Jurnal Pendidikan dan Teknologi Indonesia (JPTI)
Vol. 5, No. 7, Juli 2025, Hal. 1857-1866
DOI: https://doi.org/10.52436/1.jpti.871
p-ISSN: 2775-4227
e-ISSN: 2775-4219

Evaluasi Kinerja Model *Random Forest* Dalam Memprediksi Diabetes Berdasarkan Dataset Kesehatan di Indonesia

M. Rana Inzaghi[1], Erliyan Redy Susanto[*2], Amarudin[3], Neneng[4]

[1,2,3,4]Fakultas Teknik dan Ilmu Komputer, Universitas Teknokrat Indonesia, Indonesia
Email: [1]m_rana_inzaghi@teknokrat.ac.id, [2]erliyan.redy@teknokrat.ac.id, [3]amarudin@teknokrat.ac.id, [4]neneng@teknokrat.ac.id

Abstrak

Penyakit diabetes atau sering disebut dengan penyakit gula darah adalah sekelompok penyakit metabolik yang ditandai dengan tingginya kadar gula darah pada seseorang yang terkena, dan bertahan dalam jangka waktu lama. Di Indonesia sedikitnya terdapat 20 juta orang pada usia 20-79 tahun menderita diabetes pada tahun 2024. Hal ini disebabkan oleh kurangnya akses terhadap alat prediksi yang efektif, serta keterbatasan pada pendekatan tradisional bergantung pada diagnosis medis manual yang memakan waktu dan biaya. Permasalahan ini muncul karena kurangnya pemanfaatan teknologi berbasis data dalam menganalisis faktor risiko yang kompleks dan saling terkait. Penelitian ini bertujuan menggunakan model random forest untuk melakukan klasifikasi terhadap penyakit diabetes serta mengevaluasi nilai akurasi dengan evaluasi model menggunakan metrik seperti akurasi, presisi, recall, dan F1-score. Teknik akurasi yang digunakan yaitu confusion matrix untuk mengukur performa dalam permasalahan sehingga menghasilkan nilai akurasi yang sesuai. Hasil penelitian ini dapat memberikan wawasan praktis tentang konfigurasi optimal model untuk aplikasi dunia nyata, sehingga meningkatkan akurasi dan keandalan sistem prediksi diabetes. Model diuji menggunakan data uji yang telah dipisahkan sebelumnya dengan rasio 80:20. Hasil evaluasi kinerja model menunjukkan akurasi sebesar 0.99%, presisi 0.99%, recall 0.99%, F1-score 0.99%, Specificity 0.99% dan ROC-AUC Score 89.2%. Hasil penelitian bermanfaat untuk membantu dokter dan tenaga kesehatan serta masyarkat umum untuk mendeteksi penyakit diabetes sejak dini.

Kata kunci: *Diabetes, Evaluasi, Kerja, Random Forest.*

Evaluation of the Performance of the Random Forest Model in Predicting Diabetes Based on Health Datasets in Indonesia

Abstract

Diabetes or often referred to as blood sugar disease is a group of metabolic diseases characterized by high blood sugar levels in a person who is affected, and persists for a long time. In Indonesia, at least 20 million people aged 20-79 years suffer from diabetes in 2024. This is due to the lack of access to effective prediction tools, as well as limitations in traditional approaches that rely on manual medical diagnosis which is time-consuming and costly. This problem arises due to the lack of utilization of data-based technology in analyzing complex and interrelated risk factors. This study aims to use a random forest model to classify diabetes and evaluate the accuracy value by evaluating the model using metrics such as accuracy, precision, recall, and F1-score. The accuracy technique used is the confusion matrix to measure performance in the problem so as to produce an appropriate accuracy value. The results of this study can provide practical insights into the optimal configuration of the model for real-world applications, thereby increasing the accuracy and reliability of the diabetes prediction system. The model was tested using previously separated test data with a ratio of 80:20. The results of the model performance evaluation showed an accuracy of 0.99%, precision of 0.99%, recall of 0.99%, F1-score of 0.99%, Specificity of 0.99% and ROC-AUC Score of 89.2%. The results of the study are useful for helping doctors and health workers as well as the general public to detect diabetes early.

Keywords: *Diabetes, Evaluation, Random Forest, Work.*

1. PENDAHULUAN

Kesehatan masyarakat menjadi perhatian utama secara global karena prevalensi penyakit tidak menular yang terus meningkat, termasuk diabetes [1]. Penyakit diabetes atau sering disebut dengan penyakit gula darah

p-ISSN: 2775-4227
e-ISSN: 2775-4219

adalah sekelompok penyakit metabolik yang ditandai dengan tingginya kadar gula darah pada seseorang yang terkena, dan bertahan dalam jangka waktu lama [2]. Jumlah penderita diabetes diperkirakan akan terus meningkat jika pengetahuan masyarakat umum mengenai faktor pemicu penyakit diabetes tidak memadai [3].

Diabetes berdampak besar pada dunia kesehatan. Sebagai contoh, *International Diabetes Federation* (IDF) memperkirakan sedikitnya terdapat 20 juta orang pada usia 20-79 tahun di Indonesia menderita diabetes pada tahun 2024 [4]. Angka prevalensinya mencapai 9,3% dari total penduduk dengan usia yang sama. Berdasarkan dari jenis kelamin-nya, prevalensi diabetes di tahun 2024 mencapai 9% pada perempuan dan 9,65% pada laki-laki. Kasus di Indonesia, WHO (*World Health Organization*) berasumsi bahwa jumlah orang yang terkena DM di Indonesia dari tahun 2000 sejumlah 8,4 juta dan pada tahun 2030 diperkirakan akan meningkat hingga 21,3 juta jiwa[5].

Salah satu tantangan utama dalam pencegahan dan pengelolaan *diabetes* adalah kemampuan mendeteksi risiko secara dini sehingga langkah-langkah intervensi dapat dilakukan secara tepat. Diabetes sering kali tidak terdiagnosis pada tahap awal, yang menyebabkan penyakit ini berkembang secara bertahap hingga menyebabkan komplikasi yang parah [6]. Hal ini disebabkan oleh kurangnya akses terhadap alat prediksi yang efektif, serta keterbatasan pada pendekatan tradisional yang bergantung pada diagnosis medis manual yang memakan waktu dan biaya. Dengan semakin berkembangnya data kesehatan dalam jumlah besar, penting untuk memanfaatkan metode yang dapat mengolah data ini secara efektif guna menghasilkan model prediksi yang andal dan mudah diinterpretasikan [7].

Permasalahan ini muncul karena kurangnya pemanfaatan teknologi berbasis data dalam menganalisis faktor risiko yang kompleks dan saling terkait. Faktor-faktor seperti usia, indeks massa tubuh (BMI), tekanan darah, dan riwayat keluarga merupakan beberapa variabel yang sangat mempengaruhi risiko diabetes, namun sering kali sulit diprediksi dengan pendekatan konvensional [8]. Gejala penyakit diabetes sangat bervariasi pada setiap pasien, sehingga sulit dikenali. Menurut [9] 1 diantara 2 orang penyandang diabetes masih belum terdiagnosis dan belum menyadari bahwa dirinya diabetes, karena gejalanya mirip dengan kondisi sakit biasa, shingga banyak orang yang tidak menyadari bahwa mereka mengidap penyakit diabetes dan bahkan sudah mengarah pada komplikasi. Dilihat dari angka kematian yang tinggi yang diakibatkan oleh diabetes, prediksi begitu penting dilakukan untuk menekan angka kematian. Model prediksi dapat digunakan sebagai alat bantu bagi tenaga medis dan masayarakat awam untuk memperkirakan apakah seseorang mengidap diabetes atau tidak [10]. Salah satunya dengan evaluasi kerja model *Random Forest*.

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 [11]. 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.

Sebelumnya model *random forest* telah diteliti oleh [12] hasil analisis diperoleh bahwa model dengan error klasifikasi terkecil adalah dengan menggunakan mtry 2 dan ntree 500. Model yang dihasilkan dievaluasi dengan menggunakan confusion matrix dimana diperoleh bahwa varian minuman kategori *coffee based* lebih diminati daripada *signature coffee* dengan nilai akurasi sebesar 94,12%. Selanjutnya penelitian [13] Hasil penelitian menunjukkan bahwa XGBoost memiliki akurasi lebih tinggi (90,6%) dengan generalisasi yang baik sedangkan Random Forest lebih unggul dalam efisiensi waktu pelatihan. Analisis fitur mengidentifikasi faktor utama seperti Age, Blood Glucose Levels dan Weight Gain During Pregnancy yang mempengaruhi prediksi. Gap penelitian ini yaitu penelitian belum memberikan panduan model yang akurat dan transparan untuk mendukung pengambilan keputusan medis.

Berdasarkan latar belakang, penelitian ini bertujuan menggunakan model *random forest* untuk melakukan klasifikasi terhadap penyakit diabetes serta mengevaluasi nilai akurasi dengan evaluasi model menggunakan metrik seperti akurasi, presisi, *recall*, dan F1-score. Teknik *confusion matrix* akan digunakan untuk memastikan ketangguhan hasil akurasi, dan analisis pentingnya fitur akan dilakukan untuk meningkatkan interpretabilitas model. Tujuan penelitian ini adalah untuk membantu dokter dan tenaga *kesehatan* serta masyarkat umum untuk mendeteksi penyakit *diabetes* sejak dini. Hasil penelitian ini dapat memberikan wawasan praktis tentang konfigurasi optimal model untuk aplikasi dunia nyata, sehingga meningkatkan akurasi dan keandalan sistem prediksi diabetes. Selain itu, identifikasi faktor risiko utama melalui analisis pentingnya fitur berpotensi menginformasikan strategi pencegahan dan intervensi yang tepat sasaran, yang pada akhirnya berkontribusi pada hasil kesehatan yang lebih baik. Studi ini juga memajukan pemahaman teoretis tentang aplikasi *machine learning* di bidang kesehatan dengan mengatasi kesenjangan penelitian kritis terkait generalisasi dan interpretabilitas model.

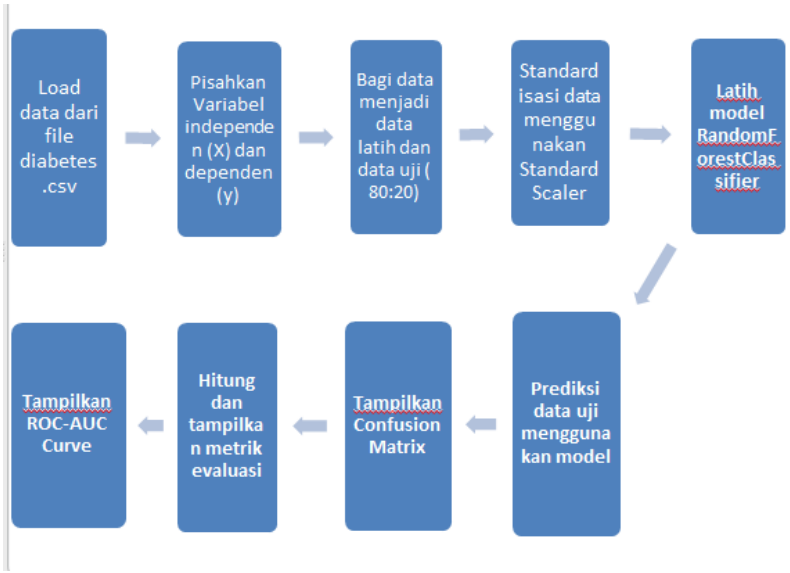
p-ISSN: 2775-4227
e-ISSN: 2775-4219

2. METODE

Penelitian ini disajikan dengan menggunakan metode studi kasus (*case study research*) dimana metode ini berhubungan dengan satu tujuan peneliti yang berfokus pada analisis penelitian [14][15][16]. Penulis juga**.** mengumpulkan data yang bersifat asli dan sudah terverifikasi serta menggunakan metode analisis dalam penelitian yaitu *random forest***.**

2.1. Tahap Studi

Berikut ini adalah tahapan dari metode *random forest* dapat dilihat pada Gambar 1



Gambar 1 Tahapan Metode *Random Forest*

2.2. Sumber Data

Berikut adalah penjelasan Gambar 1 mengenai tahapan metode *Random Forest* dengan menggunakan sumber data yang didapat:

- 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 [17][18]. 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.

- Pemisahan Variabel

Setelah data dikumpulkan dan dibersihkan, langkah selanjutnya adalah memisahkan variabel independen (X) dan variabel dependen (Y). Variabel independen mencakup semua fitur atau atribut yang digunakan sebagai input untuk memprediksi diabetes [19][20], seperti tingkat glukosa, tekanan darah, indeks massa tubuh (BMI), usia, dan faktor-faktor kesehatan lainnya. Sementara itu, variabel dependen (Y) merupakan label atau target yang menunjukkan status diabetes, biasanya dalam bentuk biner (0 untuk tidak diabetes dan 1 untuk diabetes). Pemisahan ini penting untuk memastikan bahwa model dapat mempelajari hubungan antara fitur-fitur tersebut dan hasil prediksi yang diinginkan [21]. Proses ini dilakukan dengan memisahkan kolom-kolom yang relevan dari dataset, di mana variabel independen digunakan untuk melatih model, dan variabel dependen digunakan sebagai acuan untuk mengukur akurasi prediksi model.

- Pembagian Data

Setelah variabel independen dan dependen dipisahkan, dataset kemudian dibagi menjadi dua subset, yaitu data latih (training data) dan data uji (testing data), dengan rasio 80:20. Pembagian ini dilakukan secara acak untuk memastikan bahwa kedua subset memiliki distribusi data yang representatif. Data latih, yang mencakup 80% dari total dataset, digunakan untuk melatih model Random Forest [22]. Sementara itu, data uji, yang

p-ISSN: 2775-4227
e-ISSN: 2775-4219

mencakup 20% dari dataset, digunakan untuk menguji kinerja model setelah proses pelatihan selesai. Pembagian data ini penting untuk memastikan bahwa model dapat digeneralisasi dengan baik dan tidak hanya bekerja optimal pada data latih tetapi juga mampu memberikan prediksi yang akurat pada data yang belum pernah dilihat sebelumnya. Dengan demikian, evaluasi model dapat dilakukan secara lebih objektif dan mendekati kondisi nyata.

- Standardisasi Data

Setelah data dibagi menjadi data latih dan data uji, langkah selanjutnya adalah melakukan standardisasi data menggunakan ***Standard Scaler***. Standardisasi adalah proses transformasi data sehingga memiliki mean (rata-rata) nol dan deviasi standar satu. Proses ini penting karena fitur-fitur dalam dataset sering kali memiliki skala yang berbeda-beda, misalnya, tingkat glukosa mungkin memiliki rentang nilai yang jauh lebih besar dibandingkan dengan tekanan darah. Perbedaan skala ini dapat memengaruhi performa model, terutama pada algoritma yang sensitif terhadap perbedaan magnitude, seperti Random Forest. Dengan melakukan standardisasi, semua fitur akan berada pada skala yang seragam, sehingga model dapat memproses data dengan lebih efisien dan mengurangi bias yang mungkin timbul akibat perbedaan skala. Standard Scaler digunakan karena kemampuannya dalam mengubah distribusi data tanpa mengubah struktur aslinya, sehingga informasi penting dalam data tetap terjaga. Proses ini diterapkan pada data latih, dan parameter yang sama kemudian digunakan untuk mentransformasi data uji, memastikan konsistensi dalam proses pelatihan dan pengujian model.

- Pelatihan Model

Setelah data distandardisasi, model ***Random Forest Classifier*** [23] [24] dilatih menggunakan data latih. Random Forest dipilih sebagai algoritma utama dalam penelitian ini karena kemampuannya yang kuat dalam menangani data yang kompleks dan non-linear, serta kemampuannya untuk mengurangi risiko overfitting. Random Forest bekerja dengan membangun banyak pohon keputusan (decision trees) selama proses pelatihan dan menggabungkan hasil prediksi dari semua pohon tersebut untuk menghasilkan prediksi akhir yang lebih akurat dan stabil. Selama pelatihan, model mempelajari pola dan hubungan antara variabel independen (fitur) dan variabel dependen (label diabetes). Hyperparameter dari model, seperti jumlah pohon (n_estimators) dan kedalaman maksimum pohon (max_depth), dapat diatur untuk mengoptimalkan performa model. Proses pelatihan ini bertujuan untuk membangun model yang mampu menggeneralisasi dengan baik, sehingga dapat memberikan prediksi yang akurat pada data yang belum pernah dilihat sebelumnya. Setelah model selesai dilatih, model tersebut siap untuk diuji menggunakan data uji guna mengevaluasi kinerjanya.

- Evaluasi Model

Setelah model *Random Forest* selesai dilatih, langkah selanjutnya adalah mengevaluasi kinerjanya menggunakan data uji. Evaluasi ini dilakukan untuk mengukur seberapa baik model dapat memprediksi status diabetes pada data yang belum pernah dilihat sebelumnya. Beberapa metrik evaluasi yang digunakan meliputi **akurasi**, **presisi**, **recall**, dan **F1-score**. Akurasi mengukur persentase prediksi yang benar secara keseluruhan, sementara presisi menilai proporsi prediksi positif yang benar di antara semua prediksi positif. Recall mengukur kemampuan model dalam mengidentifikasi semua kasus positif, dan F1-score merupakan harmonisasi antara presisi dan recall, memberikan gambaran yang seimbang tentang performa model. Selain itu, **ROC-AUC Curve** (Receiver Operating Characteristic - Area Under Curve) juga digunakan untuk mengevaluasi kemampuan model dalam membedakan antara kelas positif (diabetes) dan kelas negatif (tidak diabetes).

  - **ROC-AUC Curve**: Kurva ROC-AUC ditampilkan untuk mengevaluasi kemampuan model dalam membedakan antara kelas positif dan negative [25].
  - **Metrik Evaluasi**: Beberapa metrik evaluasi seperti akurasi, presisi, recall, dan F1-score dihitung dan ditampilkan untuk memberikan gambaran menyeluruh tentang kinerja model. Alasan memilih matrix ini adalah metrik ini menyediakan ukuran objektif untuk mengukur dan membandingkan kualitas, akurasi
  - **Confusion Matrix**: Confusion matrix ditampilkan untuk memberikan detail tentang jumlah prediksi benar dan salah untuk setiap kelas [26] (diabetes dan non-diabetes).

- Prediksi Data Uji

Setelah model dievaluasi menggunakan metrik-metrik yang relevan, langkah terakhir adalah menggunakan model yang telah dilatih untuk memprediksi data uji. Data uji, yang sebelumnya tidak digunakan selama proses pelatihan, berfungsi sebagai simulasi dari data baru yang akan dihadapi model dalam kondisi nyata. Dengan menerapkan model Random Forest pada data uji, prediksi status diabetes (positif atau negatif) dihasilkan untuk setiap entri dalam dataset. Hasil prediksi ini kemudian dibandingkan dengan label asli dari data uji untuk mengukur akurasi dan keandalan model. Proses ini tidak hanya menguji kemampuan model dalam menggeneralisasi data, tetapi juga memberikan gambaran tentang bagaimana model akan berperforma ketika diterapkan pada data riil di masa depan.

p-ISSN: 2775-4227
e-ISSN: 2775-4219

2.3. Variabel

Dalam penelitian ini, model *Random Forest* dievaluasi untuk menilai kemampuannya dalam memprediksi diabetes berdasarkan dataset kesehatan. Dataset yang digunakan terdiri dari beberapa variabel yang berhubungan dengan faktor risiko diabetes, seperti jumlah kehamilan, kadar glukosa dalam darah, tekanan darah, indeks massa tubuh (BMI), dan faktor keturunan.

Tabel 1. Variabel menampilkan fitur-fitur utama yang digunakan dalam analisis ini.

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)

2.4. *Random Forest*

Random Forest adalah model *ensemble* yang menggabungkan banyak ***Decision Tree*** untuk meningkatkan akurasi dan mengurangi *overfitting* [27].

2.4.1. *Decision Tree* dalam *Random Forest*

Setiap pohon keputusan dalam Random Forest mengikuti aturan pemisahan berdasarkan **entropy** atau **Gini Index [28]**.

1. Entropy (Information Gain)
   Entropy mengukur impurity (ketidakteraturan) dalam dataset, dengan rumus:

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

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:

$$IG(S, A) = H(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} H(S_v) \qquad (2)$$

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.

2. Gini Index
   Alternatif lain untuk memilih fitur adalah $Gini\ Impurity$, yang didefinisikan sebagai:

$$Gini(S) = 1 - \sum_{i-1}^{c} p_i^2 \qquad (3)$$

Semakin rendah nilai **Gini**, semakin "murni" data dalam node.

2.4.2. *Voting* dalam *Random Forest*

p-ISSN: 2775-4227
e-ISSN: 2775-4219

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

$$\hat{y} = mode\{T_1(X), T_2(X), \dots, T_N(X)\} \quad (4)$$

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

2.5. *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 [29].

$$\begin{bmatrix} \text{TP} & \text{FN} \\ \text{FP} & \text{TN} \end{bmatrix} \quad (5)$$

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:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (6)$$

b. *Precision*
Menunjukkan seberapa banyak prediksi positif yang benar:

$$Precision - \frac{TP}{TP+FP} \quad (7)$$

c. *Recall* (*Sensitivity*)
Menunjukkan seberapa banyak kasus **positif sebenarnya** yang dapat dideteksi:

$$\text{Recall} = \frac{TP}{TP+FN} \quad (8)$$

d. *Specificity*
Mengukur seberapa banyak kelas **negatif** yang diklasifikasikan dengan benar:

$$\text{Specificity} = \frac{TN}{TN+FP} \quad (9)$$

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

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (10)$$

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.

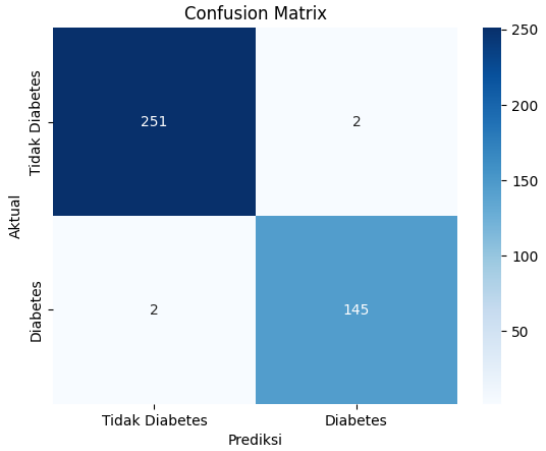
3. HASIL DAN PEMBAHASAN

Dalam studi ini kami menggunakan metode *Random Forest*.

p-ISSN: 2775-4227
e-ISSN: 2775-4219

3.1. Metode *Random Forest*

Berikut ini optimasi model *Random Forest* untuk prediksi penyakit diabet. *Confusion Matrix* algoritma Random Forest disajikan pada Gambar 2



Gambar 2 *Confusion Matrix Random Forest Evaluasi* Kinerja Model *Random Forest* Dalam Memprediksi Diabetes Berdasarkan Dataset Kesehatan

Berdasarkan Gambar 2 *confusion matrix* yang ditampilkan, model Random Forest menunjukkan kinerja yang sangat baik dalam memprediksi diabetes. Model ini berhasil mengklasifikasikan 251 individu sebagai tidak diabetes dengan benar (True Negative) dan 145 individu sebagai diabetes dengan benar (True Positive). Kesalahan klasifikasi yang terjadi sangat kecil, yaitu hanya 2 kasus False Positive (individu yang sebenarnya tidak memiliki diabetes tetapi diprediksi sebagai diabetes) dan 2 kasus False Negative (individu yang sebenarnya memiliki diabetes tetapi diprediksi sebagai tidak memiliki diabetes). Dengan hasil ini, model memiliki akurasi sebesar 99%, menunjukkan bahwa model dapat mengklasifikasikan sebagian besar sampel dengan benar. Selain itu, presisi dan recall untuk kelas diabetes sama-sama mencapai sekitar 98.64%, yang menandakan bahwa model tidak hanya mampu mendeteksi hampir semua kasus diabetes dengan baik tetapi juga jarang memberikan prediksi positif yang salah. Spesifisitas model juga sangat tinggi, sekitar 99.21%, yang berarti model sangat efektif dalam mengenali individu yang benar-benar tidak memiliki diabetes. Skor F1 sebesar 0.99 menunjukkan keseimbangan yang optimal antara presisi dan recall.

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

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

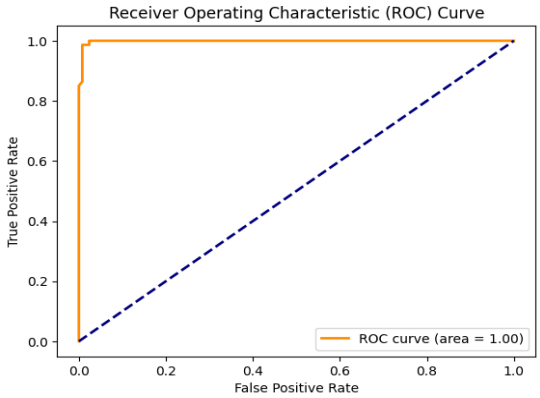
Confusion Matrix	***Accuracy***	***Precision***	***Recall***	***Specificity***	***F1-Score***
Random Forest	0.99	0.99	0.99	0.99	0.99

Berdasarkan Tabel 2 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 0.99%, presisi 0.99%, recall 0.99%, F1-score 0.99%, Specificity 0.99% dan ROC-AUC Score 89.2%. Berdasarkan hasil yang didapat menurut penelitian [12] dengan nilai diatas 0.90% dan dibawah nilai 100 % maka 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.

Hasil eksperimen menunjukkan bahwa model Random Forest mampu memberikan prediksi yang cukup akurat untuk diagnosis diabetes. Beberapa faktor utama yang mempengaruhi kinerja model ini antara lain pemilihan fitur, hyperparameter tuning, ketidakseimbangan data, perbandingan dengan model lain, dan interpretabilitas model. Analisis pentingnya fitur menunjukkan bahwa kadar glukosa dalam darah, BMI, dan riwayat keluarga (*Diabetes Pedigree Function*) adalah faktor-faktor yang paling berpengaruh dalam prediksi diabetes. Fitur-fitur ini memiliki kontribusi yang signifikan dalam membedakan pasien yang menderita diabetes

p-ISSN: 2775-4227
e-ISSN: 2775-4219

dan yang tidak. Selain itu, model mengalami peningkatan akurasi setelah dilakukan tuning terhadap parameter jumlah pohon keputusan (n_estimators) dan kedalaman maksimum pohon (max_depth), menunjukkan bahwa pemilihan parameter yang optimal sangat penting untuk meningkatkan kinerja model.



Gambar 3 *Receiver Operating Characteristic (ROC) Curve* Kinerja Model *Random Forest*

Gambar 3 menunjukkan ROC curve yang digunakan untuk mengevaluasi kinerja model Random Forest dalam memprediksi diabetes berdasarkan dataset kesehatan. ROC Curve menggambarkan hubungan antara True Positive Rate (TPR) dan *False Positive Rate* (FPR) pada berbagai threshold klasifikasi. Nilai TPR yang mencapai 0.8 hingga 1.0 menunjukkan bahwa model memiliki kemampuan yang sangat baik dalam mengidentifikasi kasus positif (pasien diabetes) dengan akurasi tinggi. Area Under Curve (AUC) sebesar 1.00 mengindikasikan bahwa model Random Forest ini memiliki performa sempurna dalam membedakan antara pasien diabetes dan non-diabetes. Berdasarkan ROC dan AUC Curva sama-sama menghasilkan nilai sebesar 1.00 sehingga hasil ini membuktikan bahwa model tidak hanya akurat tetapi juga sangat andal dalam klasifikasi, tanpa menghasilkan false positive yang signifikan. Dengan demikian, model Random Forest ini dapat diandalkan untuk aplikasi klinis atau penelitian lebih lanjut terkait prediksi diabetes.

4. DISKUSI

Dalam Studi ini, kami melakukan analisis prediksi diabetes menggunakan dataset diabetes.csv dengan memanfaatkan algoritma Random Forest. Dataset tersebut terdiri dari beberapa variabel independen seperti kadar glukosa, tekanan darah, indeks massa tubuh (BMI), dan lainnya, serta variabel dependen (Outcome) yang menunjukkan apakah seseorang menderita diabetes (1) atau tidak (0). Langkah pertama yang dilakukan adalah memisahkan variabel independen (X) dan dependen (y), kemudian membagi data menjadi data latih dan data uji dengan proporsi 80:20. Hal ini dilakukan untuk memastikan model dapat dievaluasi secara objektif pada data yang belum pernah dilihat sebelumnya.

Setelah pemisahan data, dilakukan standardisasi menggunakan StandardScaler untuk menormalisasi nilai-nilai fitur agar memiliki mean 0 dan standar deviasi 1. Proses ini penting untuk memastikan bahwa semua fitur memiliki skala yang sama, sehingga tidak ada fitur yang mendominasi proses pelatihan model. Selanjutnya, model Random Forest dilatih menggunakan data latih. Random Forest dipilih karena kemampuannya dalam menangani data yang kompleks dan mengurangi risiko overfitting berkat mekanisme bagging yang digunakan.

Setelah model dilatih, dilakukan prediksi pada data uji. Hasil prediksi kemudian dievaluasi menggunakan beberapa metrik, seperti accuracy, precision, recall, specificity, dan F1-score. Accuracy menunjukkan seberapa akurat model dalam memprediksi kedua kelas, sementara precision mengukur proporsi prediksi positif yang benar. Recall (atau sensitivity) mengukur seberapa baik model mengidentifikasi kasus positif, sedangkan specificity mengukur kemampuan model dalam mengidentifikasi kasus negatif. F1-score merupakan harmonisasi dari precision dan recall, yang berguna ketika kelas target tidak seimbang.

Selain metrik-metrik tersebut, kami juga menampilkan *Confusion Matrix* untuk memvisualisasikan performa model dalam memprediksi kelas positif dan negatif. Visualisasi ini membantu dalam memahami distribusi kesalahan prediksi, seperti false positive dan false negative. Selain itu, kami juga menggambar kurva ROC (*Receiver Operating Characteristic*) dan menghitung nilai AUC (*Area Under Curve*) untuk mengevaluasi kemampuan model dalam membedakan antara kedua kelas. Kurva ROC yang mendekati sudut kiri atas dan nilai AUC yang tinggi menunjukkan performa model yang baik.

Hasil analisis menunjukkan bahwa model *Random Forest* mampu memprediksi diabetes dengan cukup baik, meskipun masih ada ruang untuk perbaikan, terutama dalam meningkatkan recall untuk mengidentifikasi

p-ISSN: 2775-4227
e-ISSN: 2775-4219

lebih banyak kasus positif. Artikel ini menyoroti pentingnya pemilihan model yang tepat, evaluasi yang komprehensif, dan interpretasi hasil yang mendalam untuk memastikan bahwa model prediktif dapat digunakan secara efektif dalam konteks dunia nyata.

5. KESIMPULAN

Berdasarkan uraian sebelumnya maka penulis dapat memberikan kesimpulan yaitu hasil analisis menunjukkan bahwa model *Random Forest* mampu memprediksi diabetes dengan memajukan kinerja secara optimal, meskipun masih ada ruang untuk perbaikan, terutama dalam meningkatkan recall untuk mengidentifikasi lebih banyak kasus positif. Hasil penelitian menunjukkan bahwa model Random Forest mampu memberikan prediksi yang sangat akurat, dengan akurasi, presisi, recall, dan F1-score yang mencapai 99%. Secara keseluruhan, model Random Forest memiliki potensi yang besar dalam prediksi diabetes, terutama karena kemampuannya dalam menangani data non-linear dan memberikan interpretabilitas yang baik melalui analisis pentingnya fitur. Namun, penelitian lebih lanjut diperlukan untuk mengatasi tantangan seperti ketidakseimbangan data dan meningkatkan sensitivitas model terhadap kasus positif. Dengan demikian, model ini dapat menjadi alat yang efektif dalam mendukung diagnosis dini dan manajemen diabetes, serta memberikan wawasan berharga bagi tenaga medis dalam mengidentifikasi faktor risiko utama. Studi dapat memajukan pemahaman teoretis tentang aplikasi *machine learning* di bidang kesehatan dengan mengatasi kesenjangan penelitian kritis terkait generalisasi dan interpretabilitas model.

DAFTAR PUSTAKA

[1] R. Kusumastuti, "Prediksi Risiko Diabetes Menggunakan Algoritma Decision Tree Dengan Aplikasi Rapid Miner," No. November, Pp. 14–24, 2024. Doi: 10.9644/scp.v1i1.332.

[2] R. A. Siallagan And Fitriyani, "Prediksi Penyakit Diabetes Mellitus Menggunakan Algoritma C4.5," *J. Responsif Ris. Sains Dan Inform.*, Vol. 3, No. 1, Pp. 44–52, 2021, Doi: 10.51977/Jti.V3i1.407.

[3] S. U. Putri, "Implementasi Data Mining Untuk Prediksi Penyakit Diabetes Dengan Algoritma C4.5," *J. Penerapan Sist. Inf.*, Vol. 1, No. 1, Pp. 1–10, 2021.Doi: 10.32409/jikstik.23.1.3507

[4] A. M. E. R. I. C. A. N. D. I. A. B. E. T. E. S. A. S. S. O. C. I. A. T. I. O. N, "Standards Of Medical Care In Diabetes," 2024.

[5] R. F. N. Iskandar, D. H. Gutama, D. P. Wijaya, And D. Danianti, "Klasifikasi Menggunakan Metode Random Forest Untuk Awal Deteksi Diabetes Melitus Tipe 2," *J. Tek. Ind. Terintegrasi*, Vol. 7, No. 3, Pp. 1620–1626, 2024, Doi: 10.31004/Jutin.V7i3.26916.

[6] F. Johnson And L. White, "Application Of Data Mining In Health Sector: A Review Of Literature," *Int. J. Adv. Comput. Sci. Appl.*, Vol. 11, No. 5, Pp. 175–183, 2020.

[7] C. H. Lee And M. W. Yoon, "Trends In The Epidemiology Of Diabetes In Asia-Pacific Regions," *Curr. Diab. Rep.*, Vol. 16, No. 5, Pp. 465–473, 2020.Doi: 10.1177/1010539516663938

[8] E. Safitri, D. Rofianto, N. Purwati, H. Kurniawan, And S. Karnila, "Prediksi Penyakit Diabetes Melitus Menggunakan Algoritma Machine Learning Diabetes Mellitus Disease Prediction Using Machine Learning Algorithms," Vol. 12, No. 4, Pp. 760–766, 2024, Doi: 10.26418/Justin.V12i4.84620.

[9] Kemankes, *Kemankes*. 2024.

[10] M. Salsabil, N. Lutvi, And A. Eviyanti, "Implementasi Data Mining Dalam Melakukan Prediksi Penyakit Diabetes Menggunakan Metode Random Forest Dan Xgboost," *J. Ilm. Komputasi*, Vol. 23, No. 1, Pp. 51–58, 2024, Doi: 10.32409/Jikstik.23.1.3507.

[11] 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.

[12] 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.

[13] Mubaraqah, A. N. Puteri, And A. Sumardin, "Comparison Of Random Forest And Xgboost For Diabetes Classification With Shap And Lime Interpretation," *Jtera*, Vol. 9, No. 2, Pp. 121–130, 2025, Doi: 10.31544/Jtera.V9.I1.2024.121-130.

[14] A. D. Wahyudi And A. R. Isnain, "Penerapan Metode Topsis Untuk Pemilihan Distributor Terbaik," *Jati*, Vol. 1, No. 2, Pp. 59–70, 2023. Doi :org/10.34010/jati.v15i1

p-ISSN: 2775-4227
e-ISSN: 2775-4219

[15] Z. Abdussamad, *Metodologi Penelitian Kualitatif*. Makasar: Cv. Syakir Media Press, 2021.

[16] Hardani, *Metodologi Penelitian*. Jakarta: Cv. Pustaka Ilmu, 2020.

[17] 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.

[18] 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.

[19] D. Setyawan And A. Suradi, "Implementasi Web Service Dan Analisis Kinerja Algoritma Klasifikasi Data Mining Untuk Memprediksi Diabetes Mellitus," *Simetris J. Tek. Mesin, Elektro Dan Ilmu Komput.*, Vol. 8, No. 2, P. 701, Nov. 2017, Doi: 10.24176/Simet.V8i2.1584.

[20] C. Fiarni, E. M. Sipayung, And S. Maemunah, "Analysis And Prediction Of Diabetes Complication Disease Using Data Mining Algorithm," *Procedia Comput. Sci.*, Vol. 161, Pp. 449–457, 2019, Doi: 10.1016/J.Procs.2019.11.144.

[21] A. Viloria, Y. Herazo-Beltran, D. Cabrera, And O. B. Pineda, "Diabetes Diagnostic Prediction Using Vector Support Machines," *Procedia Comput. Sci.*, Vol. 170, Pp. 376–381, 2020, Doi: 10.1016/J.Procs.2020.03.065.

[22] W. Apriliah, I. Kurniawan, M. Baydhowi, And T. Haryati, "Prediksi Kemungkinan Diabetes Pada Tahap Awal Menggunakan Algoritma Klasifikasi Random Forest," *Sistemasi*, Vol. 10, No. 1, P. 163, Jan. 2021, Doi: 10.32520/Stmsi.V10i1.1129.

[23] P. Palimkar, R. N. Shaw, And A. Ghosh, "Machine Learning Technique To Prognosis Diabetes Disease: Random Forest Classifier Approach," 2022, Pp. 219–244. Doi: 10.1007/978-981-16-2164-2_19.

[24] Y. Ginanjar, I. Damayanti, And I. Permana, "Pengaruh Senam Diabetes Terhadap Penurunan Kadar Gula Darah Pada Penderita Diabetes Mellitus Di Wilayah Kerja Pkm Ciamis Kabupaten Ciamis Tahun 2021," *J. Keperawatan Galuh*, Vol. 4, No. 1, P. 19, Mar. 2022, Doi: 10.25157/Jkg.V4i1.6408.

[25] A. Pramudyantoro, E. Utami, And D. Ariatmanto, "Penggabungan K-Nearest Neighbors Dan Lightgbm Untuk Prediksi Diabetes Pada Dataset Pima Indians: Menggunakan Pendekatan Exploratory Data Analysis," *Jipi (Jurnal Ilm. Penelit. Dan Pembelajaran Inform.*, Vol. 9, No. 3, Pp. 1133–1144, Aug. 2024, Doi: 10.29100/Jipi.V9i3.4966.

[26] X. Wang *Et Al.*, "Exploratory Study On Classification Of Diabetes Mellitus Through A Combined Random Forest Classifier," *Bmc Med. Inform. Decis. Mak.*, Vol. 21, No. 1, P. 105, Dec. 2021, Doi: 10.1186/S12911-021-01471-4.

[27] 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.

[28] R. Maulana, M. F. Hasan, F. Raehan, And M. Ramzy, "Literature Review : Penerapan Algoritma Random Forest Untuk Klasifikasi Penyakit Diabetes," *Bima*, Vol. 2, No. 3, Pp. 550–555, 2024. DOI: 10.37638/bima.5.1.43-50

[29] 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.

[30] 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.