



Analisis Perbandingan Algoritma Random Forest Dan Decision Tree Pada Prediksi Penyakit Diabetes

Ardilla Ekrinifda¹, Salsabila Aulia Ramadhan², Shera Marvella³, Maulana Fansyuri⁴

^{1,2,3,4}Ilmu Komputer, Teknik Informatika, Universitas Pamulang, Tangerang Selatan, Indonesia

Email: ¹ardillaekrinifda@gmail.com, ²salsabilaauliar1@gmail.com, ³sheramv1115@gmail.com,

⁴dosen02359@unpam.ac.id

(* : coressponding author)

Abstrak - Diabetes melitus adalah penyakit kronis dengan prevalensi yang meningkat secara global. Diagnosis dini sangat penting untuk mencegah komplikasi, tetapi metode tradisional sering memerlukan waktu dan akses yang terbatas. Penelitian ini membandingkan algoritma Random Forest dan Decision Tree dalam memprediksi diabetes menggunakan dataset dari Kaggle. Algoritma ini dipilih karena kemampuan mereka menangani data kesehatan yang kompleks. Random Forest menggunakan pendekatan ensemble learning untuk meningkatkan akurasi dan stabilitas, sedangkan Decision Tree menawarkan interpretasi hasil yang lebih intuitif. Evaluasi dilakukan dengan metrik seperti akurasi, presisi, recall, dan Area Under Curve (AUC). Hasil menunjukkan Random Forest unggul dalam akurasi (78,78%) dan stabilitas dibandingkan Decision Tree (77,34%). Namun, Decision Tree lebih efisien secara komputasi dan mudah diinterpretasi. Analisis ini memberikan wawasan dalam memilih algoritma prediksi diabetes yang sesuai berdasarkan kebutuhan klinis dan sumber daya. Penelitian ini merekomendasikan Random Forest untuk keandalan prediksi dan Decision Tree untuk skenario yang memerlukan efisiensi dan interpretabilitas. Implementasi lebih lanjut diharapkan membantu tenaga medis dalam pengambilan keputusan yang lebih baik.

Kata Kunci: *Random Forest; Decision Tree; Diabetes; RapidMiner; Prediksi*

Abstract – Diabetes mellitus is a chronic disease with increasing prevalence globally. Early diagnosis is essential to prevent complications, but traditional methods often require limited time and access. This research compares the Random Forest and Decision Tree algorithms in predicting diabetes using datasets from Kaggle. These algorithms were chosen for their ability to handle complex health data. Random Forest uses an ensemble learning approach to improve accuracy and stability, while Decision Tree offers a more intuitive interpretation of results. Evaluation is done with metrics such as accuracy, precision, recall, and Area Under Curve (AUC). The results show that Random Forest is superior in accuracy (78.78%) and stability compared to Decision Tree (77.34%). However, Decision Trees are more computationally efficient and easier to interpret. This analysis provides insight into selecting an appropriate diabetes prediction algorithm based on clinical needs and resources. This research recommends Random Forest for prediction reliability and Decision Tree for scenarios that require efficiency and interpretability. Further implementation is expected to help medical personnel make better decisions.

Keywords: *Random Forest; Decision Tree; Diabetes; RapidMiner; Prediksi*

1. PENDAHULUAN

Deteksi dini penyakit diabetes menjadi sangat krusial dalam upaya pencegahan komplikasi dan penanganan yang tepat. Namun, proses diagnosis diabetes tradisional seringkali membutuhkan waktu yang lama dan melibatkan berbagai tes laboratorium yang kompleks. Hal ini dapat menyebabkan keterlambatan dalam diagnosis dan penanganan, yang pada akhirnya dapat memperburuk kondisi pasien (Alrasyid et al., 2024). Selain itu, keterbatasan akses ke fasilitas kesehatan di beberapa daerah juga menjadi tantangan tersendiri dalam proses diagnosis diabetes secara konvensional (Aditya et al., 2024).

Random Forest dan Decision Tree merupakan dua algoritma klasifikasi yang populer dalam machine learning karena kemampuannya dalam menangani data kompleks dan memberikan hasil yang dapat diinterpretasi (Sapriadi, 2024). Decision Tree menawarkan pendekatan yang intuitif dengan membangun pohon keputusan berdasarkan fitur-fitur yang paling informatif, sementara Random Forest menggunakan konsep ensemble learning yang menggabungkan multiple decision tree untuk menghasilkan prediksi yang lebih akurat dan robust. Kedua algoritma ini memiliki karakteristik dan kelebihan masing-masing yang perlu dievaluasi dalam konteks prediksi penyakit diabetes (Ivandari et al., 2023).



Implementasi algoritma machine learning dalam prediksi diabetes memerlukan evaluasi yang komprehensif untuk memastikan keandalan dan efektivitasnya(Suryanegara et al., 2021). Metrik evaluasi seperti accuracy, precision, recall, dan Area Under Curve (AUC) menjadi parameter penting dalam membandingkan performa kedua algoritma tersebut. Analisis perbandingan ini tidak hanya berfokus pada akurasi prediksi, tetapi juga mempertimbangkan aspek-aspek praktis seperti kompleksitas komputasi, interpretabilitas hasil, dan kemampuan generalisasi model(Andi et al., 2023).

Penelitian mengenai perbandingan algoritma Random Forest dan Decision Tree dalam prediksi diabetes menjadi relevan mengingat kebutuhan akan sistem prediksi yang tidak hanya akurat tetapi juga efisien dan dapat diimplementasikan dalam praktik klinis(Susanti et al., 2023). Hasil dari analisis perbandingan ini diharapkan dapat memberikan wawasan berharga bagi pengembangan sistem prediksi diabetes yang lebih baik dan membantu tenaga medis dalam proses pengambilan keputusan klinis. Selain itu, pemahaman mendalam tentang kelebihan dan keterbatasan masing-masing algoritma dapat menjadi dasar untuk pengembangan dan optimasi model prediksi di masa depan(H et al., 2024).

2. METODE

2.1 Sumber Data

Dataset yang digunakan dalam penelitian ini diperoleh dari platform Kaggle, sebuah komunitas data science yang menyediakan berbagai dataset untuk keperluan analisis. Dataset diabetes ini merupakan kumpulan data medis pasien yang mencakup berbagai parameter kesehatan yang berkaitan dengan diagnosis diabetes. Data ini telah melalui proses anonymization untuk menjaga privasi pasien dan telah divalidasi kualitasnya oleh komunitas Kaggle(Suryanegara et al., 2021).

2.2. Karakteristik Data

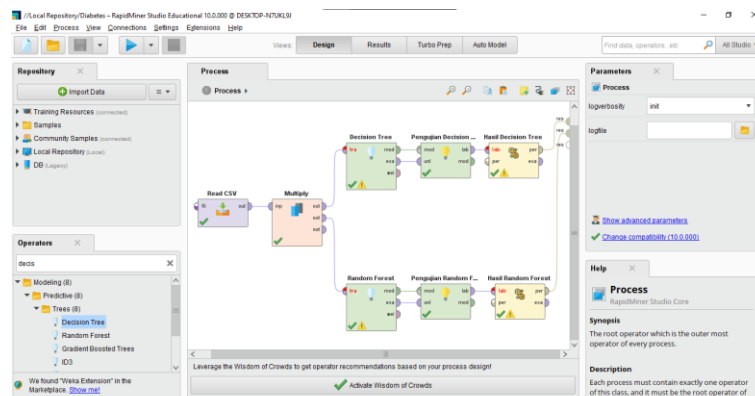
- a. Jumlah total instance/record dalam dataset
- b. Terdiri dari dua kelas target: Positif (penderita diabetes) dan Negatif (bukan penderita diabetes)
- c. Mencakup berbagai atribut kesehatan seperti:
 - Glucose: Kadar glukosa dalam darah
 - Blood Pressure: Tekanan darah
 - BMI (Body Mass Index): Indeks massa tubuh
 - Age: Usia pasien
 - Pregnancy: Jumlah kehamilan (untuk pasien wanita)
 - Insulin: Kadar insulin
 - DiabetesPedigreeFunction: Riwayat diabetes dalam keluarga
 - Skin Thickness: Ketebalan lipatan kulit

2.3. Persiapan Data

	Pregnant	Glucose	BloodPres	SkinThick	Insulin	BMI	DiabetesF	Age	Outcome
1	1	85	66	29	0.256	0.351	31	Negatif	
2	1	89	66	23	94.28.1	0.167	21	Negatif	
3	0	137	40	35	168.43.1	2.288	33	Postif	
4	1	116	74	0	0.25.6	0.201	30	Negatif	
5	3	78	50	32	88	31.0.248	26	Postif	
6	10	115	0	0	0.35.3	0.134	29	Negatif	
7	2	197	70	45	543.30.5	0.158	53	Postif	
8	8	125	96	0	0	0.0.232	54	Postif	
9	4	110	92	0	0.37.6	0.191	30	Negatif	
10	10	168	74	0	0	38.6.537	34	Postif	
11	10	139	80	0	0.27.1	1.441	57	Negatif	
12	1	189	60	23	846.30.1	0.398	59	Postif	
13	5	166	72	19	175.25.8	0.587	51	Postif	
14	7	100	0	0	0	30.0.484	32	Postif	
15	0	118	84	47	230.45.8	0.551	31	Postif	
16	7	107	74	0	0.29.6	0.254	31	Postif	
17	1	101	30	36	81.43.3	0.183	33	Negatif	
18	1	115	70	30	96.34.6	0.520	32	Postif	
19	3	126	88	41	235.39.3	0.704	27	Negatif	
20	8	99	86	0	0.30.4	0.383	30	Negatif	

Gambar 1. Dataset Diabetes

- Import dataset dalam format CSV ke RapidMiner
- Pengecekan missing values dan outliers
- Verifikasi tipe data untuk setiap atribut



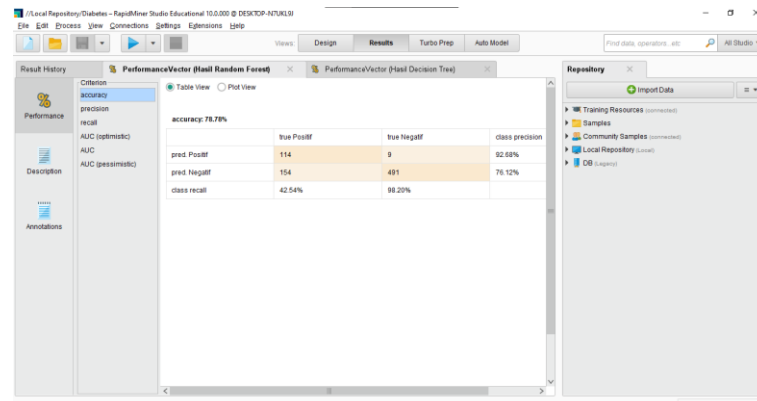
Gambar 2. Tampilan Rapidmanner

3. ANALISA DAN PEMBAHASAN

3.1 Analisis Performa Algoritma

Berikut analisa performa algoritma Random Forest dan Algoritma Decision Tree dalam memprediksi penyakit diabetes.

3.1.1 Random Forest



Gambar 3. Hasil Analisis Random Forest

Tabel 1 Hasil Random Forest

	True Positif	True Negatif	Class Precision
Pred. Positif	114	9	92,68%
Pred. Negatif	154	491	76,12%
Class Recall	42,54%	98,20%	

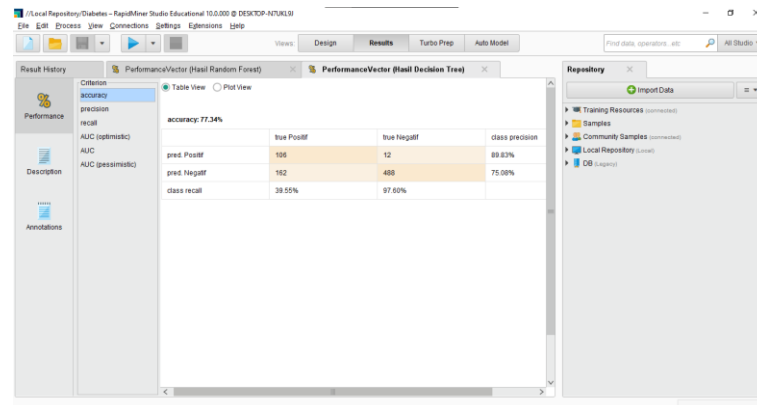
- Hasil Evaluasi Metrik:
 - Accuracy: 78.78%
 - Precision: 76.12%
 - Recall: 98.20%
 - AUC: 0.914
- Analisis Confusion Matrix:
 - True Positive (TP): 114 kasus
 - True Negative (TN): 491 kasus
 - False Positive (FP): 9 kasus
 - False Negative (FN): 154 kasus
- Interpretasi Hasil:
 - Model menunjukkan kemampuan yang baik dalam mengidentifikasi kasus negatif (spesifisitas tinggi)
 - Tingkat false positive yang rendah (9 kasus) menunjukkan model jarang melakukan kesalahan dalam mengidentifikasi kasus positif
 - Nilai AUC yang tinggi (0.914) mengindikasikan model memiliki kemampuan diskriminatif yang baik

Hasil analisis menunjukkan bahwa algoritma Random Forest unggul dalam akurasi dengan nilai 78,78%. Algoritma ini juga memiliki nilai AUC yang tinggi sebesar 0,914, yang mengindikasikan kemampuan diskriminasi yang sangat baik dalam membedakan antara kelas positif dan negatif. Selain itu, Random Forest menunjukkan spesifisitas tinggi, dengan tingkat true negative yang signifikan dibandingkan false positive, hanya sebanyak 9 kasus. Hal ini menunjukkan

algoritma ini sangat handal dalam meminimalkan kesalahan dalam mengklasifikasikan kasus negatif sebagai positif.

Kelebihan lain dari Random Forest adalah kemampuannya untuk menangani data yang memiliki banyak atribut tanpa mengalami masalah overfitting. Namun, kelemahan utama algoritma ini adalah kompleksitas komputasi yang lebih tinggi dibandingkan Decision Tree, sehingga memerlukan waktu eksekusi yang lebih lama dan sumber daya komputasi yang lebih besar.

3.1.2 Decision Tree



Gambar 4. Hasil Analisis Decision Tree

Tabel 2 Hasil Decision Tree

	True Positif	True Negatif	Class Precision
Pred. Positif	106	12	89,83%
Pred. Negatif	162	488	75,08%
Class Recall	39,55%	97,60%	

- Hasil Evaluasi Metrik:
 - Accuracy: 77.34%
 - Precision: 75.08%
 - Recall: 97.60%
 - AUC: 0.796 (average)
- Analisis Confusion Matrix:
 - True Positive (TP): 106 kasus
 - True Negative (TN): 488 kasus
 - False Positive (FP): 12 kasus
 - False Negative (FN): 162 kasus
- Interpretasi Hasil:
 - Model menunjukkan performa yang cukup stabil dalam klasifikasi
 - Terdapat variasi yang signifikan dalam nilai AUC, menunjukkan potensi ketidakstabilan model
 - Jumlah false negative yang lebih tinggi dibandingkan Random Forest



Decision Tree, meskipun memiliki akurasi lebih rendah (77,34%) dibandingkan Random Forest, menawarkan keunggulan dari segi efisiensi komputasi dan interpretasi hasil yang lebih mudah. Model ini memberikan nilai presisi yang cukup baik sebesar 75,08% dan recall sebesar 97,60%. Nilai recall yang tinggi menunjukkan bahwa Decision Tree lebih efektif dalam mengidentifikasi kasus positif, meskipun disertai dengan jumlah false negative yang lebih tinggi (162 kasus) dibandingkan Random Forest.

Nilai AUC rata-rata sebesar 0,796 mengindikasikan bahwa Decision Tree masih memiliki performa yang stabil, tetapi potensi variasi atau ketidakstabilan dalam performa model lebih besar dibandingkan Random Forest. Hal ini dapat menjadi faktor yang perlu dipertimbangkan, terutama untuk implementasi jangka panjang yang memerlukan hasil konsisten.

4. KESIMPULAN

- a. **Keunggulan Performa:** Random Forest menunjukkan keunggulan dalam hal akurasi dan stabilitas dibandingkan Decision Tree di semua metrik evaluasi. Meski perbedaannya tidak terlalu signifikan, performa Random Forest secara konsisten lebih baik, menjadikannya algoritma yang lebih dapat diandalkan untuk tugas prediksi.
- b. **Stabilitas Model:** Random Forest memiliki stabilitas yang lebih tinggi dibandingkan Decision Tree, terutama berdasarkan nilai AUC (Area Under the Curve). Hal ini membuat Random Forest lebih cocok untuk implementasi jangka panjang di mana konsistensi hasil sangat dibutuhkan. Sebaliknya, Decision Tree cenderung memiliki variasi performa yang lebih besar, yang dapat menjadi tantangan dalam aplikasi praktis.
- c. **Trade-off Penggunaan:** Keunggulan Random Forest dalam akurasi dan stabilitas datang dengan konsekuensi berupa kompleksitas yang lebih tinggi, baik dari segi komputasi maupun waktu eksekusi. Sebagai alternatif, Decision Tree lebih sederhana dan mudah diinterpretasikan, sehingga lebih efisien untuk digunakan, terutama jika interpretabilitas menjadi prioritas utama.
- d. **Pertimbangan Implementasi:** Kedua algoritma layak dipertimbangkan untuk implementasi, tergantung pada kebutuhan spesifik dan keterbatasan aplikasi. Jika kebutuhan utama adalah akurasi tinggi dan stabilitas, Random Forest adalah pilihan yang lebih baik. Namun, jika efisiensi komputasi dan interpretabilitas lebih penting, Decision Tree dapat menjadi alternatif yang sesuai.

REFERENCES

- Aditya, M. F., Pramuntadi, A., Wijaya, D. P., & Wicaksono, Y. (2024). Implementasi Metode Decision Tree pada Prediksi Penyakit Diabetes Melitus Tipe 2. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, 4(3), 1104–1110. <https://doi.org/10.57152/malcom.v4i3.1284>
- Alrasyid, H., Homaidi, A., Kom, M., Fatah, Z., & Kom, M. (2024). *Comparison Support Vector Machine and Random Forest Algorithms in Detect Diabetes*. 1(1), 447–453.
- Andi, Thamrin, Susanto, A., Wijaya, E., & Djohan, D. (2023). Analysis of the random forest and grid search algorithms in early detection of diabetes mellitus disease. *Jurnal Mantik*, 7(2), 2685–4236.
- H, M. I. A., Amran, A., Desiani, A., & Napitu, M. J. (2024). *Comparison of Classification Results of SVM , KNN , Decision Tree , and Ensemble Methods in Diabetes Diagnosis*. 76–82. <https://doi.org/10.37034/medinftech.v2i3.62>
- Ivandari, Much. Rifqi Maulana, Muhammad Faizal Kurniawan, & Al Karomi, M. A. (2023). Komparasi Algoritma Data Mining untuk Klasifikasi Penyakit Diabetes. *Bulletin of Computer Science Research*, 3(5), 343–350. <https://doi.org/10.47065/bulletincsr.v3i5.280>
- Sapriadi. (2024). *K-Nearest Neighbors, decision trees and random forest for diabetes prediction*. 7(4).
- Suryanegara, G. A. B., Adiwijaya, & Purbolaksono, M. D. (2021). Peningkatan Hasil Klasifikasi pada Algoritma Random Forest untuk Deteksi. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, 1(10), 114–122.
- Susanti, Z., Sirait, P., & Panjaitan, E. S. (2023). Peningkatan Kinerja Random Forest Melalui Seleksi Fitur Secara Pca Untuk Mendeteksi Penyakit Diabetes Tahap Awal. *Sains Dan Teknologi*, 4(3), 51–56.