 ISSN 2987-484X

Perbandingan Algoritma Random Forest dan Naive Bayes dalam Memprediksi Penyakit Diabetes

Muhammad Kholish[1], Anggi Herdianto[2], Rizqi Fakhri Setiawan[3], Riza Samsinar[4]
[1, 2, 3, 4] Teknik Elektro, Fakultas Teknik, Universitas Muhammadiyah Jakarta
Jl. Cempaka Putih Tengah 27, Cempaka Putih, Jakarta Pusat 10510
Telp. 0214244016
E-mail: 20210410200023@student.umj.ac.id

Abstrak

Diabetes adalah penyakit kronis yang memberikan dampak signifikan terhadap kesehatan masyarakat global. Penyakit ini menjadi salah satu penyebab utama kematian di dunia dengan komplikasi serius seperti penyakit kardiovaskular, gagal ginjal, neuropati, dan kebutaan. Selain dampak kesehatan, diabetes juga menimbulkan beban ekonomi yang besar, baik bagi individu maupun sistem kesehatan. Oleh karena itu, deteksi dini menjadi langkah penting untuk meningkatkan kualitas hidup pasien, mencegah komplikasi, dan mengurangi biaya pengobatan jangka panjang. Penelitian ini bertujuan untuk membandingkan kinerja algoritma Random Forest dan Naive Bayes dalam memprediksi risiko diabetes menggunakan dataset kesehatan. Dataset yang digunakan mencakup 768 data pasien dengan berbagai parameter, seperti kadar glukosa, tekanan darah, indeks massa tubuh (BMI), usia, serta faktor risiko lainnya. Dataset ini dibagi menjadi data latih (80%) dan data uji (20%) secara acak untuk memastikan validitas model. Proses penelitian melibatkan tahapan prapemrosesan data, implementasi algoritma, dan evaluasi performa menggunakan metrik seperti Accuracy, Precision, Recall, dan F1-score. Hasil penelitian menunjukkan bahwa algoritma Naive Bayes memberikan akurasi lebih tinggi (77%) dibandingkan Random Forest (72%). Selain itu, Naive Bayes lebih unggul dalam mendeteksi kasus positif diabetes, seperti tercermin pada metrik Precision, Recall, dan F1-score yang lebih baik. Penelitian ini diharapkan berkontribusi pada pengembangan sistem deteksi dini berbasis data yang lebih andal dan akurat. Selain itu, penelitian lanjutan dapat mengeksplorasi algoritma lain, seperti XGBoost atau Gradient Boosting, untuk meningkatkan performa lebih lanjut.
Kata Kunci: Machine Learning, Perbandingan, Random Forest, Naive Bayes, Diabetes.

Abstract

Diabetes is a chronic disease that has a significant impact on global public health. This disease is one of the main causes of death in the world with serious complications such as cardiovascular disease, kidney failure, neuropathy and blindness. In addition to its health impacts, diabetes also poses a large economic burden, both for individuals and the health system. Therefore, early detection is an important step to improve patient quality of life, prevent complications, and reduce long-term treatment costs. This study aims to compare the performance of Random Forest and Naive Bayes algorithms in predicting diabetes risk using health datasets. The dataset used includes 768 patient data with various parameters, such as glucose levels, blood pressure, body mass index (BMI), age, and other risk factors. This dataset is divided into training data (80%) and test data (20%) randomly to ensure the validity of the model. The research process involves stages of data preprocessing, algorithm implementation, and performance evaluation using metrics such as Accuracy, Precision, Recall, and F1-score. The research results show that the Naive Bayes algorithm provides higher accuracy (77%) than Random Forest (72%). In addition, Naive Bayes is superior in detecting positive cases of diabetes, as reflected in better Precision, Recall and F1-score metrics. This research is expected to contribute to the development of a more reliable and accurate data-based early detection system. Additionally, further research could explore other algorithms, such as XGBoost or Gradient.
Keywords: Machine Learning, Comparison, Random Forest, Naive Bayes, Diabetes.

1. Pendahuluan

Diabetes mellitus adalah salah satu masalah kesehatan global yang paling serius dan terus meningkat prevalensinya. Menurut Organisasi Kesehatan Dunia (WHO), pada tahun 2021 terdapat lebih dari 422 juta orang di seluruh dunia yang hidup dengan diabetes, dengan sekitar 1,5 juta kematian terkait setiap tahunnya (World Health Organization, 2024). Penyakit ini tidak hanya berdampak pada individu, tetapi juga menjadi beban ekonomi yang signifikan, baik pada tingkat keluarga maupun sistem kesehatan global. Dalam banyak kasus, diabetes menyebabkan komplikasi serius seperti penyakit kardiovaskular, gagal ginjal, neuropati, dan kebutaan (Nassar *et al.*, 2021). Faktor-faktor risiko utama, seperti pola makan tinggi gula, kurangnya aktivitas fisik, obesitas, dan riwayat keluarga, terus menjadi tantangan dalam upaya pencegahan penyakit ini.

Di Indonesia, diabetes menjadi ancaman yang semakin nyata. Berdasarkan laporan Riset Kesehatan Dasar (Kemenkes, 2018), prevalensi diabetes pada penduduk dewasa meningkat dari 6,9% pada 2013 menjadi 10,9% pada 2018 (Kementerian Kesehatan RI, 2018). Peningkatan ini terutama disebabkan oleh perubahan gaya hidup, termasuk pola makan tidak sehat dan urbanisasi yang mendorong gaya hidup sedenter. Diabetes tipe 2, yang menyumbang mayoritas kasus, sering kali tidak terdeteksi hingga timbul komplikasi serius. Hal ini membuat deteksi dini menjadi langkah krusial untuk mengurangi dampak buruk penyakit ini, baik dari segi kesehatan maupun ekonomi (Webber, 2013).

Kemajuan teknologi kecerdasan buatan (AI) dan pembelajaran mesin (*Machine Learning*) menawarkan solusi baru dalam mendeteksi diabetes sejak dini. Dengan kemampuan untuk menganalisis data dalam jumlah besar dan dengan kecepatan tinggi, algoritma machine learning memungkinkan prediksi yang lebih akurat berdasarkan parameter klinis seperti kadar glukosa, tekanan darah, indeks massa tubuh (BMI), dan riwayat keluarga diabetes. Pendekatan ini tidak hanya mempercepat diagnosis, tetapi juga memberikan peluang untuk mengembangkan sistem kesehatan yang lebih efisien (Nasional *et al.*, 2024).

Penelitian ini berfokus pada perbandingan dua algoritma *machine learning*, yaitu Random Forest dan Naive Bayes, dalam memprediksi risiko diabetes. Random Forest dikenal sebagai algoritma berbasis ensemble yang mampu menangani dataset kompleks dengan interaksi fitur yang rumit, sementara Naive Bayes menggunakan pendekatan probabilistik sederhana namun sangat efisien pada dataset yang lebih kecil dan independen (Kamel, Abdulah and Al-Tuwaijari, 2019). Kedua algoritma ini telah digunakan dalam berbagai penelitian sebelumnya, namun masih diperlukan studi lebih lanjut untuk mengevaluasi kinerja mereka secara langsung dalam konteks prediksi diabetes.

Dengan mengevaluasi kinerja kedua algoritma menggunakan metrik seperti akurasi, precision, recall, dan F1-score, penelitian ini bertujuan untuk menentukan algoritma yang paling sesuai untuk mendukung sistem deteksi dini diabetes. Diharapkan hasil penelitian ini tidak hanya memberikan kontribusi dalam pengembangan teknologi kesehatan berbasis data, tetapi juga membantu meningkatkan kualitas hidup pasien melalui deteksi dini yang lebih andal dan efisien (Erlin *et al.*, 2022).

2. Metodologi

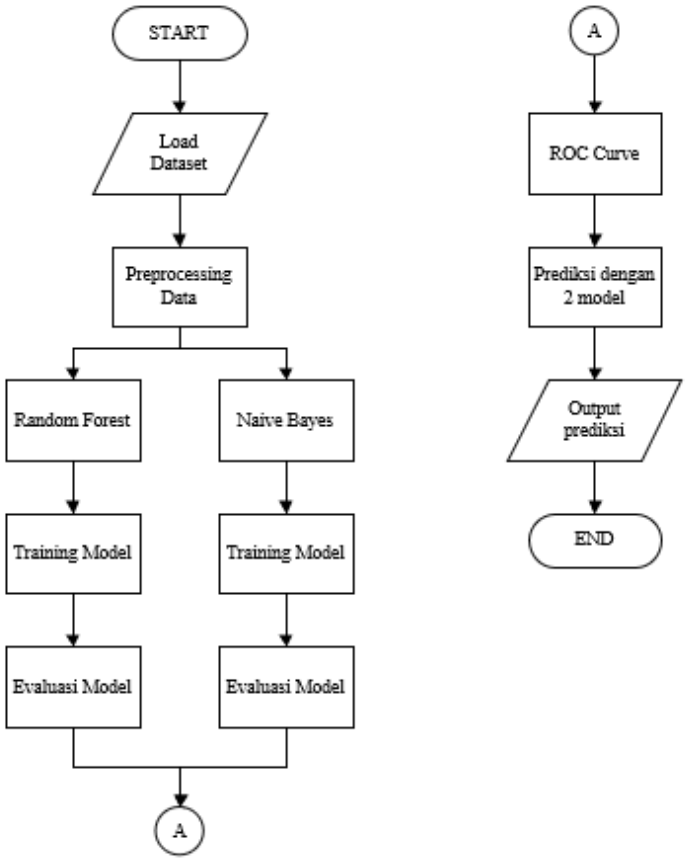
2.1. Desain penelitian

Fokus utama penelitian ini adalah membandingkan prediksi pasien berkemungkinan terkena diabetes dengan Random Forest dan Naive Bayes. Penelitian ini dilakukan dengan menggunakan dataset diabetes yang terdiri dari 768 sampel dengan delapan fitur *input* dan satu label *output* . Dataset dibagi menjadi dua set, yaitu data latih (80%) dan data uji (20%), menggunakan teknik pembagian acak (Dhawas *et al.*, 2024).

Berikut adalah alur kerja yang menggambarkan bagaimana dua algoritma (Random Forest dan Naive Bayes) diproses secara paralel untuk pelatihan, evaluasi, dan prediksi. Pendekatan ini memungkinkan perbandingan performa kedua algoritma berdasarkan

 ISSN 2987-484X

metrik evaluasi seperti akurasi dan *Area Under the Curve* (AUC), dengan hasil akhir berupa prediksi untuk mendukung analisis risiko diabetes.



Gambar 2. 1. Flowchart analisis dan prediksi diabetes

2.2. Bahan penelitian

a. Perangkat Lunak:

Visual Studio Code digunakan sebagai *software* utama untuk analisis data dan dengan bahasa pemrograman Python versi 3.11.4.

b. Perangkat Keras:

Tabel 2. 1. Spesifikasi laptop

Model Laptop	Acer Swift SF314-56G
Prosesor	Intel®Core™ i7-8565U
GPU	NVIDIA GeForce MX250
Memory	20480MB RAM
DirectX Version	DirectX 12
OS	Windows 10 Pro 64-bit

3. Hasil dan Pembahasan

3.1. Preproses data

a. Dataset:

Dataset ini adalah untuk secara diagnostik memprediksi apakah seorang pasien menderita diabetes, berdasarkan beberapa pengukuran diagnostik yang tercantum dalam dataset. Beberapa batasan diterapkan pada pemilihan ini secara khusus, semua pasien di sini adalah perempuan.

 ISSN 2987-484X

Tabel 3. 1. Tabel dataset

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1

Dataset ini berisi sebanyak 768 data dan informasi tentang faktor risiko dan diagnosis diabetes dengan kolom sebagai berikut:

- Pregnancies: Jumlah kehamilan.
- Glucose: Kadar glukosa darah. (Puasa: 70–99 mg/dL, 2 jam setelah makan: <140 mg/dL, HbA1c: <5.7%).
- BloodPressure: Tekanan darah diastolik. (60–80 mmHg diastolik, tekanan darah normal: 120/80 mmHg).
- SkinThickness: Ketebalan kulit triceps. (10–20 mm, lebih dari 25 mm bisa menunjukkan obesitas).
- Insulin: Tingkat insulin serum. (Puasa: 2–25 µU/mL).
- BMI: Indeks massa tubuh. (18.5–24.9 normal, <18.5 kurus, 25–29.9 kelebihan berat badan, ≥30 obesitas).
- DiabetesPedigreeFunction: Fungsi garis keturunan diabetes. (0.08–2.42).
- Age: Usia pasien.
- Outcome: Diagnosis diabetes (1 untuk positif dan 0 untuk negatif).

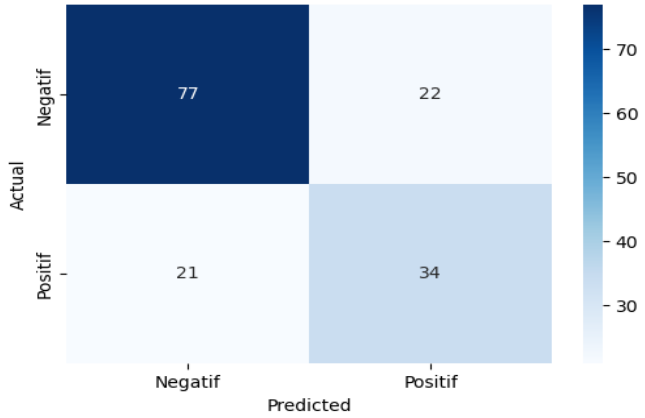
b. Pembagian dataset:

- Data latih: 0.8 atau 80% untuk melatih model.
- Data uji: 0.2 atau 20% untuk menguji performa model.
- Pembagian dilakukan secara acak dengan parameter random_state = 42 untuk memastikan hasil pembagian konsisten dan dapat direproduksi.

3.2. Implementasi algoritma

a. Random Forest:

Random Forest adalah algoritma yang terdiri dari sejumlah Decision Tree yang tumbuh hingga selesai tanpa perlu dilakukan pemangkasan. Semakin banyak pohon yang dibangun, semakin akurat hasil prediksinya dan cenderung mengurangi risiko *overfitting*. Algoritma ini menghasilkan perkiraan keseluruhan dari semua pohon dan memiliki keunggulan dalam pemilihan fitur secara otomatis (Lin *et al.*, 2017).



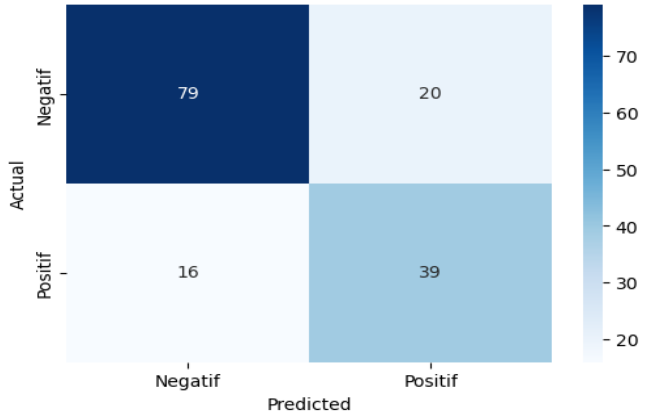
Gambar 3. 1. Confusion matrix Random Forest

b. Naive Bayes:

Naive Bayes adalah algoritma probabilistik yang menggunakan Teorema Bayes dengan asumsi independensi antar fitur, menjadikannya metode yang sederhana namun

 ISSN 2987-484X

efektif untuk berbagai tugas klasifikasi. Algoritma ini sangat cocok untuk dataset berdimensi tinggi karena efisiensinya dalam waktu komputasi. Dalam penelitian ini, digunakan Gaussian Naive Bayes untuk menangani fitur kontinu, dengan asumsi bahwa data mengikuti distribusi normal. Meskipun asumsi independensi antar fitur sering kali tidak realistis dalam data dunia nyata, algoritma ini tetap memberikan hasil yang kompetitif dalam banyak kasus (Azizah, Goejantoro and Sifriyani, 2019).



Gambar 3. 2. Confusion matrix Naive Bayes

3.3. Evaluasi

Kinerja kedua algoritma dievaluasi menggunakan metrik berikut:

a. Accuracy: Proporsi prediksi benar terhadap total prediksi.
   - Rumus Accuracy:

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

b. Precision: Proporsi prediksi positif yang benar.
   - Rumus Precision:

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

c. Recall: Kemampuan model dalam mengidentifikasi data positif.
   - Rumus Recall:

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

d. F1-Score: Keseimbangan mean dari precision dan recall.
   - Rumus F1-Score:

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

Tabel 3. 2. Tabel evaluasi

Model	Class	Precision	Recall	F1-Score	Support
Naive Bayes	0	0.83	0.80	0.81	99
	1	0.66	0.71	0.68	55
	Accuracy			0.77	154

 ISSN 2987-484X

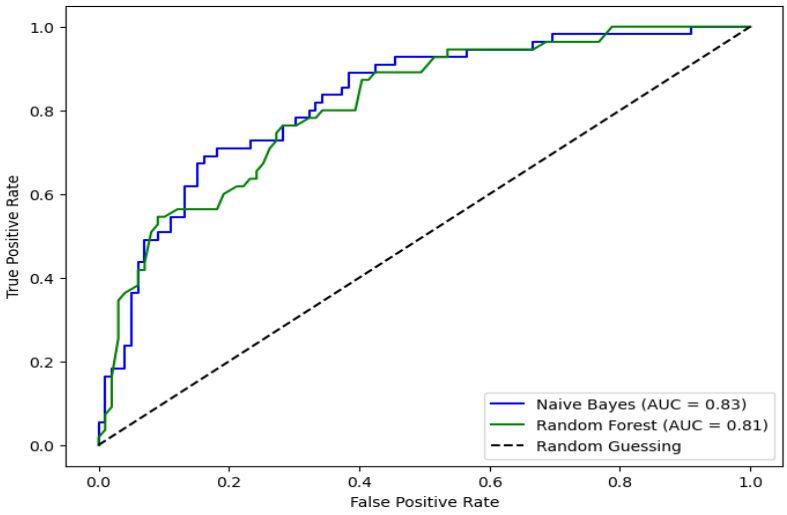
Model	Class	Precision	Recall	F1-Score	Support
	Macro Avg	0.75	0.75	0.75	154
	Weighted Avg	0.77	0.77	0.77	154
Random Forest	0	0.79	0.78	0.78	99
	1	0.61	0.62	0.61	55
	Accuracy			0.72	154
	Macro Avg	0.70	0.70	0.70	154
	Weighted Avg	0.72	0.72	0.72	154

e. ROC Curve: Grafik yang digunakan untuk menampilkan hubungan antara *True Positive Rate* (TPR) dan *False Positive* Rate (FPR).
- Rumus TPR:

$$TPR = \frac{TP}{TP + FN}$$

- Rumus FPR:

$$FPR = \frac{FP}{FP + TN}$$



Gambar 3. 3. ROC Curve

Berdasarkan analisa data di atasa bahwa Naive Bayes bekerja baik pada dataset sederhana dengan fitur yang saling independen, sementara Random Forest lebih cocok untuk menangani dataset yang lebih kompleks.

4. Kesimpulan dan Saran

4.1. Kesimpulan

a. Naive Bayes memberikan performa yang lebih baik dalam hal akurasi (77%) dibandingkan dengan Random Forest (72%) untuk memprediksi risiko diabetes. Naive Bayes unggul dalam mendeteksi kasus positif diabetes, sebagaimana tercermin dalam metrik Precision, Recall, dan F1-score yang lebih tinggi dibandingkan Random Forest.
b. Berdasarkan analisis ROC Curve, *Area Under the Curve* (AUC) Naive Bayes memiliki nilai AUC yang sedikit lebih tinggi dibandingkan Random Forest. Nilai AUC yang lebih besar mencerminkan kemampuan model yang lebih baik dalam

membedakan antara pasien dengan diabetes (positif) dan tanpa diabetes (negatif) pada berbagai ambang batas prediksi.

c. Hasil ini mencerminkan kekuatan masing-masing algoritma: Naive Bayes bekerja baik pada dataset sederhana dengan fitur yang saling independen, sementara Random Forest lebih cocok untuk menangani dataset yang lebih kompleks.

4.2. Saran

a. Naive Bayes dapat dioptimalkan melalui eksplorasi seperti Bernoulli Naive Bayes atau Multinomial Naive Bayes, yang mungkin lebih sesuai untuk karakteristik data.
b. Random Forest dapat ditingkatkan performanya melalui hyperparameter tuning, misalnya dengan menyesuaikan jumlah pohon, kedalaman pohon, atau jumlah fitur yang dipertimbangkan pada setiap split.
c. Penggunaan algoritma lain seperti Gradient Boosting atau XGBoost dapat dieksplorasi untuk dibandingkan dengan Naive Bayes dan Random Forest.

Daftar Pustaka

World Health Organization. (2024). *Diabetes*. https://www.who.int/news-room/fact-sheets/detail/diabetes

Azizah, N., Goejantoro, R. and Sifriyani (2019) 'Metode Naive Bayes Dengan PendekatanDistribusi Gauss Untuk Klasifikasi PeminatanPeserta Didik', *Prosiding Seminar Nasional Matematika dan Statistika*, 1, pp. 1–7. Available at: https://jurnal.fmipa.unmul.ac.id/index.php/SNMSA/article/view/520/217.

Dhawas, P. *et al.* (2024) 'Diabetes Detection using Machine Learning', *15th International Conference on Advances in Computing, Control, and Telecommunication Technologies, ACT 2024*, 2(06), pp. 3103–3110.

Erlin *et al.* (2022) 'Deteksi Dini Penyakit Diabetes Menggunakan Machine Learning dengan Algoritma Logistic Regression', *Jurnal Nasional Teknik Elektro dan Teknologi Informasi*, 11(2), pp. 88–96. Available at: https://doi.org/10.22146/jnteti.v11i2.3586.

Kamel, H., Abdulah, D. and Al-Tuwaijari, J.M. (2019) 'Cancer Classification Using Gaussian Naive Bayes Algorithm', *Proceedings of the 5th International Engineering Conference, IEC 2019*, pp. 165–170. Available at: https://doi.org/10.1109/IEC47844.2019.8950650.

Kemenkes (2018) 'Laporan Riskesdas 2018 Nasional.pdf', *Lembaga Penerbit Balitbangkes*, p. hal 156.

Lin, W. *et al.* (2017) 'An ensemble random forest algorithm for insurance big data analysis', *IEEE Access*, 5(JULY), pp. 16568–16575. Available at: https://doi.org/10.1109/ACCESS.2017.2738069.

Nasional, J. *et al.* (2024) 'Komparasi Algoritma Naïve Bayes dan Gradient Boosting untuk Prediksi Pasien Diabetes', 02, pp. 118–125.

Nassar, M. *et al.* (2021) 'Diabetes Mellitus and COVID-19: Review Article', *Diabetes and Metabolic Syndrome: Clinical Research and Reviews*, 15(6). Available at: https://doi.org/10.1016/j.dsx.2021.102268.

Webber, S. (2013) International *Diabetes Federation, Diabetes Research and Clinical Practice*. Available at: https://doi.org/10.1016/j.diabres.2013.10.013.