EXPERT JURNAL MANAJEMEN SISTEM INFORMASI DAN TEKNOLOGI

Analisis Kinerja Algoritma Random Forest Dengan Model Machine Learning Pada Dataset Penyakit Diabetes

Onassis Yusuf Inonu[1*], Kardita Magda[2], Amarudin[3]
[123]Fakultas Teknik dan Ilmu Komputer, Magister Ilmu Komputer, Universitas Teknokrat Indonesia, Lampung, Indonesia.
[1]onassis_yusuf_inonu@teknokrat.ac.id, [2]kardita_magda@teknokrat.ac.id, [3]amarudin@teknokrat.ac.id.

ABSTRACT – *Diabetes is a metabolic disease that is one of the major health problems in the world. Early detection and accurate diagnosis are essential to prevent long-term complications. With the development of machine learning technology, data-based diabetes risk prediction has become more effective and efficient. This study aims to analyze the performance of the Random Forest algorithm in predicting diabetes status using the Pima Indians Diabetes dataset. The research stages include data pre-processing, model training, performance evaluation, and visualization of results. The dataset used consists of 154 samples with eight clinical features and one target variable. Pre-processing is carried out to handle zero values, data normalization, and division of training and test data. The Random Forest model is trained and evaluated using accuracy, precision, recall, F1-score, confusion matrix, and ROC-AUC curve metrics. The results show that the model produces an accuracy of 78%, with an AUC value of 0.821, indicating excellent discrimination ability between positive and negative diabetes patients. Confusion Matrix and ROC curve visualization help provide a clear interpretation of the model's performance graphically. Based on these results, it can be concluded that the Random Forest algorithm has great potential as a decision support in the medical field, especially for diabetes risk prediction. The use of this model can improve the efficiency and accuracy of early diagnosis, as well as assist medical personnel in making faster and more objective decisions.*

Keywords: Diabetes; Machine Learning; Model Evaluation; Prediction, Random Forest; UCI dataset.

ABSTRAK – Diabetes merupakan penyakit metabolik yang menjadi salah satu masalah kesehatan utama di dunia. Deteksi dini dan diagnosis yang akurat sangat penting untuk mencegah komplikasi jangka panjang. Dengan perkembangan teknologi machine learning, prediksi risiko diabetes berbasis data menjadi lebih efektif dan efisien. Penelitian ini bertujuan untuk menganalisis kinerja algoritma *Random Forest* dalam memprediksi status diabetes menggunakan dataset Pima Indians Diabetes. Tahapan penelitian meliputi pra-pemrosesan data, pelatihan model, evaluasi kinerja, serta visualisasi hasil. Dataset yang digunakan terdiri dari 154 sampel dengan delapan fitur klinis dan satu variabel target. Pra-pemrosesan dilakukan untuk menangani nilai nol, normalisasi data, serta pembagian data latih dan uji. Model Random Forest dilatih dan dievaluasi menggunakan metrik akurasi, presisi, recall, F1-score, confusion matrix, dan kurva ROC-AUC. Hasil menunjukkan bahwa model menghasilkan akurasi sebesar 78%, dengan nilai AUC sebesar 0.82, menandakan kemampuan diskriminasi yang sangat baik antara pasien positif dan negatif diabetes. Visualisasi *Confusion Matrix* dan kurva ROC membantu memberikan interpretasi yang jelas mengenai performa model secara grafis. Berdasarkan hasil tersebut, dapat disimpulkan bahwa algoritma Random Forest memiliki potensi besar sebagai pendukung keputusan dalam bidang medis, khususnya untuk prediksi risiko diabetes. Penggunaan model ini dapat meningkatkan efisiensi dan akurasi diagnosis awal, serta membantu tenaga medis dalam pengambilan keputusan yang lebih cepat dan objektif.

Kata kunci: Diabete; Evaluasi Model; *Machine Learning*; Prediksi, *Random Forest;* UCI *dataset.*

1. PENDAHULUAN

Perkembangan teknologi informasi dan ilmu data telah membawa dampak yang signifikan dalam berbagai bidang, termasuk di sektor kesehatan. Salah satu penerapan machine learning yang menunjukkan potensi besar adalah dalam prediksi dan diagnosis penyakit, seperti diabetes. Diabetes merupakan penyakit kronis yang membutuhkan deteksi dini untuk mencegah komplikasi lebih lanjut. Dengan pendekatan machine learning, analisis terhadap dataset klinis dapat dilakukan untuk mengidentifikasi pola dan memberikan prediksi yang akurat.

Di tengah pesatnya perkembangan teknologi dan data, penerapan *machine learning* dalam bidang kesehatan semakin menunjukkan perannya, terutama dalam mendukung diagnosis penyakit secara lebih cepat dan akurat. Salah satu

penyakit yang menjadi tantangan global adalah diabetes melitus, sebuah penyakit metabolik kronis yang memiliki dampak signifikan terhadap kualitas hidup masyarakat serta beban ekonomi pada sistem kesehatan. Berdasarkan data WHO tahun 2024, jumlah penderita diabetes di seluruh dunia mencapai lebih dari 422 juta orang [1], dengan tingkat prevalensi yang terus meningkat setiap tahunnya. Fenomena ini menunjukkan bahwa diabetes bukan lagi sekadar masalah kesehatan individu, melainkan isu global yang memerlukan solusi inovatif dan efisien dalam deteksi serta penanganannya. Peningkatan prevalensi ini juga menyoroti urgensi untuk mengembangkan alat bantu diagnosis yang dapat diakses secara luas. Fenomena saat ini menunjukkan bahwa proses diagnosis masih banyak dilakukan secara manual oleh tenaga medis, sehingga rentan terhadap keterlambatan atau kesalahan diagnosa karena keterbatasan waktu dan sumber daya manusia.

Sebagai solusi, berbagai penelitian telah mengusulkan penggunaan model prediksi berbasis *machine learning* untuk membantu deteksi dini diabetes. Menurut [2], algoritma seperti *Decision Tree, Naïve Bayes*, dan *Support Vector Machine* (SVM) telah diimplementasikan dalam beberapa studi sebelumnya. Namun, algoritma *Random Forest* dinilai lebih unggul dalam hal ketahanan terhadap *overfitting* dan kemampuan generalisasi yang baik, sebagaimana disampaikan oleh [3] Studi yang dilakukan oleh [4] juga menunjukkan bahwa *Random Forest* memberikan performa yang stabil dalam memprediksi penyakit kompleks seperti diabetes menggunakan dataset Pima Indians Diabetes. Keunggulan Random Forest ini menjadikannya kandidat yang menjanjikan untuk aplikasi klinis, terutama di mana interpretasi model dan keandalan prediksi sangat penting.

Model *Random Forest* adalah salah satu algoritma machine learning yang termasuk dalam kategori ensemble learning. Ensemble learning melibatkan penggabungan hasil dari beberapa model untuk meningkatkan kinerja dan ketepatan prediksi dibandingkan dengan penggunaan satu model tunggal. Dalam konteks Random Forest, model yang digunakan adalah pohon keputusan (*decision trees*), pendekatan ini umumnya digunakan untuk menyisipkan vektor acak dalam pembentukan pohon – pohon keputusan [5]. Namun, Random Forest meminimalkan masalah ini dengan menggabungkan hasil dari sejumlah pohon, yang masing-masing dilatih pada subset data acak yang berbeda. Metode ini membuat Random Forest sangat andal dan tidak mudah terpengaruh oleh gangguan. Keunggulan Random Forest meliputi kemampuan menangani data dengan dimensi tinggi, toleransi terhadap overfitting, serta pengurangan varians dalam prediksi. Oleh karena itu, Random Forest banyak digunakan dalam berbagai penelitian, termasuk dalam bidang medis seperti prediksi penyakit diabetes. Pemanfaatan Random Forest dalam diagnosis diabetes dapat mengurangi beban kerja tenaga medis dan mempercepat proses diagnosis, yang pada akhirnya dapat meningkatkan kualitas perawatan pasien secara keseluruhan.

Berdasarkan latar belakang dan kajian pustaka tersebut, penelitian ini bertujuan untuk menganalisis kinerja algoritma *Random Forest* sebagai model *machine learning* dalam memprediksi risiko diabetes menggunakan dataset standar. Metode penelitian mencakup tahapan pra-pemrosesan data, pelatihan model, evaluasi kinerja dengan metrik seperti akurasi, presisi, *recall* , dan F1-*score* , serta perbandingan hasil dengan algoritma lain sebagai baseline. Hasil penelitian ini diharapkan dapat memberikan kontribusi berupa analisis komprehensif mengenai efektivitas *Random Forest* dalam konteks prediksi penyakit diabetes, serta menjadi referensi bagi pengembangan sistem pendukung keputusan berbasis data di sektor kesehatan. Penelitian ini juga akan menguraikan secara detail setiap tahapan, mulai dari pengumpulan data hingga visualisasi hasil, untuk memastikan transparansi dan reproduktifitas.

2. DASAR TEORI

Machine learning merupakan cabang dari ilmu kecerdasan buatan (*artificial intelligence*) yang berfokus pada pengembangan model atau algoritma yang mampu belajar dari data dan membuat prediksi atau keputusan tanpa diprogram secara eksplisit . Dalam konteks kesehatan, machine learning banyak digunakan untuk mendeteksi penyakit seperti diabetes melitus, yang merupakan salah satu penyakit kronis dengan prevalensi tinggi di seluruh dunia[6]. Penerapan machine learning memungkinkan identifikasi pola tersembunyi dalam dataset medis yang luas, yang mungkin tidak dapat diidentifikasi oleh analisis manual. Hal ini mengarah pada diagnosis yang lebih cepat dan akurat, serta membantu dalam penentuan strategi perawatan yang lebih personal dan efektif

Random Forest adalah algoritma ensemble learning yang menggunakan pendekatan bagging (*bootstrap aggregating*) untuk meningkatkan akurasi dan mengurangi overfitting . Algoritma ini bekerja dengan membangun sejumlah besar decision tree dan menggabungkan hasil dari setiap pohon untuk memberikan prediksi akhir yang lebih stabil dan akurat. Kelebihan utama dari *Random Forest* antara lain:

- Kemampuan menangani data dengan dimensi tinggi.
- Tahan terhadap overfitting karena penggunaan bootstrapping dan averaging.
- Mampu menangani data yang tidak seimbang (imbalanced data), yang sering ditemukan dalam dataset medis. Fleksibilitas dan robustansi Random Forest menjadikannya pilihan yang sangat baik untuk masalah klasifikasi yang kompleks di bidang medis, di mana data seringkali tidak terstruktur atau memiliki fitur yang saling tergantung.

Random Forest adalah model *ensemble* yang menggabungkan banyak *Decision Tree* untuk meningkatkan akurasi dan mengurangi *overfitting* [7].

DOI http://dx.doi.org/10.36448/expert.v15i1.4312
e-ISSN 2745-7265 p-ISSN 2088-5555 EXPERT Vol. 15 No. 1
June 30, 2025 – Hal. 2

A. *Decision Tree* dalam *Random Forest*

Setiap pohon keputusan dalam Random Forest mengikuti aturan pemisahan berdasarkan entropy atau Gini Index [2].

1. *Entropy (Information Gain)*

   Entropy mengukur impurity (ketidakteraturan) dalam dataset, dengan rumus:

$$H(S) = -\sum_{i=1}^{c} p_i \log_2 p_i$$

   di mana:

   - S = himpunan data
   - c = jumlah kelas dalam dataset
   - p_i = probabilitas suatu kelas iii

   Information Gain (IG)**,** digunakan untuk memilih fitur terbaik dalam pemisahan node:

$$IG(S,A) = H(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} H(S_v)$$

   di mana:

- A = fitur yang diuji
- S_u = subset data setelah dibagi berdasarkan fitur A

   Semakin tinggi IG , semakin baik fitur tersebut dalam membagi data. Information Gain membantu dalam menentukan atribut mana yang paling efektif dalam memisahkan data menjadi kelas-kelas yang lebih murni, sehingga pohon keputusan dapat membuat keputusan yang lebih tepat.

2. *Gini Index*

   Alternatif lain untuk memilih fitur adalah ***Gini Impurity,*** yang didefinisikan sebagai:

$$Gini(S) = 1 - \sum_{i-1}^{c} p_i^2$$

   Semakin rendah nilai Gini**,** semakin "murni" data dalam node. Mirip dengan entropy, Gini Index juga mengukur tingkat ketidakmurnian dalam suatu node. Perbedaannya, Gini Index cenderung lebih cepat dalam komputasi karena tidak melibatkan perhitungan logaritma, menjadikannya pilihan yang efisien dalam pembangunan pohon keputusan yang banyak seperti pada Random Forest.

B. *Voting* dalam *Random Forest*

Prediksi akhir dalam *Random Forest* diperoleh dari **mayoritas suara (majority voting)** dari semua pohon:

$$\hat{y} = mode\{T_1(X), T_2(X), \dots, T_N(X)\}$$

di mana:

- $T_i(X)$ = prediksi dari pohon ke-iii
- $\hat{y}$ = hasil akhir prediksi

Mekanisme voting ini merupakan inti dari kekuatan Random Forest. Dengan menggabungkan prediksi dari banyak pohon yang dilatih secara independen pada subset data yang berbeda, model dapat mengurangi bias dan varians, menghasilkan prediksi yang lebih robust dan generalisasi yang lebih baik pada data baru. Ini juga membantu mengatasi masalah overfitting yang sering terjadi pada decision tree tunggal.

C. *Confusion Matrix*

Confusion Matrix adalah tabel yang digunakan untuk mengevaluasi kinerja model klasifikasi dengan membandingkan hasil prediksi model dengan nilai aktual (data sebenarnya). *Confusion matrix* memberikan informasi tentang prediksi benar dan salah yang dilakukan model untuk setiap kelas [8].

$$\begin{bmatrix} TP & FN \\ FP & TN \end{bmatrix}$$

di mana:

- **TP (True Positive)** = Prediksi **positif** dan benar (diabetes)
- **TN (True Negative)** = Prediksi **negatif** dan benar (tidak diabetes)
- **FP (False Positive)** = Prediksi **positif**, tetapi salah (false alarm)
- **FN (False Negative)** = Prediksi **negatif**, tetapi salah (lolos deteksi)

a. Akurasi (*Accuracy*)

Mengukur sejauh mana model memprediksi dengan benar:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

b. *Precision*

Menunjukkan seberapa banyak prediksi positif yang benar:

$$Precision - \frac{TP}{TP + FP}$$

c. *Recall* (*Sensitivity*)

Menunjukkan seberapa banyak kasus positif sebenarnya yang dapat dideteksi:

$$\text{Recall} = \frac{TP}{TP + FN}$$

d. *Specificity*

Mengukur seberapa banyak kelas negatif yang diklasifikasikan dengan benar**:**

$$\text{Specificity} = \frac{TN}{TN + FP}$$

e. F1-*Score*

F1-*Score* adalah **harmonic mean** dari Precision dan Recall:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Dengan Precision = **0.90** dan Recall = **0.82**:

$$F1 - 2 \times \frac{0.90 \times 0.82}{0.90 + 0.82} - 2 \times \frac{0.738}{1.72} \approx 0.86$$

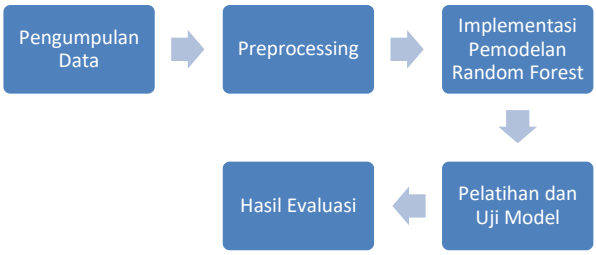
F1-Score digunakan jika keseimbangan antara Precision dan *Recall* penting.

3. METODE PENELITIAN

Penelitian ini merupakan jenis penelitian **kuantitatif eksperimental** [9][10], yang bertujuan untuk menganalisis dan membandingkan kinerja algoritma machine learning,

DOI http://dx.doi.org/10.36448/expert.v15i1.4312
e-ISSN 2745-7265 p-ISSN 2088-5555 EXPERT Vol. 15 No. 1
June 30, 2025 – Hal. 3

khususnya Random Forest, dalam memprediksi penyakit diabetes menggunakan dataset medis . Pendekatan eksperimental dilakukan melalui pengujian berbagai model machine learning serta evaluasi kuantitatif terhadap hasil prediksi yang dihasilkan melalui proses analisis data dan penerapan algoritma machine learning, khususnya *Random Forest*, untuk mengevaluasi kinerjanya dalam memprediksi penyakit *diabetes*. Desain penelitian dirancang secara sistematis agar menghasilkan model prediksi yang valid dan dapat diandalkan. Rangkaian tahapan penelitian ini dapat dilihat pada Gambar 1.



Gambar 1. Tahapan Penelitian

1. Pengumpulan data

Data yang digunakan dalam penelitian ini diambil dari file diabetes.csv, yang merupakan dataset publik yang umum digunakan dalam penelitian terkait diabetes. Dataset ini berisi rekaman kesehatan dari sejumlah individu, termasuk berbagai parameter seperti tingkat glukosa, tekanan darah, indeks massa tubuh (BMI), dan faktor-faktor lain yang relevan dengan kondisi diabetes [11][12]. Dataset yang digunakan berasal dari sumber publik seperti Kaggle , yaitu dataset Pima Indians Diabetes, yang sering digunakan dalam penelitian prediksi penyakit diabetes. Dataset ini berisi informasi klinis pasien seperti kadar gula darah, tekanan darah, BMI, usia, dan riwayat keluarga. Dataset ini dipilih karena kelengkapannya dan relevansinya dengan tujuan penelitian, yaitu memprediksi diabetes berdasarkan parameter kesehatan. Sebelum digunakan, dataset tersebut diperiksa untuk memastikan tidak ada nilai yang hilang (missing values) dan dilakukan pembersihan data jika diperlukan untuk memastikan kualitas data yang baik. Pemilihan dataset yang relevan dan berkualitas tinggi adalah langkah fundamental dalam memastikan validitas dan reliabilitas hasil penelitian. Dataset Pima Indians Diabetes secara luas diakui dalam komunitas penelitian machine learning sebagai benchmark untuk masalah prediksi diabetes, menjadikannya pilihan yang solid untuk studi ini.

Tabel 1. Variabel fitur-fitur

Variabel	Keterangan
Pregnancies	Jumlah kehamilan yang dialami
Glucose	Kadar glukosa dalam darah
BloodPressure	Tekanan darah diastolik (mm Hg)
SkinThickness	Ketebalan lipatan kulit (mm)
Insulin	Kadar insulin dalam darah
BMI	Indeks massa tubuh (kg/m²)
DiabetesPedigreeFunction	Skor riwayat keluarga terhadap diabetes
Age	Usia pasien
Outcome	Status diabetes (0 = tidak diabetes, 1 = diabetes)

2. *Preprocessing*

Sebelum membangun model *machine learning*, diperlukan tahap pra-pemrosesan data untuk memastikan bahwa dataset dalam kondisi optimal dan siap digunakan dalam pelatihan model. Dalam penelitian ini, dataset Pima Indians *Diabetes Dataset* (PIDD) yang diperoleh dari UCI *Machine Learning Repository* mengandung beberapa nilai tidak valid atau bernilai nol pada fitur-fitur tertentu, meskipun secara logika medis nilai tersebut tidak masuk akal (misalnya kadar gula darah = 0). Langkah ini sangat krusial karena kualitas data masukan secara langsung mempengaruhi kinerja dan akurasi model yang dihasilkan. Data mentah seringkali mengandung noise, missing values, atau format yang tidak konsisten yang dapat menghambat proses pembelajaran model.. Oleh karena itu, dilakukan serangkaian langkah pra-pemrosesan sebagai berikut:

a. Identifikasi Data yang Tidak Valid

Dataset awal diperiksa untuk mengetahui apakah terdapat nilai-nilai yang tidak masuk akal pada kolom-kolom seperti: Glucose (tidak boleh bernilai nol), BloodPressure (tekanan darah tidak mungkin nol), SkinThickness (ketebalan kulit triceps tidak mungkin nol), Insulin (insulin serum juga tidak realistis bernilai nol), BMI (indeks massa tubuh tidak mungkin nol). Dengan menggunakan pendekatan statistik dan visualisasi data, jumlah data dengan nilai nol di kolom-kolom tersebut diidentifikasi. Identifikasi ini penting untuk memastikan bahwa model tidak belajar dari data yang tidak akurat atau menyesatkan, yang dapat mengarah pada prediksi yang salah. Misalnya, nilai nol pada parameter medis seperti glukosa atau tekanan darah secara klinis tidak mungkin, mengindikasikan bahwa itu adalah missing value yang dicatat sebagai nol.

b. Penanganan Nilai Nol atau Missing Values

Beberapa metode digunakan untuk menangani nilai nol atau missing values yaitu dengan penghapusan baris data jika jumlah data yang bermasalah sangat kecil dan tidak signifikan, serta imputasi nilai menggunakan rata-

DOI http://dx.doi.org/10.36448/expert.v15i1.4312
e-ISSN 2745-7265 p-ISSN 2088-5555 EXPERT Vol. 15 No. 1
June 30, 2025 – Hal. 4

rata (*mean*) atau median (*median)* dari masing-masing kolom yang bersangkutan. Dalam kasus ini, imputasi median dipilih karena distribusi data cenderung tidak normal dan memiliki outlier.

c. *Normalisasi dan Standarisasi Data*

Untuk meningkatkan performa model machine learning, dilakukan standarisasi data menggunakan metode StandardScaler , yaitu proses transformasi data agar memiliki rata-rata nol dan deviasi standar satu. Hal ini penting karena Random Forest sebenarnya tidak terlalu sensitif terhadap skala data. Meskipun Random Forest tidak terlalu sensitif terhadap skala data, standarisasi tetap bermanfaat untuk algoritma lain yang mungkin digunakan dalam perbandingan atau untuk memastikan konsistensi dalam pipeline machine learning secara keseluruhan. Proses ini juga membantu dalam mempercepat konvergensi algoritma jika digunakan bersama metode optimasi berbasis gradien.

d. *Encoding Variabel Kategorikal*

Dataset Pima Indians Diabetes hanya memiliki satu variabel target (Outcome) yang merupakan label biner (0 atau 1), sedangkan semua fitur lainnya bersifat numerik. Oleh karena itu, tidak ada variabel kategorikal yang perlu di-encode ulang.

3. Implementasi Pemodelan *Random Forest*

Langkah utama penelitian ini adalah mengimplementasikan model klasifikas menggunakan *Random Forest*. Metode *Random Forest* adalah model *ensemble* yang menggabungkan banyak *Decision Tree* untuk meningkatkan akurasi dan mengurangi *overfitting* [7]. *Random Forest* memerlukan kombinasi beberapa pohon keputusan untuk memprediksi hasil secara akurat. Saat menggunakan *random forest* sebagai pengklasifikasi, setiap pohon keputusan dapat menghasilkan jawaban yang sama atau berbeda. Misalnya pohon keputusan A, B, E dan F memprediksi hasil 1. Sedangkan pohon keputusan C dan D memprediksi hasil 0. Karena banyak alternatif jawaban di pohon keputusan dan probabilitasnya tinggi maka random forest mengambil hasil prediksi tersebut. Hasil dari beberapa pohon keputusan berdasarkan suara mayoritas dan prediksi hasil yang lebih akurat.

4. Pelatihan dan Uji Model

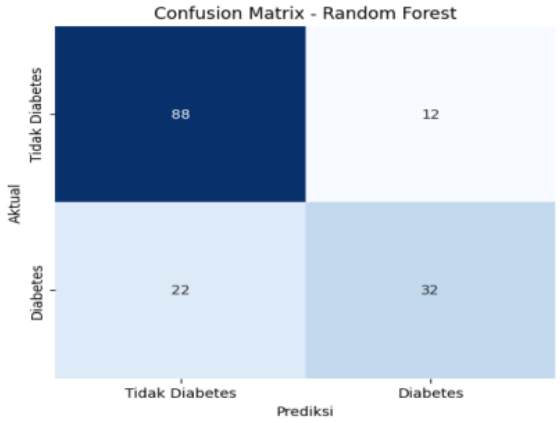
Setelah data diproses, dataset dibagi menjadi dua bagian Training set (80%) untuk melatih model Random Forest, dan Testing set (20%) untuk mengevaluasi performa model setelah dilatih. Pembagian data dilakukan secara acak dengan mempertahankan proporsi kelas (stratifikasi), sehingga distribusi antara kelas positif dan negatif tetap seimbang di kedua subset. Pembagian data yang stratifikasi sangat penting, terutama pada dataset yang tidak seimbang, untuk memastikan bahwa baik training set maupun testing set merepresentasikan distribusi kelas yang sebenarnya dalam populasi. Ini mencegah model menjadi bias terhadap kelas mayoritas dan memastikan evaluasi kinerja yang lebih realistis.

5. Hasil Evaluasi

Setelah melalui proses pra-pemrosesan data dan pelatihan model menggunakan algoritma *Random Forest*, tahap selanjutnya dalam penelitian ini adalah evaluasi kinerja model. Evaluasi dilakukan dengan mengukur kemampuan prediksi model terhadap dataset uji, serta divisualisasikan dalam bentuk grafis untuk mempermudah interpretasi hasil. Berikut ini adalah hasil-hasil yang diperoleh dari tahapan tersebut.

5.1. Visualisasi *Confusion Matrix*

Setelah model dilatih dan diuji pada data uji, performa model dikuantifikasi melalui matriks konfusi. Matriks ini memberikan gambaran tentang hasil prediksi model secara detail, termasuk jumlah prediksi benar positif (*True Positive*), salah negatif (*False Negative*), salah positif (*False Positive*), dan benar negatif (*True Negative*), disajikan pada Gambar 2



Gambar 2. *Confusion Matrix*

Berdasarkan Gambar 2 yang menampilkan matriks konfusi dalam bentuk heatmap, yang membantu dalam memvisualisasikan tingkat kesalahan dan keberhasilan prediksi model terhadap setiap kelas. Dari visualisasi ini, dapat dilihat bahwa model memiliki tingkat *True Positive Rate* (TPR) yang tinggi dan *False Positive Rate* (FPR) yang rendah, menunjukkan bahwa model memiliki kinerja yang baik dalam mengklasifikasikan kelas minoritas setelah penerapan. Model ini berhasil mengklasifikasikan 88 individu sebagai tidak kanker payudara dengan benar (True Negative) dan 32 individu sebagai kanker payudara dengan benar (True Positive). Kesalahan klasifikasi yang terjadi sangat kecil, yaitu hanya 12 kasus False Positive (individu yang sebenarnya tidak memiliki tetapi diprediksi sebagai deabetes) dan 22 kasus False Negative (individu yang sebenarnya memiliki deabetes tetapi diprediksi sebagai tidak memiliki deabetes).

Dalam penelitian sebelumnya menggunakan metode pembelajaran mesin dalam prediksi penyakit kanker payudara dengan pembagian data dan uji 80% : 20% didapatkan hasil seperti pada Tabel 2.

Tabel 2. Hasil Pengujian Pembagian Data Uji 80% : 20%

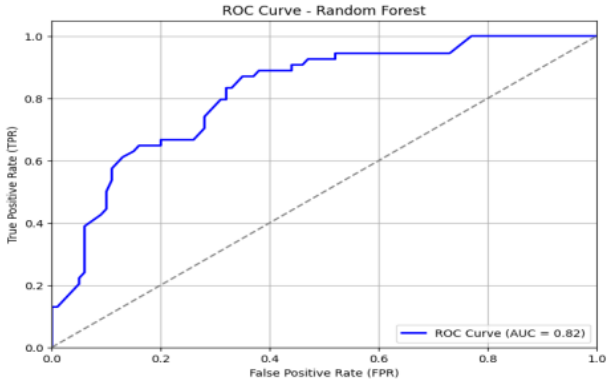
Confusion Matrix	*Accuracy*	*Precision*	*Recall*	*F1-Score*
Random Forest	0.78	0.76	0.74	0.75

DOI http://dx.doi.org/10.36448/expert.v15i1.4312
e-ISSN 2745-7265 p-ISSN 2088-5555 EXPERT Vol. 15 No. 1
June 30, 2025 – Hal. 5

Berdasarkan Tabel 1 hasil eksperimen dengan model *Random Forest* untuk memprediksi diabetes berdasarkan dataset kesehatan, hasil evaluasi menunjukkan bahwa model ini memiliki performa yang cukup baik dalam mengklasifikasikan pasien dengan dan tanpa diabetes. Model diuji menggunakan data uji yang telah dipisahkan sebelumnya dengan rasio 80 : 20. Hasil evaluasi kinerja model menunjukkan akurasi sebesar 0.78%, presisi 0.76%, recall 0.74%, F1-score 0.75%. Berdasarkan hasil yang didapat menurut penelitian [13] dengan nilai diatas 0maka hasil penelitian dikatakan baik. Selain itu, *confusion matrix* yang dihasilkan menunjukkan bahwa model mampu mendeteksi sebagian besar kasus diabetes dengan tingkat kesalahan yang rendah. Namun, terdapat beberapa kasus false positive dan false negative yang perlu diperhatikan lebih lanjut.

5.2. Visualisasi Kurve ROC-AUC

Untuk mengevaluasi kemampuan diskriminatif model secara keseluruhan, digunakan kurva ROC (*Receiver Operating Characteristic*) bersama dengan nilai AUC (*Area Under Curve*). Berdsarkan **Gambar 3** dibawah ini akan menampilkan kurva ROC yang memetakan hubungan antara True Positive Rate (TPR) dan False Positive Rate (FPR) pada berbagai threshold.



Gambar 3. *Receiver Operating Characteristic (ROC) Curve*

Berdasarkan Gambar 3 Nilai AUC yang diperoleh adalah sebesar **0.821**, menunjukkan bahwa model memiliki kemampuan prediksi yang baik. Semakin mendekati 1, semakin baik kemampuan model dalam membedakan antara kelas positif dan negatif.

5. KESIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan, dapat disimpulkan bahwa algoritma *Random Forest* menunjukkan kinerja yang baik dalam memprediksi risiko diabetes menggunakan dataset Pima Indians Diabetes. Dengan pendekatan *machine learning* dan penerapan pra-pemrosesan data yang meliputi normalisasi, penanganan nilai nol dan pembagian data latih-uji, model berhasil menghasilkan prediksi yang cukup akurat. Model *Random Forest* memberikan akurasi sebesar 78% pada data uji, dengan nilai AUC (Area Under Curve) sebesar 0.821, menunjukkan kemampuan diskriminasi antara pasien dengan dan tanpa diabetes yang tergolong sangat baik. Confusion matrix juga menunjukkan bahwa model mampu mendeteksi kasus positif diabetes dengan presisi dan recall yang seimbang, sehingga layak digunakan sebagai alat bantu dalam sistem pendukung keputusan di bidang medis. Dengan hasil yang diperoleh, dapat disimpulkan bahwa algoritma *Random Forest* merupakan salah satu model machine learning yang potensial untuk digunakan dalam prediksi penyakit kronis seperti diabetes, terutama jika dikombinasikan dengan teknik pra-pemrosesan dan evaluasi yang tepat. Hasil penelitian ini dapat menjadi dasar pengembangan sistem diagnosis berbasis AI yang cepat, objektif, dan dapat diandalkan dalam lingkungan medis.

DAFTAR PUSTAKA

[1] F. S. Nugraha, M. J. Shidiq, and S. Rahayu, "Analisis Algoritma Klasifikasi Neural Network Untuk Diagnosis Penyakit Kanker Payudara," *J. Pilar Nusa Mandiri*, vol. 15, no. 2, pp. 149–156, 2019, doi: 10.33480/pilar.v15i2.601.

[2] R. Maulana, M. F. Hasan, F. Raehan, and M. Ramzy, "Literature Review : Penerapan Algoritma Random Forest untuk Klasifikasi Penyakit Diabetes," *BIKMA*, vol. 2, no. 3, pp. 550–555, 2024.

[3] T. Zhang, Q. Wu, and Z. Zhang, "Probable Pangolin Origin of SARS-CoV-2 Associated with the COVID-19 Outbreak," *Curr. Biol.*, vol. 30, no. 7, pp. 1346-1351.e2, 2020, doi: 10.1016/j.cub.2020.03.022.

[4] S. Ramadhani and M. R. Wayahdi2, "K-Nearest Neighbor and Random Forest Algorithms in Loan Approval Prediction," *J. Minfo Polgan*, vol. 13, pp. 1307–1313, 2024.

[5] Ary Prandika Siregar, Dwi Priyadi Purba, Jojor Putri Pasaribu, and Khairul Reza Bakara, "Implementasi Algoritma Random Forest Dalam Klasifikasi Diagnosis Penyakit Stroke," *J. Penelit. Rumpun Ilmu Tek.*, vol. 2, no. 4, pp. 155–164, 2023, doi: 10.55606/juprit.v2i4.3039.

[6] I. Naive, B. Dalam, M. Penyakit, D. Mellitus, and D. Mining, "IMPLEMENTASI NAIVE BAYES DALAM MEMPREDIKSI PENYAKIT DIABETES MELLITUS Fakultas Teknologi Informasi , Universitas Hasyim Asy ' ari Tebuireng Jombang Abstrak," pp. 144–153.

[7] R. Irfannandhy, L. B. Handoko, and N. Ariyanto, "Analisis Performa Model Random Forest dan CatBoost dengan Teknik SMOTE dalam Prediksi Risiko Diabetes," *J. Pendidik. Inform.*, vol. 8, no. 2, pp. 714–723, 2024, doi: 10.29408/edumatic.v8i2.27990.

[8] H. Ma'rifah, A. P. Wibawa, and M. I. Akbar, "Klasifikasi Artikel Ilmiah Dengan Berbagai Skenario Preprocessing," *Sains, Apl. Komputasi dan Teknol. Inf.*, vol. 2, no. 2, p. 70, 2020, doi: 10.30872/jsakti.v2i2.2681.

[9] Sugiyono, *Buku Metode Penelitian*. In Metode Penelitian, 2020.

[10] Z. Abdussamad, *Metodologi Penelitian Kualitatif*. Makasar: CV. syakir Media Press, 2021.

[11] P. A. S. Banilai and M. Sakundarno, "SYSTEMATIC REVIEW: FAKTOR-FAKTOR YANG BERHUBUNGAN DENGAN KEJADIAN DIABETES MELITUS (DM) PADA PENDERITA TUBERKULOSIS (TB)," *Heal. Tadulako J. (Jurnal Kesehat. Tadulako)*, vol. 9, no. 2, pp. 205–217, May 2023, doi: 10.22487/htj.v9i2.739.

[12] Q. P. Irawan, K. D. Utami, S. Reski, and Saraheni, "Hubungan Indeks Massa Tubuh (IMT) dengan Kadar HbA1c pada Penderita Diabetes Mellitus Tipe II di Rumah Sakit Abdoel Wahab Sjahranie," *Formosa J. Sci. Technol.*, vol. 1, no. 5, pp. 459–468, Oct. 2022, doi: 10.55927/fjst.v1i5.1220.

[13] Suci Amaliah, M. Nusrang, and A. Aswi, "Penerapan Metode Random Forest Untuk Klasifikasi Varian Minuman Kopi di Kedai Kopi Konijiwa Bantaeng," *VARIANSI J. Stat. Its Appl. Teach. Res.*, vol. 4, no. 3, pp. 121–127, 2022, doi: 10.35580/variansiunm31.