SAINSTECH
INNOVATION JOURNAL

SIJ, Volume 8, Nomor 1, Mei 2025, Hal. 507-518

COMPARATIVE ANALYSIS OF DECISION TREE AND RANDOM FOREST ALGORITHMS FOR PREDICTING DIABETES MELLITUS

Nindri Lia Desmita[1)], Syahrani Lonang[1*)], Danang Tejo Kumoro[1)]

[1)] Program Studi Teknologi Informasi, Fakultas Sains dan Teknologi, Universitas Qamarul Huda Badaruddin, Indonesia

**Corresponding Author: lonang@uniqhba.ac.id , Tel: +6287808406977*

Diterima pada *2 Februari 2025*, Direvisi pertama pada *15 Maret 2025*, Disetujui pada *22 April 2025*, Diterbitkan daring pada *20 Mei 2025*

Abstract: *Diabetes mellitus (DM) is a chronic disease with an increasing number of sufferers and a risk of serious complications. Early detection is very important to prevent these risks. This study uses a public dataset from Kaggle to compare the performance of Decision Tree and Random Forest algorithms in predicting diabetes status. The dataset includes demographic and medical information such as age, hypertension, cardiac history, BMI, HbA1c, and blood glucose levels. Unbalanced data was handled using the SMOTE method, and then tested with 80:20, 70:30, and 60:40 data sharing schemes. The evaluation results showed that Random Forest excelled in all schemes, with the best performance in the 60:40 scheme (96.02% accuracy, 76.13% F1-score). This research shows that Random Forest is effective to support machine learning-based diabetes early detection system.*

Keywords: *Decision Tree, Diabetes, Classification, Random Forest, SMOTE*

Abstrak: Diabetes Mellitus (DM) merupakan penyakit kronis yang jumlah penderitanya terus meningkat dan berisiko menimbulkan komplikasi serius. Deteksi dini sangat penting untuk mencegah risiko tersebut. Penelitian ini menggunakan dataset publik dari Kaggle untuk membandingkan performa algoritma Decision Tree dan Random Forest dalam memprediksi status diabetes. Dataset mencakup informasi demografis dan medis seperti usia, hipertensi, riwayat jantung, BMI, HbA1c, dan kadar glukosa darah. Data yang tidak seimbang ditangani menggunakan metode SMOTE, lalu diuji dengan skema pembagian data 80:20, 70:30, dan 60:40. Hasil evaluasi menunjukkan bahwa Random Forest unggul dalam semua skema, dengan performa terbaik pada skema 60:40 (akurasi 96,02%, F1-score 76,13%). Penelitian ini menunjukkan bahwa Random Forest efektif untuk mendukung sistem deteksi dini diabetes berbasis machine learning.

Kata kunci: Decision Tree, Diabetes, Klasifikasi Random Forest, SMOTE

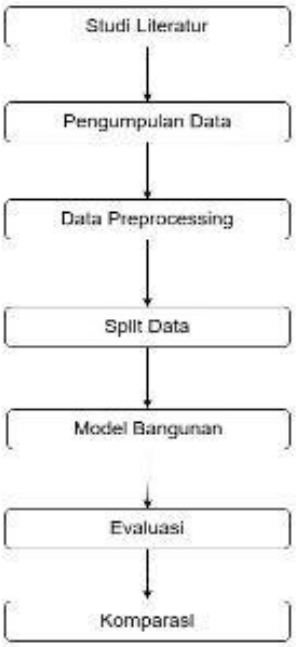
1. PENDAHULUAN

Diabetes Mellitus (DM) merupakan penyakit kronis yang ditandai oleh tingginya kadar gula dalam darah akibat gangguan produksi atau fungsi insulin [1]. Penyakit ini tergolong berbahaya karena dapat menimbulkan komplikasi jangka panjang yang mengancam jiwa, seperti gangguan kardiovaskular, kerusakan saraf, ginjal, mata, dan kulit, hingga gangguan psikologis seperti depresi[2][3]. Data International Diabetes Federation (IDF) tahun 2021 menunjukkan bahwa terdapat lebih dari 537 juta penderita diabetes di dunia, dengan 6,7 juta kematian. Indonesia sendiri menempati peringkat kelima dengan jumlah penderita mencapai 19,5 juta orang dan diperkirakan akan meningkat menjadi 28,8 juta pada tahun 2025 [4][5] . Fakta ini menegaskan bahwa DM merupakan ancaman serius secara global maupun nasional[6]z.

Langkah preventif, seperti deteksi dini, sangat penting untuk menekan risiko komplikasi dan meningkatkan efektivitas pengelolaan penyakit. Dalam hal ini, sistem berbasis kecerdasan buatan, khususnya machine learning (ML), berperan penting untuk menganalisis data besar (big data) dan mendukung proses pengambilan keputusan klinis [7]. Machine learning merupakan cabang kecerdasan buatan yang memungkinkan komputer belajar dari data dan membuat prediksi atau keputusan tanpa pemrograman eksplisit [8]. Pendekatan *supervised learning* dalam ML, seperti algoritma Decision Tree (DT) [9] dan Random Forest (RF) [10], telah banyak digunakan dalam klasifikasi medis. Decision Tree menawarkan struktur klasifikasi yang mudah diinterpretasikan, sedangkan Random Forest menggabungkan banyak pohon keputusan untuk menghasilkan prediksi yang lebih akurat dan tahan terhadap overfitting [11]. Kedua metode ini dipilih dalam penelitian ini untuk membandingkan kinerja dalam mengklasifikasikan dataset diabetes, baik dari segi akurasi prediksi maupun kemampuan menangani data dengan kompleksitas atribut yang tinggi.

2. METODE PENELITIAN

Penelitian ini menggunakan pendekatan eksperimen komparatif untuk menganalisis performa dua algoritma *supervised learning*, yaitu Decision Tree dan Random Forest, dalam melakukan klasifikasi pada dataset penderita diabetes. Penelitian dilakukan selama periode Januari hingga April 2025 dengan menggunakan data publik yang diperoleh dari situs Kaggle. Dataset yang digunakan merupakan data medis dan demografis pasien yang mencakup atribut seperti jenis kelamin, usia, tekanan darah, riwayat merokok, BMI, kadar HbA1c, kadar glukosa darah, dan status diabetes. Tujuan utama dari penelitian ini adalah membandingkan efektivitas kedua algoritma dalam melakukan prediksi terhadap status diabetes berdasarkan parameter yang tersedia. metode penelitian berisi cara, langkah, atau metode yang digunakan untuk menjawab masalah penelitian.



Gambar 1 Flowchart Penelitian

Tahapan dalam penelitian ini dapat dilihat pada Gambar 1, mencakup: studi literatur, pengumpulan data, preprocessing data, pembagian data menjadi data latih (*training*) dan data uji (*testing*), pembangunan model klasifikasi, evaluasi performa model, serta analisis komparatif hasil klasifikasi. Data yang diperoleh melalui repositori Kaggle dilakukan pembersihan terlebih dahulu dengan menghapus entri yang berpotensi menimbulkan bias, seperti kategori "Other" pada atribut *gender*. Selain itu, dilakukan normalisasi terhadap fitur numerik menggunakan metode *Min-Max Scaling* untuk memastikan kesetaraan skala antar fitur. Dataset yang telah diproses kemudian dibagi ke dalam tiga skenario proporsi data, yaitu 60:40, 70:30, dan 80:20, guna mengamati konsistensi performa model.Bab ini juga dapat berisi data yang digunakan, variabel penelitian, alat dan bahan, diagram alur atau flowchart serta lokasi penelitian. Model klasifikasi dibangun menggunakan algoritma Decision Tree dan Random Forest. Kedua model dilatih menggunakan data latih yang telah disiapkan pada masing-masing skenario, dan diuji menggunakan data uji untuk mengukur performa klasifikasinya. Evaluasi dilakukan menggunakan metrik akurasi, presisi, *recall*, spesifisitas, dan F1-score melalui *confusion matrix* [12]. Hasil dari evaluasi ini kemudian dianalisis untuk mengetahui metode mana yang lebih unggul dalam hal prediksi status diabetes. Penelitian ini diharapkan dapat memberikan kontribusi dalam pemanfaatan algoritma *machine learning* untuk diagnosis dini penyakit diabetes secara lebih akurat dan efisien.

3. HASIL DAN PEMBAHASAN

3.1 *Exploratory Data Analysis*

Eksplorasi awal dataset dilakukan untuk memperoleh pemahaman mengenai struktur data, karakteristik distribusi variabel, serta untuk mengidentifikasi adanya potensi ketidakseimbangan (imbalance) kelas target. Dataset yang digunakan berjumlah 100.000 data dari Kaggle Diabetes Prediction Dataset. Variabel target adalah diabetes yang berisi label 0 (tidak menderita diabetes) dan 1 (menderita diabetes).

Tabel 1. Dataset Penelitian

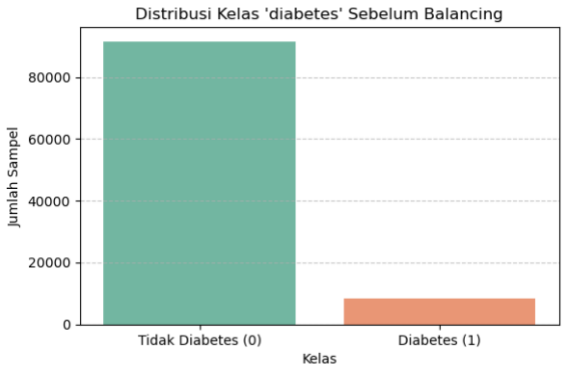
Variabel	Temuan Utama
Distribusi Kelas	Kelas tidak seimbang (mayoritas non-diabetes)
Gender	Wanita memiliki jumlah kasus diabetes lebih tinggi
Usia	Risiko meningkat secara signifikan pada usia >50 tahun
Hipertensi	Meningkatkan kemungkinan diabetes hingga dua kali lipat

Riwayat Jantung	Korelasi signifikan dengan kasus diabetes
Riwayat Merokok	Perokok aktif dan mantan perokok lebih berisiko
BMI	Rata-rata lebih tinggi pada penderita diabetes
HbA1c	Lebih tinggi pada penderita diabetes
Glukosa Darah	Signifikan lebih tinggi pada penderita diabetes

Pada Tabel 1 diatas terdapat hasil ringkasan setelah dilakukannya EDA atau *Exploratory data analysis*.

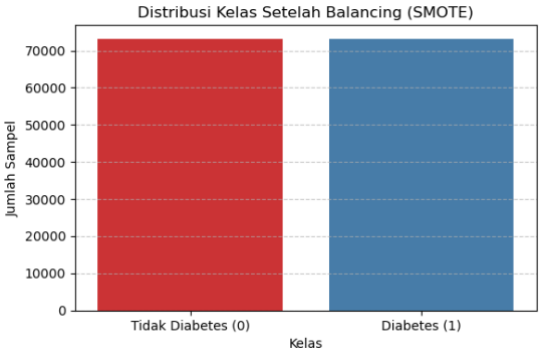
3.2 *Balancing Data*

Untuk mengatasi permasalahan kelas tidak seimbang, digunakan metode **SMOTE (Synthetic Minority Over-sampling Technique)** [13]. SMOTE menghasilkan data sintetis untuk kelas minoritas tanpa melakukan duplikasi langsung [14][15].



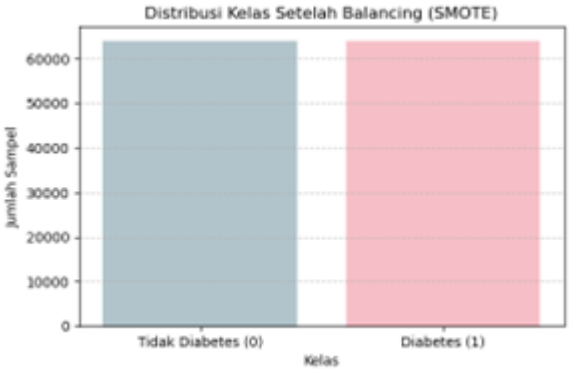
Gambar 2 Distribusi Kelas Diabetes Sebelum Balancing

Gambar 2 menunjukkan hasil distribusi data yang tidak seimbang atau *imbalance* class. Terlihat ketimpangan yang sangat signifikan antara dua kelas yang ada, yaitu tidak diabetes (0) dan diabetes (1). Berdasarkan hasil analisisnya, jumlah data untuk kelas tidak diabetes (label 0) adalah sekitar 91.580 sampel, sedangkan untuk kelas diabetes (label 1) hanya terdiri dari 8.403. Ketimpangan ini menunjukkan bahwa hanya sekitar 8,4% data yang merepresentasikan penderita diabetes.



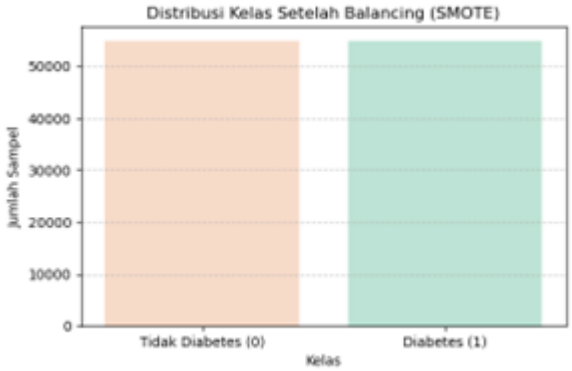
Gambar 3 Distribusi Kelas Setelah Balancing Skema 80:20

Visualiasasi pada gambar 3 merupakan distribusi kelas setelah dilakukannya balancing data dengan skema 80:20. Pada gambar tersebut terlihat bahwa distribusi kelas setelah proses balancing menggunakan metode SMOTE dengan skema 80:20 telah seimbang. Kedua kelas, yaitu tidak diabetes (label 0) dan diabetes (label 1), masing-masing memiliki jumlah data sebesar 73.186 sampel. Hal ini menunjukkan bahwa metode SMOTE berhasil mengatasi ketimpangan kelas yang sebelumnya terjadi, seperti ditunjukkan pada gambar 2. Tinggi batang yang sejajar pada visualisasi tersebut menandakan distribusi yang sudah merata antara kedua kelas.



Gambar 4 Distribusi Kelas setelah Balancing Skema 70:30

Berdasarkan gambar 4 dapat dilihat bahwa distribusi kelas diabetes setelah dilakukannya *balancing data* dengan menggunakan metode SMOTE skema 70:30, jumlah sampel pada kedua kelas tersebut menjadi seimbang, yaitu masing-masing sebanyak 64.038 sampel data. Ini menunjukkan bahwa metode SMOTE berhasil mengatasi ketimpangan pada jumlah data yang tidak seimbang. Hal ini ditunjukkan pada visualisasi hasil balancing yang memperlihatkan kedua kelas memiliki tinggi bar yang sama dalam diagram batang, menandakan tidak adanya lagi ketimpangan antar kelas.

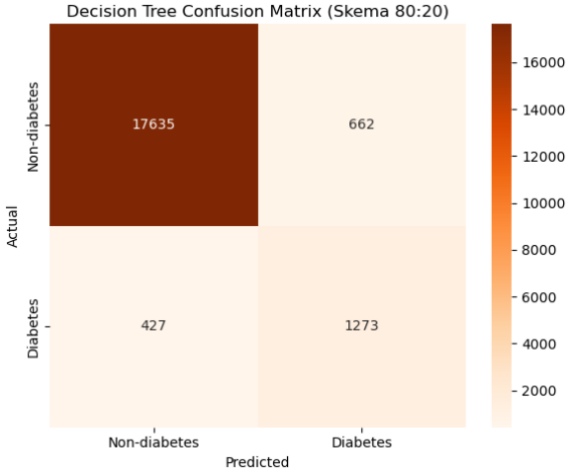


Gambar 5 Distribusi Kelas setelah Balancing Skema 60:40

Berdasarkan gambar 5 menunjukkan distribusi kelas setelah melakukan proses *balancing* menggunakan metode SMOTE skema 60:40 yang menghasilkan jumlah sampel data yang seimbangan antara kelas tidak diabetes dan kelas diabetes, masing-masing sebanyak 54.889 sampel.

3.3 Evaluasi Model: *Decision Tree*

Model dievaluasi menggunakan beberapa metrik untuk mengukur performa model. Metrik yang digunakan sebagai alat ukur adalah akurasi, presisi, recall, dan f1-score.

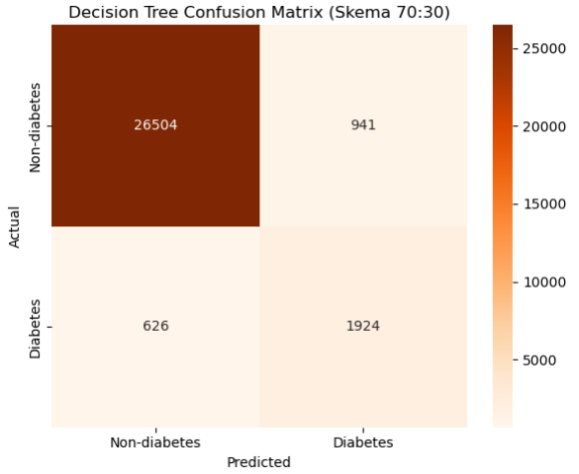


Gambar 6 *Decision Tree Confusion Matrix* Skema 80:20

Pada gambar 6 confusion matrix diatas menggambarkan hasil prediksi model Decision Tree terhadap data uji menggunakan skema 80:20. Berdasarkan confusion matrix tersebut, dapat diketahui hasil klasifikasi, model menghasilkan True Positive (TP) sebanyak 1.273 dan True Negative (TN) sebesar 17.635, yang menunjukkan bahwa model mampu mengklasifikasikan sebagian besar data dengan benar. Namun, masih terdapat False Positive (FP) sebanyak 662, di mana data yang seharusnya bukan penderita diabetes diklasifikasikan sebagai penderita, serta False Negative (FN) sebanyak 427, yaitu kasus penderita diabetes yang tidak berhasil terdeteksi. Nilai-nilai ini mencerminkan kinerja model dalam mengenali kedua kelas, baik yang positif maupun negatif.

Berdasarkan hasil confusion matrix tersebut dapat diketahui bahwa hasil dari model decision tree skema 80:20 yang digunakan dalam penelitian ini menunjukkan performa yang cukup baik, dengan akurasu tinggi yaitu 94,55. Namun, nilai presisi yang tidak terlalu tinggi yaitu 65,79% yang

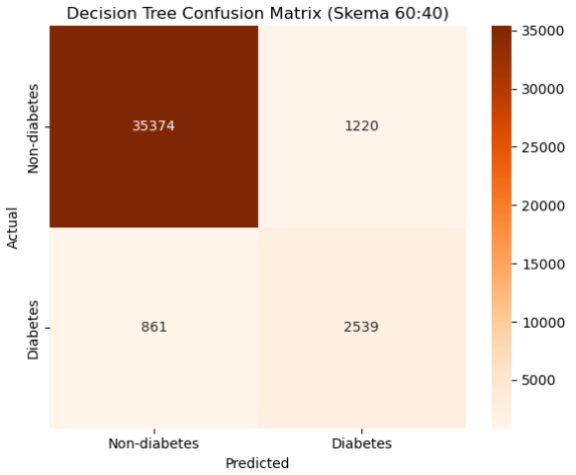
mengindikasikan adanya jumlah false positive yang cukup signifikan. Walaupun demikian, hasil recall yang cukup baik yaitu 74,88% menunjukkan bahwa model relatif handal dalam mendeteksi penderita diabetes. F1-Score sebesar 70.04% menegaskan bahwa model memiliki performa seimbang, meskipun masih ada ruang untuk peningkatannya.



Gambar 7 *Decision Tree Confusion Matrix* Skema 70:30

Gambar 7 confusion matrix dapat diketahui jumlah dari hasil klasifikasi menggunakan model decision tree skema 70:30 adalah model berhasil mengidentifikasi 1.924 kasus positif secara benar (True Positive) dan 26.504 kasus negatif dengan tepat (True Negative). Meski demikian, masih terdapat 941 kasus yang salah diklasifikasikan sebagai positif padahal seharusnya negatif (False Positive) serta 626 kasus positif yang tidak terdeteksi oleh model (False Negative). Hasil ini menunjukkan bahwa model memiliki kemampuan yang cukup baik dalam mengenali kedua kelas, meskipun masih terdapat kesalahan klasifikasi yang perlu diperhatikan dalam evaluasi performa secara menyeluruh.

Berdasarkan hasil evaluasi tersebut dapat disimpulkan bahwa model Decision Tree yang diuji dengan skema pembagian data 70:30 menunjukkan performa yang sangat baik secara keseluruhan, dengan akurasi mencapai 94,78%. Namun, presisi sebesar 67,16% menunjukkan bahwa terdapat sejumlah prediksi positif yang tidak akurat (false positive), yang berpotensi menyebabkan kesalahan diagnosis. Meskipun demikian, recall sebesar 75,45% menunjukkan kemampuan model dalam menangkap sebagian besar kasus diabetes yang sebenarnya. Kombinasi presisi dan recall tersebut menghasilkan nilai F1-score sebesar 71,06%, yang cukup memuaskan dan menunjukkan bahwa model bekerja seimbang antara mengidentifikasi kasus diabetes dan meminimalkan kesalahan.



Gambar 8 *Decision Tree Confusion Matrix* Skema 60:40

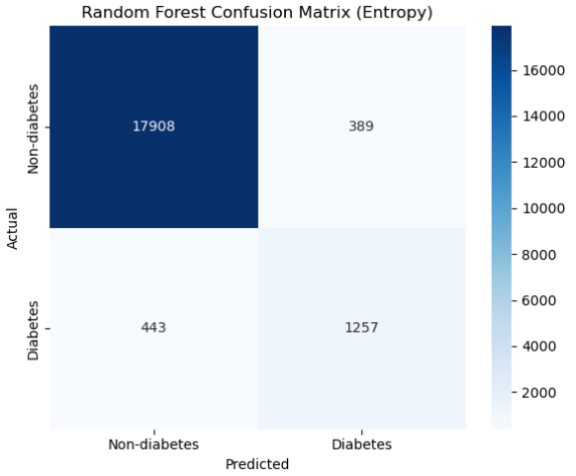
Gambar 8 confusion matrix diatas menggambarkan performa klasifikasi model terhadap data uji dengan skema 60:40 dengan hasil klasifikasi, model mencatat

2.539 kasus yang teridentifikasi dengan benar sebagai positif (True Positive) dan 35.374 kasus negatif yang juga berhasil diklasifikasikan dengan tepat (True Negative). Namun, terdapat 1.220 data yang seharusnya negatif tetapi salah diklasifikasikan sebagai positif (False Positive), serta 861 kasus positif yang tidak berhasil dikenali oleh model (False Negative). Hasil ini menunjukkan bahwa performa model cukup baik dalam mendeteksi kedua kelas, meskipun masih terdapat kesalahan klasifikasi yang perlu dievaluasi lebih lanjut.

Berdasarkan hasil evaluasi tersebut dapat disimpulkan bahwa model Decision Tree dengan skema 60:40 memberikan performa yang konsisten dengan skema lainnya. Akurasi tinggi sebesar 94,80% menunjukkan kinerja keseluruhan yang baik, meskipun nilai presisi yang lebih rendah (67,54%) mengindikasikan masih adanya prediksi positif yang salah. Dengan recall sebesar 74,68%, model cukup baik dalam mendeteksi penderita diabetes. Nilai F1-score 70,93% menguatkan bahwa model ini cukup seimbang dalam mendeteksi kasus positif dan menghindari kesalahan klasifikasi.

3.4 Evaluasi Model: *Random Forest (Entropy)*

Evaluasi model menggunakan beberapa metrik untuk mengukur performa model. Metrik yang digunakan sebagai alat ukur adalah akurasi, presisi, recall, dan f1-score.



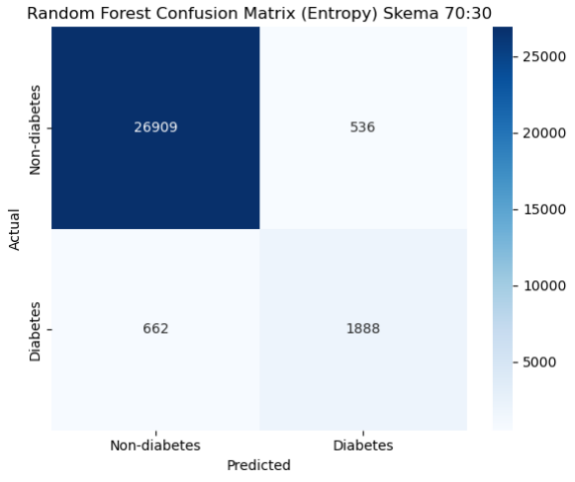
Gambar 9 *Random Forest Confusion Matrix (Entropy)* Skema 80:20

Gambar 9 confusion matrix diatas didapatkan hasil sebagai berikut:

Berdasarkan hasil klasifikasi, model berhasil mengidentifikasi 1.257 data positif dengan benar (True Positive) dan 17.908 data negatif secara tepat (True Negative). Meski demikian, masih terdapat 389 kasus yang salah diklasifikasikan sebagai positif padahal seharusnya negatif (False Positive), serta 443 kasus positif yang tidak terdeteksi oleh model (False Negative). Hasil ini menunjukkan bahwa model memiliki kinerja yang cukup baik, meskipun masih terdapat ruang untuk perbaikan dalam mengurangi kesalahan klasifikasi.

Berdasarkan hasil evaluasi dari model Random Forest dengan entropy sebagai kriteria pemisahnya menunjukkan performa klasifikasi yang sangat baik pada skema data 80:20. Akurasi yang dicapai sebesar 95,84% menunjukkan bahwa sebagian besar prediksi model adalah benar. Presisi sebesar 76,37% menunjukkan bahwa dari seluruh individu yang diprediksi menderita diabetes, sekitar 76% benar-benar positif.

Sementara itu, nilai recall sebesar 73,94% menunjukkan bahwa model mampu mengenali sebagian besar individu yang benar-benar menderita diabetes. Kombinasi keduanya menghasilkan F1-score sebesar 75,13%, yang mencerminkan keseimbangan antara kemampuan mengenali kasus positif dan menghindari kesalahan prediksi positif.

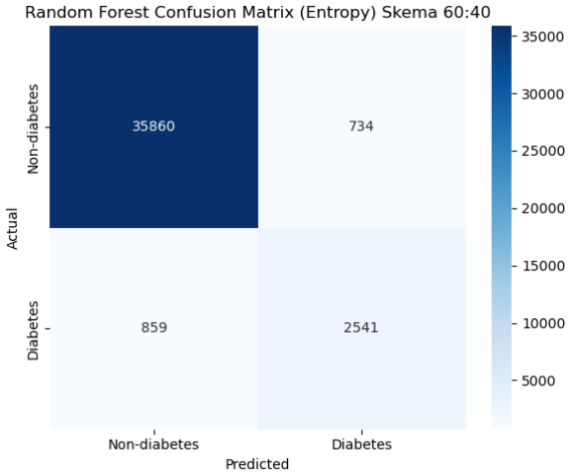


Gambar 10 *Random Forest Confusion Matrix (Entropy)* Skema 70:30

Gambar 10 confusion matrix diatas didapatkan hasil sebagai berikut:

Berdasarkan hasil klasifikasi, model berhasil mengklasifikasikan 1.888 data positif dengan benar (True Positive) dan 26.909 data negatif secara tepat (True Negative). Namun, masih terdapat 536 data yang salah diklasifikasikan sebagai positif padahal sebenarnya negatif (False Positive) serta 662 data positif yang tidak berhasil terdeteksi oleh model (False Negative). Hasil ini mencerminkan bahwa model memiliki performa yang cukup baik dalam membedakan kedua kelas, meskipun masih terdapat kesalahan klasifikasi yang perlu diperhatikan.

Berdasarkan hasil evaluasi dari model Random Forest dengan entropy sebagai fungsi pemisahan menunjukkan performa yang sangat baik pada skema 70:30. Nilai akurasi sebesar 96,01% mengindikasikan bahwa sebagian besar prediksi model sesuai dengan data aktual. Nilai presisi sebesar 77,89% menandakan bahwa dari seluruh individu yang diprediksi menderita diabetes, hampir 78% benar-benar positif. Sedangkan recall sebesar 74,04% menunjukkan bahwa model mampu mengidentifikasi sebagian besar kasus positif dari total kasus diabetes yang ada. Nilai F1-score sebesar 75,91% mencerminkan keseimbangan antara presisi dan recall, menjadikannya metrik penting terutama pada kasus data yang tidak seimbang.



Gambar 11 *Random Forest Confusion Matrix (Entropy)* Skema 60:40

Gambar 11 confusion matrix diatas menggambarkan performa klasifikasi model terhadap data uji dengan skema 60:40 yang menghasilkan jumlah klasifikasi berdasarkan confusion matrix adalah:

Hasil klasifikasi, model berhasil mengidentifikasi 2.541 data positif dengan

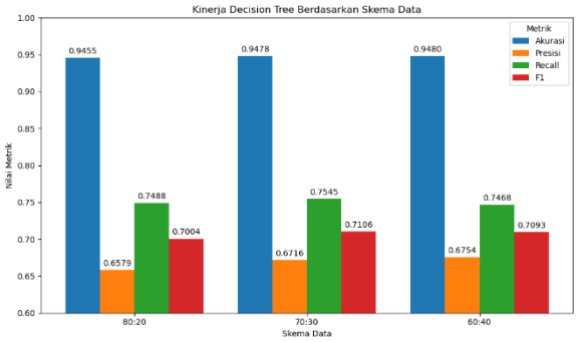
benar (True Positive) dan 35.860 data negatif secara tepat (True Negative). Meskipun demikian, terdapat 734 data yang salah diklasifikasikan sebagai positif padahal sebenarnya negatif (False Positive), serta 859 data positif yang tidak terdeteksi oleh model (False Negative). Hasil ini menunjukkan bahwa model memiliki kinerja yang baik dalam membedakan kedua kelas, meskipun masih ada kesalahan klasifikasi yang dapat diperbaiki untuk meningkatkan akurasi secara keseluruhan.

Model ini memiliki akurasi yang tinggi (96%), yang menunjukkan bahwa secara keseluruhan model cukup andal. Namun, Recall (74.74%) menunjukkan bahwa masih ada sekitar 25% penderita diabetes yang tidak terdeteksi oleh model (false negative). Precision (77.59%) menunjukkan bahwa sekitar 22% prediksi diabetes adalah salah (false positive).

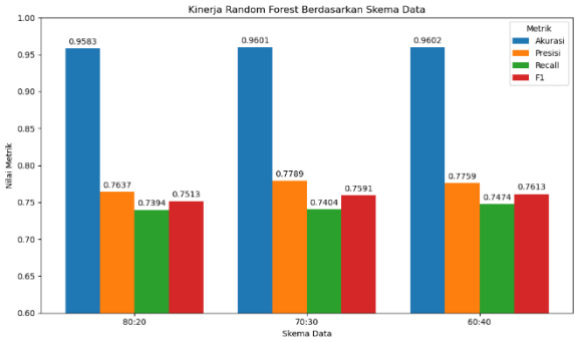
3.5 Perbandingan Model

Skema	Algoritma	Akurasi	Presisi	Recall	F1-Score
80:20	**Decision Tree**	**0.9455**	**0.6579**	**0.7488**	**0.7004**
	Random Forest	**0.9583**	**0.7637**	**0.7394**	**0.7513**
70:30	**Decision Tree**	**0.9478**	**0.6716**	**0.7545**	**0.7106**
	Random Forest	**0.9601**	**0.7789**	**0.7404**	**0.7591**
60:40	**Decision Tree**	**0.9480**	**0.6754**	**0.7468**	**0.7093**
	Random Forest	**0.9602**	**0.7759**	**0.7474**	**0.7613**

Berdasarkan tabel tersebut dapat dilihat perbandingan dari kedua metode tersebut, bahwa metode *random forest* lebih unggul dalam seluruh metriks evaluasi. Skema 60:40 menghasilkan performa tertinggi di kedua algoritma.Pemanfaatan metode SMOTE sangat penting untuk mengatasi ketimpangan kelas. Berikut ini grafik untuk memperkuat tabel tersebut adalah sebagai berikut:



Gambar 12. Grafik Kinerja Matriks dalm *Decision Tree*



Gambar 13 Grafik Kinerja Matriks dalam *Random Forest*

Gambar 12 dan gambar 13 tersebut menunjukkan perbandingan kinerja model yang menyatakan bahwa algoritma Decision Tree memiliki nilai akurasi yang tinggi di ketiga skema pembagian data (80:20, 70:30, dan 60:40), dengan nilai tertinggi sebesar 0,9480 pada skema 60:40. Namun, meskipun akurasi cukup baik, nilai presisi, recall, dan F1-score masih tergolong rendah dan kurang stabil. Presisi berkisar antara 0,6579–0,6754, recall antara 0,6716–0,7488, dan F1-score antara 0,7004–0,7106. Hal ini menunjukkan bahwa model Decision Tree kurang optimal dalam mengidentifikasi kelas minoritas,

sehingga terjadi ketidakseimbangan antara prediksi positif dan negatif, yang penting dalam konteks klasifikasi seperti deteksi diabetes.

Sebaliknya, model Random Forest menunjukkan performa yang lebih unggul dan konsisten pada seluruh skema data. Akurasi tertinggi tercapai pada skema 60:40 sebesar 0,9602, dengan nilai presisi antara 0,7394–0,7759, recall antara 0,7404–0,7474, dan F1-score antara 0,7513–0,7613. Keseimbangan antara metrik evaluasi ini mencerminkan kemampuan Random Forest dalam menangani ketidakseimbangan kelas secara lebih baik dibandingkan Decision Tree. Dengan demikian, Random Forest dapat disimpulkan sebagai algoritma yang lebih efektif dan andal dalam melakukan klasifikasi pada dataset prediksi diabetes, terutama dalam konteks distribusi data yang tidak seimbang.

4. KESIMPULAN

Penelitian ini menggunakan dataset publik yang diperoleh dari platform Kaggle untuk menganalisis dan memprediksi kemungkinan terjadinya diabetes pada individu berdasarkan sejumlah variabel klinis dan demografis. Hasil eksplorasi data menunjukkan bahwa beberapa faktor seperti usia lanjut, hipertensi, riwayat penyakit jantung, indeks massa tubuh (BMI), kadar HbA1c, kadar glukosa darah, serta riwayat merokok memiliki hubungan yang signifikan terhadap peningkatan risiko diabetes. Ketidakseimbangan kelas dalam data berhasil ditangani dengan metode SMOTE, yang mampu meningkatkan kemampuan model dalam mengenali kelas minoritas (penderita diabetes).

Model prediktif dibangun menggunakan dua algoritma machine learning, yaitu Decision Tree dan Random Forest. Hasil evaluasi menunjukkan bahwa Random Forest secara konsisten memberikan performa yang lebih unggul pada semua skema pembagian data (80:20, 70:30, dan 60:40), baik dari sisi akurasi, presisi, recall, maupun F1-score. Skema 60:40 memberikan hasil terbaik dengan akurasi mencapai 96,02% dan F1-score sebesar 76,13%. Hal ini mengindikasikan bahwa model Random Forest lebih andal dalam mengidentifikasi kasus diabetes dibandingkan Decision Tree.

Dengan demikian, studi ini membuktikan bahwa penggunaan algoritma machine learning berbasis data publik dapat menghasilkan sistem prediksi yang akurat dan potensial untuk diterapkan dalam sistem pendukung keputusan medis, khususnya dalam skrining awal dan pencegahan dini terhadap diabetes.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada semua pihak yang telah memberikan dukungan dan kontribusi dalam penyusunan penelitian ini. Ucapan terima kasih khusus disampaikan kepada dosen pembimbing atas bimbingan dan arahannya yang sangat berarti selama proses penelitian berlangsung. Penulis juga berterima kasih kepada platform Kaggle yang telah menyediakan dataset publik yang digunakan dalam penelitian ini, serta kepada rekan-rekan dan keluarga yang senantiasa memberikan semangat dan motivasi. Semoga hasil penelitian ini dapat memberikan manfaat dan kontribusi positif bagi pengembangan ilmu pengetahuan, khususnya di bidang prediksi penyakit berbasis data.ucapana terima kasih kepada

orang orang yang telah membantu penelitian (tidak termasuk author yang telah disebutkan). Selain itu disebutkan pula sumber pendanaan penelitian beserta tahun dan institusi yang memberikan.

DAFTAR PUSTAKA

[1] F. R. Aftha Harianto, Z. Alawi, and I. A. Sa'ida, "Pengaruh Komposisi Split Data Pada Akurasi Klasifikasi Penderita Diabetes Menggunakan Algoritma Machine Learning," *J. Sist. Inf. dan Inform.*, vol. 8, no. 1, pp. 36–44, 2025, doi: 10.47080/simika.v8i1.3663.

[2] A. G. Pertiwi, N. Bachtiar, R. Kusumaningrum, I. Waspada, and A. Wibowo, "Comparison of performance of k-nearest neighbor algorithm using smote and k-nearest neighbor algorithm without smote in diagnosis of diabetes disease in balanced data," in *Journal of Physics: Conference Series*, Apr. 2020, p. 012048. doi: 10.1088/1742-6596/1524/1/012048.

[3] Dewi Nasien *et al.*, "Perbandingan Implementasi Machine Learning Menggunakan Metode KNN, Naive Bayes, dan Logistik Regression Untuk Mengklasifikasi Penyakit Diabetes," *JEKIN - J. Tek. Inform.*, vol. 4, no. 1, pp. 10–17, 2024, doi: 10.58794/jekin.v4i1.640.

[4] M. S. Fiqri and H. D. Bhakti, "Klasifikasi Potensi Penyakit Diabetes Mellitus Tipe Ii Pada Pasien Menggunakan Algoritme Knn (K-Nearest Neighbor)," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 8, no. 4, pp. 7305–7313, 2024.

[5] "Text Mining untuk Analisis Kasus Stunting di Nusa Tenggara Barat".

[6] D. A. S. K. Purwaningrum and D. Agustina, "Implementation of Machine Learning Algorithm C4.5 in Classification of Patients With Type 2 Diabetes Mellitus," *Barekeng*, vol. 18, no. 1, pp. 193–204, 2024, doi: 10.30598/barekengvol18iss1pp0193-0204.

[7] S. Lonang, A. Yudhana, and M. K. Biddinika, "Analisis Komparatif Kinerja Algoritma Machine Learning untuk Deteksi Stunting," *J. Media Inform. Budidarma*, vol. 7, no. 4, p. 2109, 2023, doi: 10.30865/mib.v7i4.6553.

[8] F. Syuhada, Y. Sa'adati, L. A. Apriani, S. Lonang, and A. F. D. Putra, "Text Mining untuk Analisis Kasus Stunting di Nusa Tenggara Barat," *J. Ilm. Edutic Pendidik. dan Inform.*, vol. 12, no. 1, pp. 29–42, 2025, doi: 10.21107/edutic.v12i1.29522.

[9] Purbandini, E. Purwanti, E. Hariyanti, and F. Y. Ramadhan, "Application of the Decision Tree C4.5 Method on the Classification of Diet Types of People with Diabetes Mellitus," *AIP Conf. Proc.*, vol. 2975, no. 1, 2023, doi: 10.1063/5.0187456.

[10] I. P. Gede, A. Sudiatmika, P. S. Saputra, and M. R. Ulil, "a Comparative Study of K-Nearest Neighbors and Random Forest Classifiers on Diabetes," vol. 3, no. December, pp. 122–129, 2024.

[11] A. S. Kirono and Y. Nataliani, "Perbandingan Algoritma Machine Learning dalam Analisis Penyebab Penyakit Gagal Jantung," *J. Edukasi dan Penelit. Inform.*, vol. 10, no. 2, pp. 296–301, 2024, doi: 10.26418/jp.v10i2.

[12] M. Fahmy Amin, "Confusion Matrix in Binary Classification Problems: A Step-by-Step Tutorial," *J. Eng. Res.*, vol. 6, no. 5, pp. 0–0, 2022, doi: 10.21608/erjeng.2022.274526.

[13] F. D. Astuti and F. N. Lenti, "Implementasi SMOTE untuk Mengatasi Imbalance Class pada Klasifikasi Car Evolution Menggunakan K-NN," *J. JUPITER*, vol. 13, no. 1, pp. 89–98, 2021.

[14] E. Erlin, Y. Desnelita, N. Nasution, L. Suryati, and F. Zoromi, "Dampak SMOTE terhadap Kinerja Random Forest Classifier berdasarkan Data Tidak seimbang," *MATRIK J. Manajemen, Tek. Inform. dan Rekayasa Komput.*, vol. 21, no. 3, pp. 677–690, 2022, doi: 10.30812/matrik.v21i3.1726.

[15] R. Ridwan, E. H. Hermaliani, and M.

Ernawati, "Penerapan: Penerapan Metode SMOTE Untuk Mengatasi Imbalanced Data Pada Klasifikasi Ujaran Kebencian," *Comput. Sci.*, vol. 4, no. 1, pp. 80–88, 2024, [Online]. Available: https://jurnal.bsi.ac.id/index.php/co-science/article/view/2990