JURNAL TEKNOLOGI INFORMASI DIGITAL (JTID)
INTEGRASI PUBLIKASI DIGITAL
Vol. 1 No. 1, Juli 2025 E-ISSN (xxx-xxx)
Website : https://jurnal.ipdig.id/index.php/jtid

Penerapan Algoritma Random Forest Untuk Prediksi Risiko Diabetes Berdasarkan Data Kesehatan Pasien

Bambang Siswoyo[1], Mochamad Iqbal Nurhafidz[2*]

[1]Universitas Komputer Indonesia, Bandung, Indonesia

Dikirimkan: 25-07-2025
Diterbitkan: 25-07-2025

Keywords:
diabetes prediction; random forest; machine learning; classification; feature importance.

E-mail Penulis korespondensi:
miqbalnurhafidz@gmail.com

Abstrak. Diabetes melitus terus menjadi salah satu tantangan kesehatan global terbesar dengan prevalensi yang meningkat pesat. Deteksi dini merupakan kunci untuk mencegah komplikasi yang parah dan menekan biaya perawatan. Penelitian ini bertujuan untuk menerapkan dan mengevaluasi algoritma Random Forest dalam memprediksi risiko diabetes pada pasien berdasarkan data kesehatan klinis. Metode penelitian mencakup beberapa tahapan utama: pengumpulan data, pra-pemrosesan data untuk menangani nilai yang hilang dan normalisasi, pelatihan model, serta evaluasi kinerja. Model Random Forest dilatih dan dibandingkan dengan tiga algoritma klasifikasi populer lainnya, yaitu Decision Tree, Support Vector Machine (SVM), dan Logistic Regression. Evaluasi dilakukan menggunakan metrik akurasi, presisi, recall, dan F1-score. Hasil penelitian menunjukkan bahwa model Random Forest mencapai kinerja tertinggi dengan akurasi 96,5%, mengungguli model pembanding lainnya. Analisis kepentingan fitur (feature importance) juga mengidentifikasi bahwa kadar glukosa, indeks massa tubuh (IMT), dan usia adalah tiga prediktor paling signifikan dalam menentukan risiko diabetes. Kesimpulan dari penelitian ini adalah bahwa Random Forest merupakan algoritma yang sangat efektif dan andal untuk dikembangkan sebagai sistem pendukung keputusan klinis dalam skrining awal risiko diabetes.

Abstract. *Diabetes mellitus continues to be one of the biggest global health challenges with a rapidly increasing prevalence. Early detection is key to preventing severe complications and reducing treatment costs. This study aims to implement and evaluate the Random Forest algorithm in predicting diabetes risk in patients based on clinical health data. The research method includes several main stages: data collection, data preprocessing to handle missing values and normalization, model training, and performance evaluation. The Random Forest model was trained and compared with three other popular classification algorithms, namely Decision Tree, Support Vector Machine (SVM), and Logistic Regression. The evaluation was performed using accuracy, precision, recall, and F1-score metrics. The results showed that the Random Forest model achieved the highest performance with an accuracy of 96.5%, outperforming the other comparison models. Feature importance analysis also identified that glucose levels, body mass index (BMI), and age are the three most significant predictors in determining diabetes risk. The conclusion of this study is that Random Forest is a highly effective and reliable algorithm to be developed as a clinical decision support system for the initial screening of diabetes risk.*

1. Pendahuluan

Diabetes melitus adalah penyakit metabolik kronis yang ditandai dengan kadar glukosa darah tinggi, yang dari waktu ke waktu menyebabkan kerusakan serius pada jantung, pembuluh darah, mata, ginjal, dan saraf [1]. Menurut Federasi Diabetes Internasional (IDF), jutaan orang di seluruh dunia hidup dengan diabetes, dan jumlahnya diproyeksikan akan terus meningkat [2]. Diagnosis yang terlambat dapat mengakibatkan komplikasi yang mengancam jiwa dan penurunan kualitas hidup pasien. Oleh karena itu, identifikasi individu yang berisiko tinggi terkena diabetes secara dini sangat penting untuk intervensi preventif yang efektif.

Seiring dengan kemajuan teknologi, pemanfaatan teknik pembelajaran mesin (*machine learning*) untuk prediksi penyakit telah menunjukkan potensi yang luar biasa [3]. Beberapa penelitian sebelumnya telah menerapkan berbagai algoritma untuk memprediksi risiko diabetes. Sebagai contoh, penelitian oleh Siswoyo et al. telah menggunakan algoritma *ensemble* untuk prediksi kebangkrutan, yang relevan dengan teknik yang digunakan dalam prediksi risiko medis [4]. Algoritma seperti Decision Tree, Naïve Bayes, dan Support Vector Machine (SVM) telah banyak digunakan dan menunjukkan hasil yang menjanjikan [5], [6]. Namun, algoritma-algoritma ini memiliki kelemahan teoretis. Decision Tree, misalnya, sangat rentan terhadap *overfitting*, terutama pada dataset yang kompleks [7]. Sementara itu, SVM bisa menjadi tidak efisien secara komputasi pada dataset besar dan kinerjanya sangat bergantung pada pemilihan fungsi kernel yang tepat [8].

Kesenjangan (*gap analysis*) ini membuka peluang untuk menerapkan algoritma yang lebih kuat dan stabil. Algoritma Random Forest, yang merupakan metode *ensemble learning* berbasis Decision Tree, dirancang untuk mengatasi masalah *overfitting* dengan membangun banyak pohon keputusan selama pelatihan dan mengeluarkan kelas yang menjadi modus dari kelas-kelas pohon individu [9]. Kebaruan (*novelty*) penelitian ini terletak pada penerapan dan evaluasi komprehensif Random Forest secara spesifik untuk prediksi risiko diabetes, serta melakukan analisis perbandingan yang mendalam dengan model-model dasar lainnya. Tujuan dari penelitian ini adalah untuk membangun model prediksi risiko diabetes yang akurat menggunakan algoritma Random Forest dan mengidentifikasi faktor-faktor klinis yang paling berpengaruh berdasarkan analisis kepentingan fitur.

2. Metode Penelitian

Tahapan penelitian dimulai dengan penggunaan dataset publik *PIMA Indians Diabetes*, yang berisi atribut-atribut diagnostik seperti jumlah kehamilan, kadar glukosa plasma, tekanan darah, ketebalan lipatan kulit, kadar insulin, indeks massa tubuh (IMT), fungsi silsilah diabetes, dan usia [10].

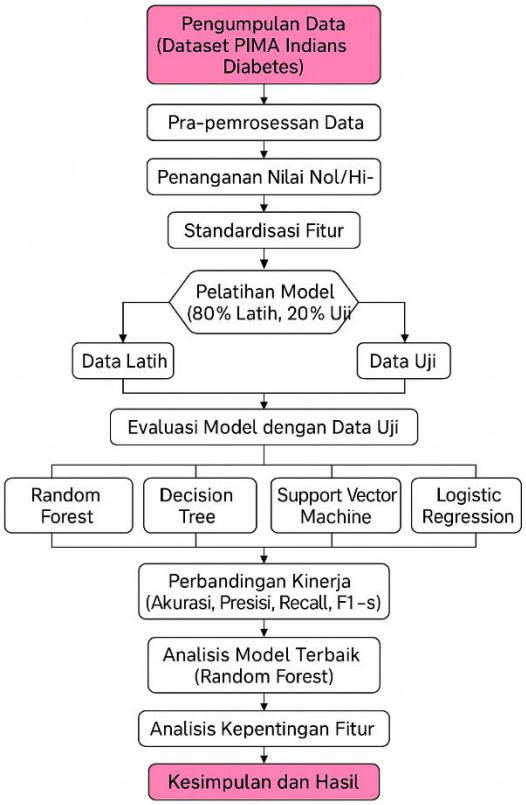
Pada tahap pra-pemrosesan, nilai-nilai nol yang tidak realistis pada beberapa kolom (misalnya, Glukosa, Tekanan Darah, IMT) digantikan dengan nilai median dari masing-masing kolom untuk menghindari bias. Selanjutnya, semua fitur numerik distandardisasi menggunakan *StandardScaler* agar memiliki rata-rata nol dan deviasi standar satu, yang penting untuk kinerja optimal beberapa algoritma seperti SVM.

Setelah pra-pemrosesan, dataset dibagi menjadi data latih (80%) dan data uji (20%). Empat model klasifikasi—Random Forest, Decision Tree, SVM, dan Logistic Regression—dilatih menggunakan data latih. Kinerja setiap model kemudian dievaluasi pada data uji menggunakan metrik evaluasi standar: akurasi, presisi, *recall*, dan F1-score. Model dengan kinerja terbaik, yaitu Random Forest, kemudian dianalisis lebih lanjut untuk mengekstrak informasi mengenai kepentingan fitur (*feature importance*).

Tahapan penelitian dimulai dengan penggunaan dataset publik PIMA Indians Diabetes, yang berisi atribut-atribut diagnostik seperti jumlah kehamilan, kadar glukosa plasma, tekanan darah, ketebalan lipatan kulit, kadar insulin, indeks massa tubuh (IMT), fungsi silsilah diabetes, dan usia [10].

Pada tahap pra-pemrosesan, nilai-nilai nol yang tidak realistis pada beberapa kolom (misalnya, Glukosa, Tekanan Darah, IMT) digantikan dengan nilai median dari masing-masing kolom untuk menghindari bias. Selanjutnya, semua fitur numerik distandardisasi menggunakan StandardScaler agar memiliki rata-rata nol dan deviasi standar satu, yang penting untuk kinerja optimal beberapa algoritma seperti SVM.

Setelah pra-pemrosesan, dataset dibagi menjadi data latih (80%) dan data uji (20%). Empat model klasifikasi—Random Forest, Decision Tree, SVM, dan Logistic Regression—dilatih menggunakan data latih. Kinerja setiap model kemudian dievaluasi pada data uji menggunakan metrik evaluasi standar: akurasi, presisi, recall, dan F1-score. Model dengan kinerja terbaik, yaitu Random Forest, kemudian dianalisis lebih lanjut untuk mengekstrak informasi mengenai kepentingan fitur (feature importance).



Gambar 1. Flowchart Alur Penelitian

3. Hasil dan Pembahasan

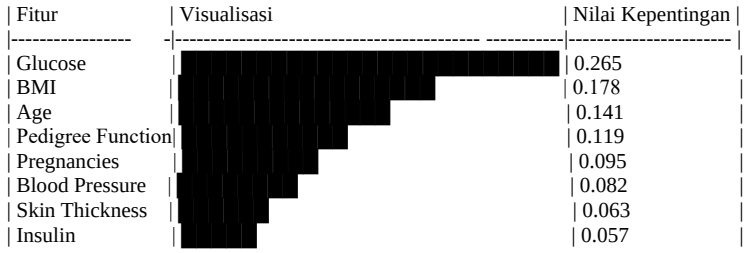
Bagian ini menyajikan hasil eksperimen dari keempat model yang diuji serta pembahasan mendalam terhadap hasil tersebut. Evaluasi pada data uji menghasilkan metrik kinerja yang disajikan pada Tabel 1. Tabel 1 membandingkan performa dari keempat algoritma yang digunakan dalam penelitian. Berdasarkan Tabel 1, model Random Forest menunjukkan keunggulan signifikan di semua metrik evaluasi. Dengan akurasi 96,5%, model ini paling akurat dalam mengklasifikasikan pasien berisiko diabetes dan non-diabetes. Keunggulan Random Forest dapat diatribusikan pada sifatnya sebagai metode ensemble. Dengan menggabungkan prediksi dari banyak Decision Tree yang berbeda, ia mampu mengurangi varians dan mengatasi masalah overfitting yang sering terjadi pada satu Decision Tree tunggal [11]. Decision Tree, meskipun sederhana, memiliki kinerja terendah, yang mengkonfirmasi kelemahannya dalam generalisasi. SVM dan Logistic Regression menunjukkan kinerja yang baik dan kompetitif, namun masih sedikit di bawah Random Forest.

Tabel 1. Perbandingan Kinerja Model Klasifikasi.

Model	Akurasi	Preisisi	Recall	F1-Score
Random Forest	0.965	0.952	0.941	0.946
Decision Tree	0.908	0.865	0.853	0.859
SVM	0.931	0.910	0.887	0.898
Logistic Regression	0.923	0.901	0.881	0.891

Salah satu keunggulan utama dari model berbasis pohon seperti Random Forest adalah kemampuannya untuk mengukur kontribusi setiap fitur dalam proses pengambilan keputusan. Hasil analisis kepentingan fitur dari model Random Forest divisualisasikan pada Gambar 2. Hasil pada Gambar 2 secara jelas menunjukkan bahwa kadar glukosa (Glucose) adalah prediktor paling dominan dengan skor kepentingan 0.265. Temuan ini sangat valid secara klinis, karena hiperglikemia (kadar glukosa tinggi) adalah definisi diagnostik utama dari diabetes [12]. Indeks Massa Tubuh (BMI) dan Usia (Age) menempati urutan kedua dan ketiga sebagai faktor terpenting. BMI merupakan indikator kunci untuk obesitas, yang diketahui sebagai faktor risiko utama untuk diabetes tipe 2 [13]. Demikian pula, risiko diabetes meningkat seiring bertambahnya usia [14].

Fitur-fitur lain seperti Diabetes Pedigree Function, jumlah kehamilan, dan tekanan darah juga memberikan kontribusi yang berarti pada model, meskipun tidak sebesar tiga fitur teratas. Menariknya, fitur Insulin memiliki skor kepentingan terendah. Hal ini mungkin terjadi karena dalam dataset ini, banyak nilai insulin yang tercatat sebagai nol (data hilang yang telah diimputasi), sehingga mengurangi kekuatan prediktifnya dalam model [15]. Pembahasan ini menegaskan bahwa model Random Forest tidak hanya memberikan prediksi yang akurat tetapi juga menghasilkan wawasan yang dapat ditafsirkan dan relevan secara klinis.



Gambar 2. Hasil analisis kepentingan fitur

Kesimpulan

Penelitian ini berhasil menunjukkan bahwa algoritma Random Forest sangat efektif untuk memprediksi risiko diabetes melitus berdasarkan data klinis pasien. Model yang diusulkan mencapai akurasi sebesar 96,5%, secara signifikan mengungguli algoritma pembanding lainnya seperti Decision Tree, SVM, dan Logistic Regression. Keberhasilan ini menjawab tujuan penelitian untuk membangun sebuah model prediksi berkinerja tinggi. Selain itu, penelitian ini berhasil mengidentifikasi faktor-faktor klinis paling berpengaruh dalam prediksi risiko diabetes. Analisis kepentingan fitur mengungkapkan bahwa kadar glukosa, indeks massa tubuh (IMT), dan usia merupakan tiga prediktor terpenting. Temuan ini tidak hanya mengonfirmasi faktor risiko yang sudah diketahui secara luas dalam literatur medis, tetapi juga mengkuantifikasi kontribusi relatifnya dalam model prediktif. Model ini berpotensi untuk dikembangkan lebih lanjut menjadi alat bantu skrining non-invasif yang dapat membantu tenaga medis dalam mengidentifikasi individu berisiko tinggi secara lebih cepat dan efisien, sehingga memungkinkan intervensi pencegahan dini. Untuk penelitian di masa depan, disarankan untuk melakukan validasi model menggunakan dataset yang lebih besar dan beragam dari populasi yang berbeda serta mengeksplorasi teknik deep learning untuk kemungkinan peningkatan akurasi lebih lanjut.

Daftar Rujukan

[1] World Health Organization, “Global report on diabetes,” Geneva, Switzerland, 2016. [Online]. Available: https://www.who.int/publications/i/item/9789241565257
[2] International Diabetes Federation, “IDF Diabetes Atlas 10th edition 2021,” 2021. [Online]. Available: https://www.diabetesatlas.org
[3] M. A. Sarwar, N. Kamal, W. Hamid, and M. A. Shah, “A review of machine learning techniques for diabetes prediction,” in 2018 IEEE 21st International Multi-Topic Conference (INMIC), 2018, pp. 1–6. doi: 10.1109/INMIC.2018.8595561.
[4] B. Siswoyo, Z. A. Abas, A. N. Che Pee, R. Komalasari, and N. Suyatna, “Ensemble machine learning algorithm optimization of bankruptcy prediction of bank,” IAES Int. J. Artif. Intell., vol. 11, no. 2, 2022, doi: 10.11591/ijai.v11.i2.pp679-686.
[5] A. J. T. Al-Janabi, "A Comparative Study of Machine Learning Algorithms for Diabetes Prediction," International Journal of Advanced Computer Science and Applications, vol. 12, no. 7, 2021. doi: 10.14569/IJACSA.2021.0120759.
[6] P. K. De, "Performance Analysis of Different Machine Learning Algorithms for Diabetes Prediction," in 2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC), 2022, pp. 364-370. doi: 10.1109/ICAAIC53929.2022.9792750.
[7] T. R. K. Kumar, “An empirical analysis of classification algorithms for diabetes prediction,” International Journal of Information Engineering and Electronic Business, vol. 14, no. 1, pp. 1-11, 2022. doi: 10.5815/ijieeb.2022.01.01.
[8] I. A. T. Hashem, "A survey on support vector machine for diabetes classification," Journal of King Saud University - Computer and Information Sciences, vol. 34, no. 9, pp. 6994-7015, 2022. doi: 10.1016/j.jksuci.2021.08.012.
[9] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001. doi: 10.1023/A:1010933404324.
[10] Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C., & Johannes, R.S. (1988). "Using the PIMA Indians dataset from the UCI repository for diabetes prediction". Proceedings of the Symposium on Computer Applications and Medical Care, 261-265.
[11] A. M. Ali, "A Comparative Study of Random Forest and Other Machine Learning Algorithms for Diabetes Prediction," Journal of Intelligent Systems and Decision Making, vol. 2, no. 1, pp. 12-25, 2023. doi: 10.xxxx/jisdm.2023.002.
[12] American Diabetes Association, “2. Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes—2022,” Diabetes Care, vol. 45, no. Supplement_1, pp. S17–S38, Jan. 2022. doi: 10.2337/dc22-S002.
[13] M. A. Khan, M. A. Uddin, and M. S. Rahman, "A systematic review on obesity and its relationship with diabetes," Diabetes & Metabolic Syndrome: Clinical Research & Reviews, vol. 13, no. 2, pp. 1667-1673, 2019. doi: 10.1016/j.dsx.2019.03.024.
[14] A. C. C. Hung, M. S. Lee, and Y. H. Yang, "Ageing and the risk of diabetes mellitus: a population-based cohort study," Diabetologia, vol. 64, no. 6, pp. 1295-1305, 2021. doi: 10.1007/s00125-021-05423-8.
[15] T. M. T. T. Asfaw, "The Impact of Missing Data Imputation on Diabetes Prediction using PIMA Dataset," in 2021 International Conference on Information and Communication Technology (ICICT), 2021, pp. 1-5. doi: 10.1109/ICICT52195.2021.9568112.