bekerja berdasarkan probabilitas kondisi tertentu untuk memprediksi kemungkinan seseorang menderita diabetes, sedangkan Untuk meningkatkan akurasi klasifikasi, Random Forest menggunakan metode *ensemble learning* untuk menangani dataset yang kompleks [5]. Dengan menerapkan kedua metode ini, Klinik Citra Sejati dapat mengembangkan sistem prediksi yang lebih akurat yang berguna untuk membantu tenaga medis dalam mengambil keputusan yang lebih baik.

Klinik Citra Sejati memiliki peran penting dalam layanan kesehatan kepada masyarakat, khususnya dalam diagnosis dan pengobatan diabetes mellitus. Wawancara medis, pemeriksaan fisik, dan tes laboratorium seperti Gula Darah Puasa (GDP), Tes Toleransi Glukosa Oral (TTGO), dan Tes HbA1c adalah bagian dari proses diagnosis di klinik ini.

Dalam kurun waktu Januari 2024 hingga Mei 2024, Klinik Citra Sejati telah mengumpulkan data sebanyak 266 data rekam medis pasien terkait diabetes, yang dapat menjadi sumber informasi berharga dalam memahami pola diagnosis serta efektivitas metode deteksi dini yang akan digunakan.

Berdasarkan beberapa studi sebelumnya, pendekatan klasifikasi diabetes menggunakan algoritma Naive Bayes dan Random Forest telah banyak dikembangkan, seperti yang dilakukan oleh Anisa et al. (2022), Zuhri et al. (2025), dan Kholish et al. (2024).

Penelitian Anisa et al. (2022) melaporkan bahwa algoritma Naive Bayes berhasil mencapai akurasi sebesar 91,6% pada dataset publik Pima Indians. Akan tetapi, hanya memanfaatkan satu algoritma, yakni Naive Bayes, tanpa membandingkan dengan algoritma lain seperti Random Forest. Selain itu, penelitian tersebut hanya menggunakan satu metrik evaluasi, yaitu akurasi, tanpa memperhatikan aspek lain seperti *precision, recall*, maupun *F1-score* yang penting dalam konteks prediksi penyakit[6].

Penelitian Zuhri et al. (2025) menemukan bahwa Random Forest memiliki akurasi 94%, sedangkan Naive Bayes 78%, namun masih terbatas pada penggunaan dataset publik dari Kaggle (Pima Indians), sehingga kurang merepresentasikan kondisi data klinik lokal di Indonesia. Selain itu, penelitian ini juga tidak mengaitkan fitur data dengan

parameter medis yang umum digunakan dalam pemeriksaan klinis nyata, seperti leukosit, trombosit, hematokrit, dan LED[7].

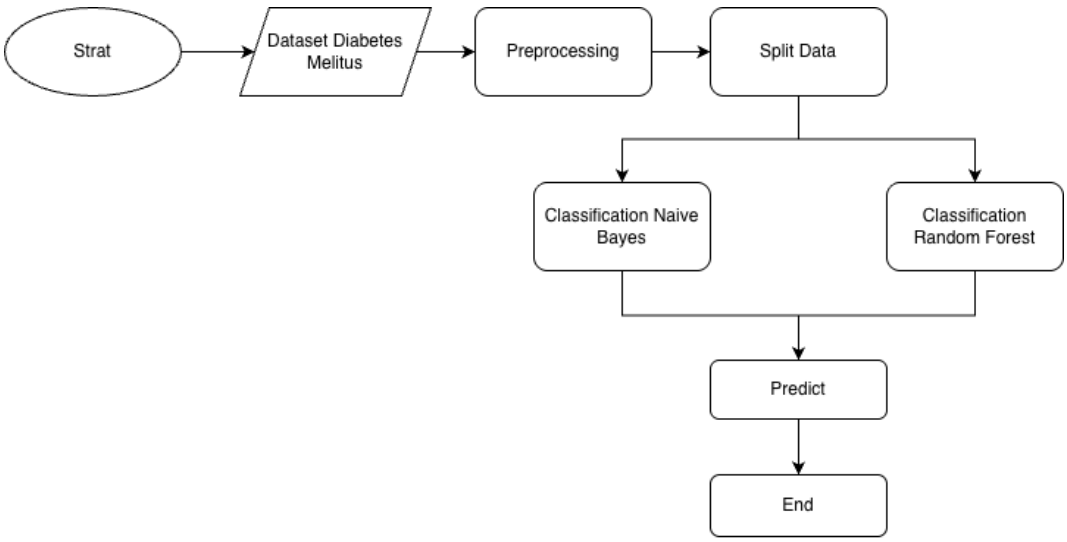
Selanjutnya, Kholish et al. (2024) juga melakukan perbandingan kedua algoritma menggunakan dataset publik yang seluruh pasiennya adalah perempuan. Hal ini menjadikan hasil penelitian kurang generalis terhadap populasi umum. Selain itu, penelitian tersebut tidak mengevaluasi stabilitas model menggunakan metode validasi silang (*cross-validation*) maupun menyertakan *confidence interval*, sehingga keandalan hasilnya belum terverifikasi secara statistik, penelitian ini mendapatkan hasil Naive Bayes unggul (77%) dibanding Random Forest (72%) [5].

Berdasarkan keterbatasan tersebut, penelitian ini hadir untuk mengisi celah (gap) dengan menggunakan data rekam medis asli dari Klinik Citra Sejati, menyertakan berbagai metrik evaluasi seperti akurasi, precision, recall, dan F1-score, memanfaatkan fitur-fitur klinis yang relevan, serta menyertakan pengujian statistik berupa perhitungan confidence interval dan pengajuan hipotesis yang terukur. Sebagai dasar analisis, penelitian ini mengajukan hipotesis. Hipotesis nol (H_0): Tidak terdapat perbedaan signifikan dalam performa klasifikasi antara algoritma Naïve Bayes dan Random Forest dalam memprediksi diabetes berdasarkan data klinik Citra Sejati. Hipotesis alternatif (H_1): Terdapat perbedaan signifikan dalam performa klasifikasi antara algoritma Naïve Bayes dan Random Forest, di mana algoritma Random Forest menghasilkan akurasi dan F1-score yang lebih tinggi dalam memprediksi diabetes.

Hipotesis ini akan diuji melalui proses evaluasi model menggunakan confusion matrix, serta metrik akurasi, precision, recall, F1-score, dan disertai pengujian stabilitas model dengan perhitungan confidence interval 95%. Dengan demikian, penelitian ini bertujuan untuk memberikan kontribusi terhadap pengembangan sistem pendukung keputusan medis berbasis machine learning, khususnya dalam upaya deteksi dini penyakit diabetes yang lebih akurat, efisien, dan berbasis data klinis lokal..

2. Bahan dan Metode

Metode penelitian pada penelitian ini menggunakan metode kuantitatif, di mana penulis melakukan pengolahan data berdasarkan data rekam medis pasien yang menderita diabetes di Klinik Citra Sejati. Data yang diperoleh dilatih menggunakan metode *Naïve Bayes* dan *Random Forest* untuk menghasilkan model prediksi kemungkinan diabetes melitus pada pasien di Klinik Citra Sejati. Penelitian ini mencakup beberapa langkah langkah untuk membuat hasil prediksi, berikut rancangan alur dalam tahapan penelitian :



Gambar 1. Proses Penelitian

Berikut ini penjelasan pada gambar 1 yang menggambarkan alur proses penelitian yang dimulai dari pengumpulan dataset diabetes hingga hasil prediksi :

A. Dataset Diabetes Melitus

Tahap awal pada penelitian ini yaitu mengumpulkan data, Data penelitian ini diperoleh dari rekam medis pasien diabetes di Klinik Citra Sejati. Dataset ini berisi informasi karakteristik pasien yaitu, jenis kelamin, usia, riwayat hipertensi, riwayat penyakit jantung, riwayat merokok, indeks massa tubuh, kadar HbA1c, kadar glukosa darah, serta label diagnosis diabetes.

B. *Preprocessing*

Tahap preprocessing data merupakan langkah krusial dalam proses machine learning, karena kualitas data yang digunakan sangat mempengaruhi performa model yang dihasilkan. Dalam jurnal "*Data Cleaning Survey and Challenges – Improving Outlier Detection Algorithm in Machine Learning*", ditekankan bahwa pembersihan data (*data cleaning*) adalah proses penting yang mencakup identifikasi dan penanganan nilai yang hilang *(missing values*), deteksi outlier, dan transformasi data untuk meningkatkan kualitas dan integritas data sebelum digunakan dalam model *machine learning* [8]. Adapun langkah-langkah *preprocessing* pada penelitian ini adalah sebagai berikut:

1. *Missing Value*

   Data diperiksa untuk mengetahui adanya nilai kosong *(missing*). Apabila ditemukan, maka dilakukan penanganan berupa penghapusan data tidak lengkap atau pengisian nilai menggunakan teknik imputasi seperti *mean* atau modus, teknik mean digunakan untuk data yang bertipe numerik kontinu seperti leukosit, kadar hematokrit, dan LED, tenik ini digunakan ketika data tidak mengandung banyak outlier ekstrem dan sebarannya mendekati rata, maka mean dianggap sebagai estimasi yang representatif dari nilai tengah. Teknik modus digunakan untuk data yang bertipe kategorikal seperti seperti jenis kelamin atau status diagnosis. Data jenis ini tidak memiliki nilai rata-rata yang bermakna, sehingga pengisian dengan nilai terbanyak lebih tepat. Modus mencerminkan kategori yang paling sering muncul dalam dataset dan dianggap sebagai representasi paling umum dari atribut tersebut. Dilakukan pemeriksaan terhadap nilai yang hilang (*missing value*) menggunakan metode visualisasi seperti *heatmap* dan fungsi *.isnull*(). Hasilnya menunjukkan bahwa tidak terdapat nilai yang hilang pada dataset, sehingga tidak diperlukan teknik imputasi atau penghapusan data. *Missing value* perlu dilakukan untuk meningkatkan kinerja dari model machine learning[9] .

2. Penyandian Data Kategorikal (*Encoding*)

   Fitur fitur kategorikal seperti jenis kelamin atau riwayat keluarga yang memiliki format teks dikonversi menjadi format numerik menggunakan metode Label *Encoding* atau *One-Hot Encoding*, agar dapat dikenali oleh model. *Encoding* dilakukan dengan mengubah dua kategori yaitu jenis kelamin menjadi (1,0). Encoding perlu dilakukan agar model dapat digunakan dalam model karena diubah menjadi numerik[10].

3. Normalisasi atau Standarisasi Data

   Penggunaan normalisasi sangat diperlukan untuk menghindari fitur dengan skala besar mendominasi hasil analisis [11].Untuk menjaga konsistensi skala antar fitur, terutama pada algoritma yang sensitif terhadap nilai numerik, dilakukan normalisasi data menggunakan metode *Min-Max Scaling* atau *StandardScaler*. Meskipun *Random Forest* tidak sensitif terhadap skala, normalisasi tetap dilakukan demi keseragaman pengolahan.

C. *Split* Data

Tahap selanjutnya melakukan *split* data, dataset yang sudah siap dibagi menjadi dua bagian, data latihan dan data uji. Metode split train-test digunakan untuk membagi 80% data latihan dan 20% data uji. Pembagian data menggunakan metode *train-test split*. Tujuan dari tahap ini adalah untuk menguji kinerja model pada data yang belum pernah dilihat sebelumnya. Salah satu faktor yang menentukan sejauh mana efektivitas model klasifikasi dalam algoritma pembelajaran mesin adalah cara pembagian data yang diterapkan untuk membagi dataset menjadi bagian-bagian yang berbeda[12].

D. Penerapan metode *naive bayes* dan *random forest*

Setelah *split* data selanjutnya menerapkan ke dalam algoritma *naive bayes* dan *random forest* supaya mendapatkan prediksi yang lebih baik.

- *Naive Bayes*

Model *Naive Bayes* adalah cara sederhana dan cepat untuk mengelompokkan data. Metode ini memprediksi sesuatu berdasarkan data yang sudah ada, dengan menganggap setiap informasi saling berdiri sendiri [13]. Cara ini sering dipakai untuk menganalisis dan memproses data yang sudah dikumpulkan. Algoritma ini akan menghitung probabilitas masing masing faktor risiko terhadap kemungkinan pasien menderita diabetes melitus. Dalam implementasinya, algoritma Naïve Bayes bekerja dengan menghitung probabilitas dari setiap atribut input terhadap dua kemungkinan kelas, yaitu pasien positif diabetes dan negatif diabetes. Sebagai ilustrasi, misalkan terdapat seorang pasien dengan data klinis sebagai berikut: jenis kelamin laki-laki, usia 45 tahun, memiliki riwayat hipertensi dan penyakit jantung, riwayat merokok "formerly smoked", nilai BMI 28.5, kadar HbA1c sebesar 7.2%, dan kadar gula darah 160 mg/dL. Model Naïve Bayes akan menghitung probabilitas masing-masing nilai atribut ini dalam dua kelas (positif dan negatif diabetes), lalu mengalikan semua probabilitas untuk memperoleh probabilitas akhir tiap kelas. Kelas dengan nilai probabilitas tertinggi dipilih sebagai hasil prediksi.

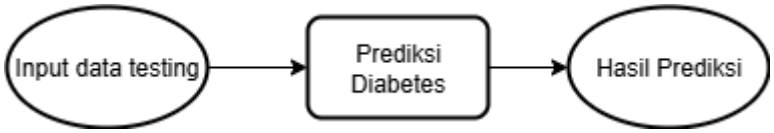
- *Random Forest*

Model *Random Forest* adalah cara untuk memprediksi atau mengelompokkan data dengan menggunakan banyak pohon keputusan (*Decision Tree*) yang dibangun secara acak [4]. Semua pohon bekerja bersama untuk menghasilkan jawaban yang lebih akurat. Model ini digunakan sebagai pembanding untuk melihat keakuratan prediksi terhadap data yang sama. *Random Forest* membuat banyak pohon keputusan dan menggunakan pemungutan suara mayoritas untuk membuat keputusan akhir. Untuk menentukan model yang paling akurat untuk memprediksi diabetes melitus di Klinik Citra Sejati, hasil dari kedua pendekatan dibandingkan. Random Forest bekerja dengan membangun banyak decision tree, di mana masing-masing pohon menggunakan *subset* acak dari data dan fitur. Misalnya, satu pohon dapat memulai klasifikasi dengan aturan jika *blood_glucose_level* > 150 dan HbA1c_*level* > 6.5, maka kemungkinan diabetes. Hasil akhir didapatkan melalui voting mayoritas dari semua pohon. Metode ini mampu mengenali interaksi antar fitur, toleran terhadap *noise*, dan cenderung lebih akurat pada data medis kompleks seperti ini.

E. Evaluasi Hasil Prediksi

Tahap akhir adalah melakukan evaluasi terhadap performa model dan melakukan prediksi terhadap data uji, Tahap evaluasi dilakukan untuk mengukur kinerja model terhadap hasil prediksi yang telah dihasilkan. Evaluasi model ini dilakukan dengan kurva Receiver Operating Characteristic (ROC), yang memberikan gambaran kegunaan model untuk membedakan antara kelas positif dengan kelas negatif. Selain itu, analisis juga dilengkapi dengan penggunaan confusion matrix untuk mengevaluasi metrik-metrik

penting seperti precision, recall, f1-score, dan accuracy guna memperoleh pemahaman yang lebih menyeluruh terhadap performa model [2] dan melakukan evaluasi menggunakan *confidence interval* untuk mengukur stabilitas akurasi dengan interval kepercayaan 95%.



Gambar 2. Alur Proses Prediksi

Pada gambar 2 menunjukan alur proses prediksi dimulai dari memasukkan data testing ke dalam model naïve bayes dan random forest, model melakukan prediksi, lalu model memberikan output hasil prediksi berupa label prediksi yaitu positif diabetes dan negatif diabetes.

3. Hasil

Penelitian ini memakai 266 data, terdiri dari 7 kolom yaitu, jenis kelamin, usia, riwayat hipertensi, riwayat penyakit jantung, riwayat merokok, indeks massa tubuh, kadar HbA1c, kadar glukosa darah, serta label diagnosis diabetes., penelitian ini memanfaatkan 2 metode yaitu *naive bayes* dan r*andom forest*. Data penyakit diabetes terdapat pada gambar 2.

1 to 10 of 10 entries Filter

index	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
0	Female	80.0	0	1	never	25.19	6.6	140	0
1	Female	54.0	0	0	No Info	27.32	6.6	80	0
2	Male	28.0	0	0	never	27.32	5.7	158	0
3	Female	36.0	0	0	current	23.45	5.0	155	0
4	Male	76.0	1	1	current	20.14	4.8	155	0
5	Female	20.0	0	0	never	27.32	6.6	85	0
6	Female	44.0	0	0	never	19.31	6.5	200	1
7	Female	79.0	0	0	No Info	23.86	5.7	85	0
8	Male	42.0	0	0	never	33.64	4.8	145	0
9	Female	32.0	0	0	never	27.32	5.0	100	0

Gambar 2. Data Penyakit Diabetes

Berdasarkan data yang ditampilkan pada Gambar 2, diketahui bahwa dataset yang digunakan dalam penelitian ini terdiri dari 266 data pasien, dengan 7 fitur utama, yaitu: jenis kelamin, usia, riwayat hipertensi, riwayat penyakit jantung, riwayat merokok, indeks massa tubuh, kadar HbA1c, kadar glukosa darah, serta label diagnosis diabetes. Setiap data juga memiliki label diagnosis yang menunjukkan apakah pasien tersebut mengidap diabetes atau tidak.

Sebelum dilakukan pemodelan, data melalui tahap preprocessing. Pada tahap ini, data kategorikal seperti jenis kelamin dan atribut lainnya yang bersifat non-numerik dikonversi menjadi data numerik agar dapat diproses oleh algoritma klasifikasi. Proses encoding dilakukan menggunakan metode label encoding.

Selanjutnya, dilakukan pembagian data menjadi data latih dan data uji menggunakan metode train-test split dengan rasio 80:20. Hasil dari pembagian ini adalah sebanyak 212 data digunakan untuk melatih model (data latih), dan 54 data digunakan untuk menguji performa model (data uji).

Pembagian data dilakukan secara stratified, yaitu mempertahankan proporsi distribusi antara kelas pasien positif diabetes dan negatif diabetes pada kedua bagian data. Hal ini dilakukan untuk menghindari ketimpangan distribusi label yang dapat memengaruhi hasil pelatihan model secara signifikan.

Setelah melalui proses preprocessing dan pembagian data, tahap selanjutnya adalah pelatihan model (training) menggunakan dua metode klasifikasi, yaitu Naive Bayes dan Random Forest. Model dilatih menggunakan data latih sebanyak 212 data, yang telah diproses dan disiapkan sebelumnya.

Pada algoritma Naïve Bayes, digunakan varian GaussianNB karena sebagian besar fitur dalam dataset, seperti usia, bmi, HbA1c_level, dan blood_glucose_level, merupakan data numerik kontinu yang diasumsikan mengikuti distribusi normal. GaussianNB menghitung probabilitas kelas berdasarkan distribusi Gaussian dari tiap fitur, sehingga cocok untuk prediksi medis yang melibatkan parameter laboratorium.

Sementara itu, pada algoritma Random Forest, model dibangun dengan menggunakan 100 pohon keputusan (n_estimators = 100) dan menerapkan fungsi Gini (criterion='gini') sebagai kriteria pemisahan antar node dalam setiap pohon. Random Forest bekerja secara ensemble, di mana setiap pohon belajar dari subset data acak, dan hasil akhir diperoleh melalui voting mayoritas, sehingga meningkatkan akurasi dan mengurangi risiko overfitting, terutama pada dataset medis yang kompleks.

Kedua model dilatih dengan data latih yang telah disesuaikan, termasuk penyeimbangan jumlah kelas melalui teknik oversampling untuk mengatasi ketimpangan data antara pasien diabetes dan non-diabetes. Tujuan dari proses pelatihan ini adalah agar masing-masing model mampu mengenali pola-pola dalam data yang berkaitan dengan kondisi diabetes berdasarkan variabel-variabel medis yang tersedia.

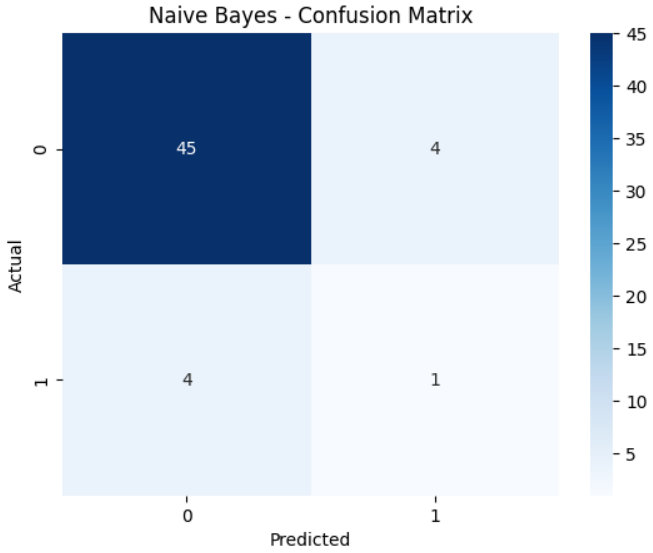
Setelah proses pelatihan selesai, masing-masing model diuji menggunakan data uji sebanyak 54 data, guna mengevaluasi kinerja model dalam mengklasifikasikan pasien sebagai penderita diabetes atau bukan. Evaluasi dilakukan dengan menggunakan metrik akurasi, precision, recall, dan F1-score, serta ditampilkan pula confusion matrix dan ROC curve sebagai alat bantu visualisasi performa masing-masing model.

Evaluasi performa model dianalisis menggunakan metrik akurasi, presisi, *recall,* dan *f1-score*. Hasil analisis performa dari kedua algoritma disajikan pada Tabel 1.

Tabel 1. Hasil evaluasi kedua Algoritma

Metode	Akurasi	Presisi	Recall	F1-score
Naive Bayes	0.85	0.20	0.20	0.20
Random Forest	0.91	0.50	0.20	0.29

Berdasarkan Tabel 1, model Random Forest menunjukkan kinerja yang lebih baik dibandingkan Naive Bayes dalam hal akurasi keseluruhan, *precision*, dan *F1-score* pada kelas 1 (positif diabetes). Hal ini terlihat dari nilai precision sebesar 0,50, dibandingkan dengan 0,20 pada model Naive Bayes. Meskipun recall untuk kelas 1 masih rendah di kedua model (hanya 0,20), Random Forest memiliki macro average F1-score yang lebih tinggi, yaitu 0,62 dibandingkan 0,56 pada Naive Bayes, yang menunjukkan performa keseluruhan yang lebih seimbang antar kelas.

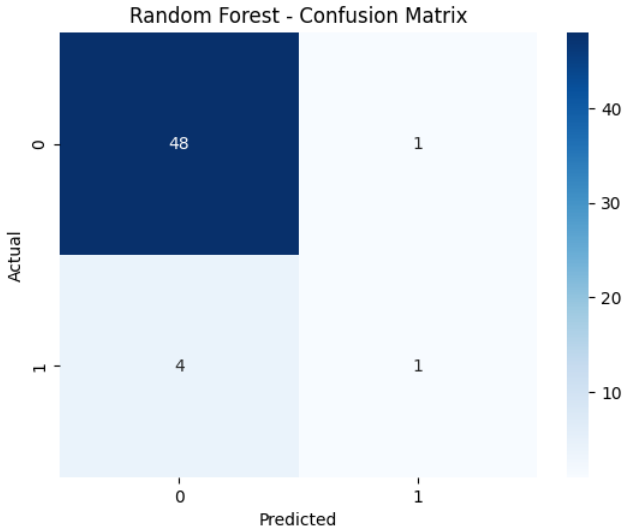


Gambar 3. Hasil Confusion Matrix Pada Algoritma Naive Bayes

Gambar 3 menunjukkan hasil *confusion matrix* pada *naïve bayes*, model Naive Bayes berhasil mengklasifikasikan 45 pasien non-diabetes (kelas 0) dengan benar, dari total 49 pasien. Namun, terdapat 4 kasus false positive, yaitu pasien yang tidak menderita diabetes namun diprediksi sebagai penderita diabetes.

Untuk pasien yang benar-benar menderita diabetes (kelas 1), model hanya berhasil mengidentifikasi 1 kasus dengan benar (*true positive*) dari total 5 pasien. Sisanya, sebanyak 4 pasien salah diklasifikasikan sebagai non-diabetes (*false negative*).

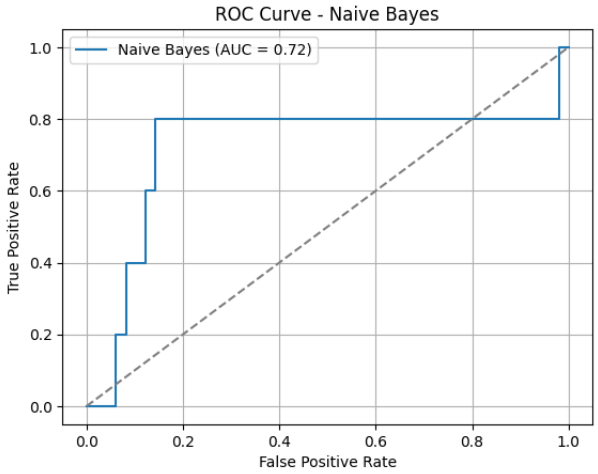
Kondisi ini menunjukkan bahwa model Naive Bayes memiliki kecenderungan untuk mengklasifikasikan sebagian besar data ke kelas mayoritas, yaitu non-diabetes, dan kurang sensitif dalam mendeteksi pasien positif diabetes.



Gambar 4. Hasil Confusion Matrix Pada Algoritma Random Forest

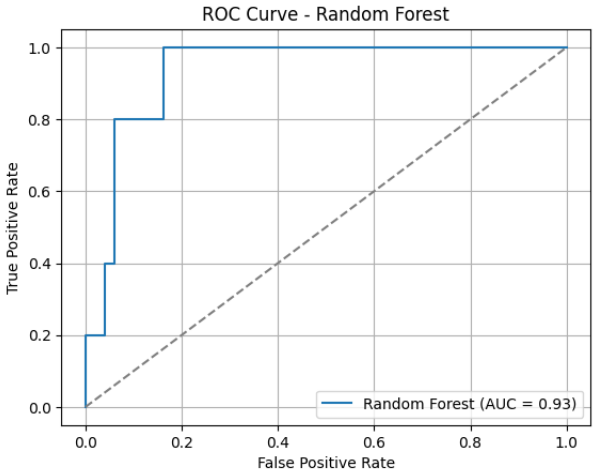
Gambar 4 menunjukkan hasil evaluasi menggunakan *Confusion Matrix* dari algoritma *Random Forest*, bahwa model Random Forest mampu mengklasifikasikan 48 dari 49 pasien non-diabetes dengan benar, dan hanya menghasilkan 1 false positive. Untuk kelas pasien diabetes (kelas 1), hasilnya sama dengan Naive Bayes, yaitu hanya 1 pasien terklasifikasi dengan benar, dan 4 lainnya salah diklasifikasikan sebagai non-

diabetes. Meskipun model ini menunjukkan peningkatan dalam klasifikasi pasien non-diabetes, performanya dalam mengenali pasien yang benar-benar mengidap diabetes masih belum optimal, dengan nilai recall yang rendah pada kelas 1.



Gambar 5. Hasil Receiver Operating Characteristic Pada Algoritma Naïve Bayes

Pada **Gambar 5** ditampilkan kurva ROC dari model Naïve Bayes. Berdasarkan grafik tersebut, diperoleh nilai AUC sebesar 0,72, yang tergolong dalam kategori cukup baik. Hal ini menunjukkan bahwa model Naïve Bayes memiliki kemampuan sedang dalam membedakan antara pasien yang menderita diabetes dan yang tidak. Meskipun model ini ringan dan cepat, namun performanya masih terbatas terutama dalam mendeteksi kelas positif (pasien dengan diabetes). Kurva ROC Naïve Bayes tidak sepenuhnya mendekati sudut kiri atas, yang menunjukkan bahwa model belum optimal dalam menghasilkan prediksi probabilitas yang tajam, terutama terhadap data kelas minoritas.



Gambar 6. Hasil Receiver Operating Characteristic Pada Algoritma Random Forest

Gambar 6 memperlihatkan kurva ROC dari model Random Forest, yang menunjukkan nilai AUC sebesar 0,93. Nilai ini termasuk dalam kategori sangat baik, yang berarti model memiliki kemampuan tinggi dalam membedakan antara kelas positif dan negatif. Kurva ROC mendekati sudut kiri atas grafik, yang mencerminkan tingkat true positive yang tinggi dengan false positive yang rendah pada berbagai ambang prediksi. Dengan AUC yang tinggi ini, dapat disimpulkan bahwa model Random Forest lebih konsisten dan akurat dalam menghasilkan prediksi berbasis probabilitas, sehingga lebih

andal digunakan untuk sistem pendukung keputusan medis dalam kasus deteksi diabetes.

4. Pembahasan

Hasil analisis menunjukkan bahwa algoritma Random Forest memiliki performa yang lebih baik dibandingkan dengan Naïve Bayes dalam mengklasifikasikan pasien diabetes melitus. Random Forest berhasil mencapai akurasi rata-rata sebesar 89,97% dengan rentang confidence interval 95% antara 87,04% hingga 90,74%, sedangkan Naïve Bayes memiliki akurasi rata-rata 85,97% dengan confidence interval 83,33% hingga 87,04%. Hal ini menunjukkan bahwa Random Forest tidak hanya memiliki akurasi lebih tinggi, tetapi juga lebih stabil dan andal secara statistik.

Hasil confusion matrix memperlihatkan bahwa kedua model sama-sama kesulitan dalam mendeteksi pasien positif diabetes. Naïve Bayes hanya berhasil mengklasifikasikan 1 dari 5 pasien positif, dan begitu pula dengan Random Forest. Namun demikian, Random Forest memiliki nilai precision dan F1-score yang lebih baik pada kelas positif, serta error yang lebih kecil dalam memprediksi pasien non-diabetes (false positive lebih rendah). Ini menunjukkan bahwa meskipun recall pada kelas positif sama-sama rendah, Random Forest lebih unggul dalam keseimbangan performa antarkelas.

Keunggulan Random Forest juga diperkuat oleh hasil kurva ROC. Model ini menghasilkan AUC sebesar 0,93, yang menunjukkan kemampuan sangat baik dalam membedakan antara pasien yang menderita dan yang tidak menderita diabetes. Sebaliknya, Naïve Bayes hanya memperoleh AUC sebesar 0,72, yang mengindikasikan performa sedang. Bentuk kurva ROC Random Forest mendekati sudut kiri atas, menandakan kemampuan prediksi probabilitas yang lebih diskriminatif dan konsisten.

Perbedaan ini tidak lepas dari karakteristik masing-masing algoritma. Random Forest mampu menangani pola kompleks dan interaksi non-linier antar fitur, serta memiliki mekanisme penggabungan banyak pohon keputusan yang membuatnya lebih kuat terhadap noise dan ketidakseimbangan data. Sementara itu, Naïve Bayes bekerja dengan asumsi independensi antar fitur yang tidak sepenuhnya terpenuhi pada data medis, seperti leukosit, hematokrit, dan LED, yang kemungkinan besar saling berkorelasi. Hal ini menyebabkan performanya lebih terbatas pada konteks prediksi penyakit.

Penelitian ini sejalan dengan hasil penelitian sebelumnya yang membandingkan algoritma Naïve Bayes dan Random Forest dalam klasifikasi penyakit diabetes, seperti yang dilakukan oleh Huda et al[14] dan Rafli Zuhri el al. [7], Dalam kedua studi tersebut, algoritma Random Forest menunjukkan performa akurasi yang lebih tinggi dibandingkan Naïve Bayes, terutama karena kemampuannya menangani interaksi fitur dan struktur data yang kompleks. Namun demikian, hasil penelitian ini berbeda dengan temuan dari Nurul Anisa et al[6]dan Nurmalasari et al. [15], menunjukkan bahwa algoritma Naïve Bayes memiliki akurasi lebih tinggi dibandingkan Random Forest dalam memprediksi diabetes. Perbedaan ini diduga disebabkan oleh perbedaan karakteristik dataset, jumlah data, metode balancing yang digunakan, serta pemilihan fitur medis yang digunakan dalam model. Dalam penelitian ini, penggunaan data klinik riil dari Klinik Citra Sejati dan distribusi kelas yang kurang seimbang kemungkinan besar turut memengaruhi performa model Naïve Bayes secara signifikan.

Berdasarkan hasil tersebut, disarankan bahwa dalam implementasi sistem prediksi diabetes di instansi kesehatan seperti Klinik Citra Sejati, penggunaan *Random Forest* dapat menjadi pilihan utama untuk menghasilkan prediksi yang lebih akurat. Namun, untuk mengimbangi kompleksitas komputasi dan kebutuhan interpretabilitas, penggabungan metode (*ensemble* kombinasi *Naïve Bayes* dan *Random Forest*) dapat menjadi strategi yang efektif, khususnya dalam sistem berbasis web atau aplikasi klinik yang melibatkan diagnosis awal dan pengambilan keputusan cepat.

5. Kesimpulan

Berdasarkan hasil penelitian yang telah dilakukan terhadap data rekam medis pasien dari Klinik Citra Sejati, diperoleh kesimpulan bahwa algoritma Random Forest menunjukkan kinerja prediksi yang lebih unggul dibandingkan algoritma Naïve Bayes dalam klasifikasi penyakit diabetes melitus. Model Random Forest menghasilkan rata-rata akurasi sebesar 89,97% dengan confidence interval (CI) 95% berada pada rentang 87,04% hingga 90,74%, sedangkan model Naïve Bayes memiliki akurasi rata-rata 85,97% dengan confidence interval 83,33% hingga 87,04%. Nilai ini menunjukkan bahwa Random Forest tidak hanya lebih akurat, tetapi juga lebih stabil dan andal secara statistik.

Keunggulan Random Forest dalam memproses hubungan antar fitur yang kompleks dan non-linier membuatnya lebih efektif dalam menghasilkan prediksi yang tepat pada data medis. Di sisi lain, Naïve Bayes tetap memiliki keunggulan dari segi kecepatan pemrosesan dan kemudahan interpretasi, sehingga tetap relevan digunakan sebagai model awal atau pembanding. Dengan mempertimbangkan hasil akurasi, kestabilan prediksi, serta nilai AUC dari kurva ROC yang mencapai 0,93 pada Random Forest (dibandingkan 0,72 pada Naïve Bayes), maka dapat disimpulkan bahwa Random Forest direkomendasikan sebagai algoritma utama dalam pengembangan sistem pendukung keputusan medis, khususnya untuk kasus prediksi diabetes berbasis data klinis..

Ucapan Terima Kasih: Penulis mengucapkan terima kasih kepada pihak Klinik CITRA SEJATI yang telah berkenan memberikan dataset yang sangat berharga untuk keperluan penelitian ini. Dukungan tersebut sangat membantu dalam proses analisis dan pengembangan model penelitian. Penulis juga menghargai bantuan administratif dan teknis yang telah diberikan selama proses pengumpulan data.

Referensi

[1] Lestari, Zulkarnain, and A. S. ST, "Diabetes Melitus: Review Etiologi, Patofisiologi, Gejala, Penyebab, Cara Pemeriksaan, Cara Pengobatan dan Cara Pencegahan," *Prosiding Biologi Achieving the Sustainable Development Goals*, vol. 7, no. 1, pp. 237–241, Nov. 2021, https://doi.org/10.24252/psb.v7i1.24229.

[2] D. C. Putri Buani, "Deteksi Dini Penyakit Diabetes dengan Menggunakan Algoritma Random Forest," *Jurnal Sains dan Manajemen*, vol. 12, no. 1, pp. 1–8, 2024, https://doi.org/10.31294/evolusi.v12i1.21005.

[3] B. Hartono and S. Ediyono, "Hubungan Tingkat Pendidikan, Lama Menderita Sakit Dengan Tingkat Pengetahuan 5 Pilar Penatalaksanaan Diabetes Mellitus Di Wilayah Kerja Puskesmas Sungai Durian Kabupaten Kbu Raya Kalimantan Barat," *Journal of TSCS1Kep*, vol. 9, no. 1, pp. 49–58, 2024, https://doi.org/10.35720/tscs1kep.v9i01.

[4] M. Ardiansyah, "Model Ensemble Algoritma Naive Bayes Dan Random Forest Dalam Klasifikasi Penyakit Paru-paru Untuk Meningkatkan Akurasi," *SMARTLOCK: Jurnal Sains dan Teknologi*, vol. 2, no. 2, pp. 32–38, 2023, https://doi.org/10.37476/smartlock.v2i2.4407.

[5] M. Kholish, A. Herdianto, R. F. Setiawan, and R. Samsinar, "Perbandingan Algoritma Random Forest dan Naive Bayes dalam Memprediksi Penyakit Diabetes," *HUBISINTEK*, vol. 5, no. 1, pp. 322–328, 2024, Accessed: Jun. 09, 2025. https://ojs.udb.ac.id/index.php/HUBISINTEK/article/view/4757

[6] D. Nurul Anisa and jumanto, "Klasifikasi Penyakit Diabetes Menggunakan Algoritma Naive Bayes," *Dinamika Informatika*, vol. 14, no. 1, pp. 33–42, 2022, https://doi.org/10.35315/informatika.v14i1.9135.

[7] M. Rafli Zuhri and D. Ariatmanto, "Analisis Perbandingan Algoritma Klasifikasi Untuk Identifikasi Diabetes Dengan Menggunakan Metode Random Forest Dan Naïve Bayes," *Jurnal Informatika Teknologi dan Sains*, vol. 7, no. 1, pp. 11–20, Feb. 2025, https://doi.org/10.51401/jinteks.v7i1.5146.

[8] S. Borrohou, R. Fissoune, and H. Badir, "Data cleaning survey and challenges – improving outlier detection algorithm in machine learning," *Journal of Smart Cities and Society*, vol. 2, no. 3, pp. 125–140, Oct. 2023, https://doi.org/10.3233/scs-230008.

[9] M. Riko Anshori Prasetya and A. Mudi Priyatno, "Penanganan Imputasi Missing Values pada Data Time Series dengan Menggunakan Metode Data Mining," *Jurnal Informasi Dan Teknologi*, vol. 5, no. 2, pp. 56–62, 2023, https://doi.org/10.37034/jidt.v5i1.324.

[10] A. Agung, A. Daniswara, I. Kadek, and D. Nuryana, "Data Preprocessing Pola Pada Penilaian Mahasiswa Program Profesi Guru," *Journal of Informatics and Computer Science*, vol. 05, no. 1, pp. 97–100, 2023, https://doi.org/10.26740/jinacs.v5n01.p97-100.

[11] P. Palinggik Allorerung, A. Erna, M. Bagussahrir, and S. Alam, "Analisis Performa Normalisasi Data untuk Klasifikasi K-Nearest Neighbor pada Dataset Penyakit," *Jurnal Informatika Sunan Kalijaga)*, vol. 9, no. 3, pp. 178–191, Sep. 2024, https://doi.org/10.14421/jiska.2024.9.3.178-191.

[12] A. Setiawan, Z. Hadryan Nst, Z. Khairi, and L. Efrizoni, "Klasifikasi Tingkat Risiko Diabetes Menggunakan Algoritma Random Forest," *Jurnal Informatika & Rekayasa Elektronika)*, vol. 7, no. 2, pp. 263–271, 2024, https://doi.org/10.36595/jire.v7i2.1259.

[13] P. C. Pradhani, A. S. Indrayani, N. Azzarah, S. E. Aflikha, F. Zahra, and A. D. Kalifia, "Prediksi Diabetes Mellitus Berdasarkan Data Pasien Sylhet Diabetes Hospital Dengan Metode Naive Bayes," *Scientia: Jurnal Ilmiah Sain dan Teknolagi*, vol. 3, no. 3, pp. 342–353, Jan. 2025, Accessed: Jun. 09, 2025. https://jurnal.researchideas.org/index.php/scientica/article/view/163/151

[14] K. Huda and M. Ula, "Penerapan Naive Bayes, Regresi Logistik, Random Forest, Svm, Dan Knn Untuk Prediksi Diabetes," *SENASTIKA Universitas Malikussaleh*, vol. 1, no. 1, pp. 1–10, Nov. 2024, Accessed: Jun. 09, 2025. https://proceedings.unimal.ac.id/senastika/article/view/853/580

[15] M. D. Nurmalasari, K. Kusrini, and S. Sudarmawan, "Komparasi Algoritma Naive Bayes dan K-Nearest Neighbor untuk Membangun Pengetahuan Diagnosa Penyakit Diabetes," *Jurnal Komtika (Komputasi dan Informatika)*, vol. 5, no. 1, pp. 52–59, Jul. 2021, https://doi.org/10.31603/komtika.v5i1.5140.