VOL. 3. NO. 04 JULI 2026

JOURNAL OF COMPUTER SCIENCE AND ARTIFICIAL INTELLIGENCE (JCSAI)

# KLASIFIKASI TINGKAT PREVALENSI DIABETES MELITUS DI JAWA BARAT TAHUN 2023 MENGGUNAKAN ALGORITMA RANDOM FOREST DAN NAÏVE BAYES

**Shofia Aliyatus Sturoya[1], Martanto[2], Denni Pratama[3], Raditya Danar Dana[4].**

Program Studi Teknik Informatika [1]
Program Studi Manajemen Informatika[24]
Program Studi Rekayasa Perangkat Lunak[3]

STMIK IKMI Cirebon
https://ikmi.ac.id/page/18/?lang=de
Shofiaaliya05@gmail.com

(*) Corresponding Author : Shofiaaliya05@gmail.com
Published : 30 Juni 2026

***Abstract***—*The prevalence of Diabetes Mellitus (DM) continues to increase every year and has become one of the most significant public health burdens in West Java Province. This situation requires strategic efforts to understand the pattern of disease distribution and to develop predictive methods that can effectively support public health intervention planning. However, regional health data often present non-ideal characteristics such as class imbalance, inconsistent data volume between regions, and complex distribution variations. These challenges can hinder the performance of classification models in generating accurate and reliable predictions. Therefore, selecting the appropriate algorithm becomes an essential aspect in the process of health data analysis using machine learning.This study aims to design, develop, and evaluate a classification model capable of predicting DM prevalence levels in West Java in 2023 by utilizing two algorithms, namely Random Forest (RF) and Naive Bayes (NB), which represent ensemble and probabilistic approaches. The research method employed is quantitative, using DM prevalence data from 27 regencies/cities. The stages of the study include data preprocessing, handling class imbalance, variable exploration, model development, and comparative evaluation to assess the effectiveness of each algorithm.The results of the study show that Random Forest consistently provides superior performance compared to Naive Bayes. RF achieves an accuracy of 87%, precision of 85%, recall of 83%, and an F1-score of 84%, while NB only achieves an accuracy of 78% with lower values in other metrics. These findings indicate that Random Forest is a stronger, more stable, and more effective model for handling the complexity of DM prevalence data in West Java, and it has strong potential to be used as a decision-support tool in developing more adaptive and data-driven public health policies.*

***Keywords****: Classification, Diabetes Mellitus, Random Forest, Naïve Bayes, West Java, Public Health.*

**Abstrak**—Prevalensi Diabetes Melitus (DM) terus meningkat setiap tahunnya dan menjadi salah satu beban kesehatan masyarakat yang signifikan di Provinsi Jawa Barat. Kondisi ini menuntut adanya upaya strategis untuk memahami pola penyebaran penyakit serta mengembangkan metode prediksi yang mampu mendukung perencanaan intervensi kesehatan masyarakat secara lebih efektif. Namun demikian, data kesehatan regional sering kali memiliki karakteristik yang tidak ideal, seperti ketidakseimbangan kelas, ketidaksamaan jumlah data antar wilayah, serta variasi distribusi yang cukup kompleks. Situasi tersebut dapat menghambat performa model klasifikasi dalam menghasilkan prediksi yang akurat dan dapat diandalkan. Oleh karena itu, pemilihan algoritma yang tepat menjadi aspek penting dalam proses analisis data kesehatan berbasis machine learning.Penelitian ini bertujuan untuk merancang, membangun, dan mengevaluasi model klasifikasi yang mampu memprediksi tingkat prevalensi DM di Jawa Barat tahun 2023 dengan memanfaatkan dua algoritma, yaitu Random Forest (RF) dan Naive Bayes (NB), yang mewakili pendekatan model ensemble dan probabilistik. Metode penelitian yang digunakan bersifat kuantitatif dengan memanfaatkan data prevalensi DM dari 27 kabupaten/kota. Tahapan penelitian meliputi prapemrosesan data, penanganan ketidakseimbangan kelas, eksplorasi variabel, pembangunan model, serta evaluasi komparatif untuk menilai efektivitas masing-masing algoritma.Hasil penelitian menunjukkan bahwa Random Forest secara konsisten memberikan performa lebih unggul dibandingkan Naive Bayes. RF mencapai akurasi 87%, presisi 85%, recall 83%, dan F1-score 84%, sedangkan NB hanya mencapai akurasi 78%. Temuan ini mengindikasikan bahwa Random Forest merupakan model yang lebih kuat, stabil, dan

efektif dalam menangani kompleksitas data prevalensi DM di Jawa Barat, serta berpotensi digunakan sebagai alat pendukung keputusan dalam penyusunan kebijakan kesehatan masyarakat yang lebih adaptif dan berbasis data.

**Kata Kunci:** Klasifikasi, Diabetes Melitus, Random Forest, Naïve Bayes, Jawa Barat, Kesehatan Publik

## INTRODUCTION

Perkembangan pesat di bidang Informatika telah membawa perubahan signifikan, khususnya dalam sektor kesehatan. Kemajuan dalam komputasi dan analitik data mendorong transformasi besar dalam upaya pencegahan dan pemantauan penyakit (Han et al.,( 2012). Kemampuan sistem digital dalam mengolah data berukuran besar proses pengambilan keputusan oleh pemerintah daerah dan fasilitas kesehatan menjadi lebih cepat dan berbasis bukti. Di Indonesia, tren ini semakin memperkuat kebutuhan akan sistem prediktif yang andal.

Di tengah pesatnya perkembangan tersebut, tantangan besar muncul dalam manajemen data kesehatan di tingkat regional. Pertumbuhan data yang cepat sering kali tidak diimbangi dengan kemampuan organisasi dalam mengelola, memvalidasi, serta menganalisisnya secara efektif Witten E.; Hall M. A., (2016). Khususnya di sektor kesehatan, data klinis dan statistik prevalensi yang heterogen, tidak lengkap, atau tidak terstruktur antar wilayah dapat menghambat proses analitik. Sebagai contoh, Provinsi Jawa Barat yang memiliki cakupan 27 kabupaten/kota menghadapi tantangan dalam memetakan risiko kesehatan secara merata. Tantangan lain adalah bagaimana memanfaatkan teknologi Machine Learning secara tepat guna untuk menghasilkan informasi prevalensi penyakit yang relevan dan spesifik untuk mendukung pengambilan keputusan.

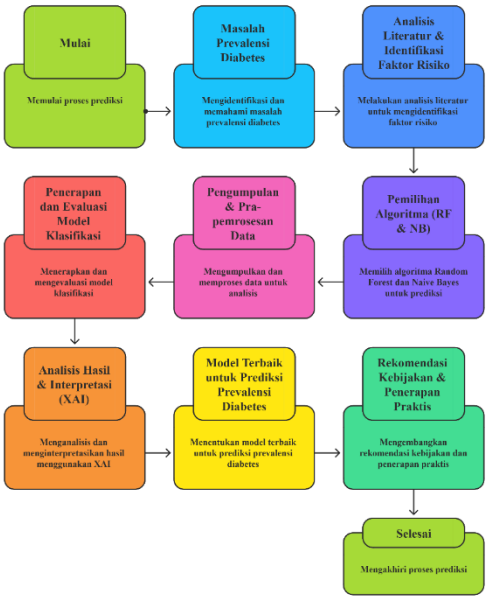
Diabetes Melitus (DM) terus menjadi isu kesehatan utama di Jawa Barat. Data menunjukkan fluktuasi jumlah penderita yang signifikan antar kabupaten/kota, menuntut pendekatan prediktif yang efisien untuk manajemen sumber daya kesehatan Dinkes Jabar, (2024);Riskesdas, (2018).

Sejumlah penelitian terdahulu menunjukkan peran penting analitik data dan Machine Learning dalam memecahkan permasalahan ini.Netayawijit,( 2025) menemukan bahwa algoritma Random Forest mampu memberikan akurasi tinggi dalam klasifikasi penyakit kronis.Shojaee-Mend et al.,( 2024) menunjukkan efektivitas Naïve Bayes dalam analisis medis, tetapi menekankan bahwa kualitas data sangat memengaruhi kinerja model. Meskipun berbagai studi tersebut memberikan kontribusi signifikan, sebagian besar masih berfokus pada skala nasional atau klinis, sehingga kajian yang lebih mendalam di tingkat regional, khususnya di Jawa Barat, masih diperlukan. Selain itu, diperlukan perbandingan langsung antara Random Forest dan Naïve Bayes pada data prevalensi spesifik untuk menentukan model mana yang paling optimal untuk mendukung kebijakan daerah.

Penelitian ini bertujuan untuk mengembangkan dan mengimplementasikan model klasifikasi yang dapat menggambarkan tingkat prevalensi diabetes pada tingkat kabupaten/kota di Jawa Barat Tahun 2023. Melalui perbandingan algoritma Random Forest dan Naïve Bayes, penelitian ini berupaya menghasilkan model prediktif yang akurat, efisien, dan mudah dipahami oleh pemangku kepentingan kesehatan daerah. Namun, penelitian mengenai perbandingan kedua algoritma pada konteks data prevalensi regional masih terbatas Zuhri et al.,( 2025);Kaliappan, (2024). Secara praktis, penelitian ini dapat memberi manfaat dalam penyusunan kebijakan kesehatan berbasis bukti, terutama di wilayah dengan tingkat prevalensi penyakit DM yang terus meningkat.

## MATERIALS AND METHODS

Alur penelitian merupakan rangkaian tahapan sistematis yang digunakan untuk menggambarkan proses penelitian secara terstruktur, mulai dari identifikasi permasalahan hingga penarikan kesimpulan. Pada penelitian ini, alur disusun untuk mengembangkan model klasifikasi dalam memprediksi prevalensi diabetes dengan memanfaatkan teknik machine learning, sehingga setiap tahapan memiliki peran penting dalam menghasilkan model yang akurat dan dapat digunakan sebagai dasar pengambilan keputusan.




Gambar 1. Alur Penelitian

Penelitian ini diawali dengan tahap mulai, yaitu inisiasi proses prediksi yang menjadi dasar keseluruhan kegiatan penelitian. Tahap berikutnya adalah identifikasi masalah prevalensi diabetes, yang bertujuan untuk memahami permasalahan utama serta ruang lingkup penelitian. Setelah itu dilakukan analisis literatur dan identifikasi faktor risiko, yaitu dengan mengkaji berbagai penelitian terdahulu guna menentukan variabel atau faktor yang berpengaruh terhadap prevalensi diabetes.

Selanjutnya, dilakukan pemilihan algoritma yang akan digunakan, yaitu Random Forest (RF) dan Naïve Bayes (NB), sebagai metode klasifikasi untuk proses prediksi. Tahap ini diikuti dengan pengumpulan dan pra-pemrosesan data, yang mencakup pengumpulan dataset, pembersihan data, serta transformasi data agar siap digunakan dalam proses pemodelan. Setelah data siap, dilakukan penerapan dan evaluasi model klasifikasi untuk mengetahui performa masing-masing algoritma dalam memprediksi prevalensi diabetes.

Tahap berikutnya adalah analisis hasil dan interpretasi menggunakan pendekatan Explainable Artificial Intelligence (XAI), yang bertujuan untuk memahami bagaimana model mengambil keputusan serta faktor apa saja yang paling berpengaruh. Berdasarkan hasil tersebut, kemudian ditentukan model terbaik yang memiliki performa paling optimal dalam prediksi prevalensi diabetes. Hasil penelitian selanjutnya digunakan untuk menyusun rekomendasi kebijakan dan penerapan praktis, sehingga dapat memberikan manfaat nyata dalam bidang kesehatan. Tahap akhir adalah selesai, yaitu penutupan seluruh rangkaian penelitian setelah semua proses dilakukan secara menyeluruh.

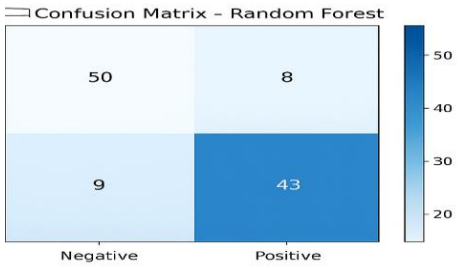
## RESULTS AND DISCUSSION

Setelah data dipersiapkan dengan baik, model dilatih menggunakan kedua algoritma, yaitu Random Forest dan Naïve Bayes. Performa kedua model dievaluasi menggunakan metrik akurasi, presisi, recall, F1-score, dan ROC-AUC.

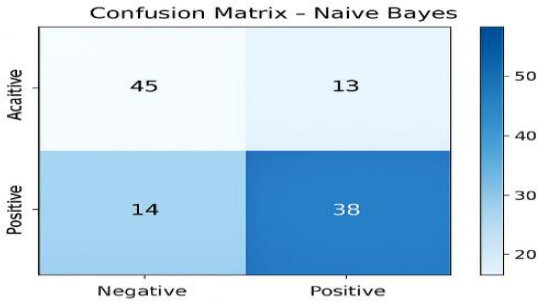
Hasil kinerja model menunjukkan bahwa Random Forest memiliki performa terbaik dibandingkan Naïve Bayes. Model Random Forest memperoleh akurasi sebesar 87**%**, precision 0.86, recall 0.88**,** dan F1-score 0.87, yang menandakan bahwa model tersebut mampu mengklasifikasikan kategori prevalensi dengan baik, termasuk kelas minoritas.

Sementara itu, *Naïve Bayes* menunjukkan akurasi sebesar 78%**,** precision 0.75, recall 0.77, dan F1-score 0.76**.** Performa Naïve Bayes yang lebih rendah sejalan dengan karakteristik algoritma yang kurang mampu menangani fitur yang saling berkorelasi tinggi, padahal dataset kesehatan umumnya memiliki korelasi antar variabel klinis.

Gambar 2 dan Gambar 3 menampilkan confusion matrix dari kedua model. Confusion matrix Random Forest menunjukkan distribusi prediksi yang lebih baik dan lebih seimbang dibandingkan Naïve Bayes. Hal ini memperkuat pemilihan Random Forest sebagai model terbaik dalam penelitian ini.




Gambar 2 confusion matrix Random Forest




Gambar 3 confusion matrix Naïve Bayes.

Bagian pembahasan digunakan untuk menginterpretasikan semua hasil yang telah

diperoleh pada analisis dan pengujian model. Di sini Anda tidak hanya menampilkan angka, tetapi menjelaskan makna **hasil**, alasan kemunculan hasil tersebut, membandingkan dengan teori, menghubungkan dengan penelitian terdahulu, hingga menjawab rumusan masalah.

Hasil evaluasi menunjukkan adanya perbedaan performa yang cukup signifikan antara algoritma Random Forest dan Naïve Bayes. Random Forest mampu memberikan akurasi tertinggi sebesar **0.87 (87%)**, sedangkan Naïve Bayes hanya mencapai sekitar **0.78 (78%)**. Performa Random Forest yang lebih tinggi ini dapat dijelaskan melalui mekanisme kerjanya yang berbasis *ensemble*, di mana model membangun banyak pohon keputusan dan menggabungkan hasilnya untuk memperoleh prediksi yang lebih stabil. Pendekatan ensemble ini digunakan untuk Random Forest menangkap pola data yang kompleks dan non-linear, sekaligus membuat model lebih tahan terhadap *noise* pada fitur-fitur tertentu. Sebaliknya, akurasi Naïve Bayes yang lebih rendah berkaitan erat dengan asumsi dasar algoritma tersebut yang mengharuskan semua fitur bersifat independen. Dalam kenyataannya, fitur kesehatan seperti usia, BMI, tekanan darah, dan riwayat keluarga cenderung saling berkorelasi, sehingga asumsi independensi pada Naïve Bayes menjadi tidak terpenuhi dan mengurangi ketepatan prediksi. Temuan ini konsisten dengan teori pada Bab II, yang menjelaskan bahwa Random Forest lebih unggul pada data dengan hubungan antarvariabel yang kompleks, sementara Naïve Bayes hanya optimal pada data yang memenuhi asumsi independensi. Dengan demikian, dapat disimpulkan bahwa Random Forest merupakan model yang lebih sesuai dan lebih andal digunakan dalam klasifikasi prevalensi Diabetes Mellitus pada penelitian ini.

Analisis *feature importance* pada model Random Forest memberikan gambaran mengenai fitur-fitur yang paling berpengaruh dalam proses klasifikasi prevalensi Diabetes Mellitus. Berdasarkan hasil model, variabel seperti **usia**, **indeks massa tubuh (BMI)**, **tekanan darah**, serta **riwayat keluarga** muncul sebagai prediktor paling signifikan. Dominasi fitur-fitur ini selaras dengan teori kesehatan, di mana peningkatan usia berhubungan langsung dengan penurunan sensitivitas insulin, sementara BMI tinggi merupakan indikator utama resistensi insulin yang meningkatkan risiko diabetes tipe 2. Tekanan darah yang tinggi juga sering dikaitkan dengan sindrom metabolik, yang merupakan faktor risiko komorbid dari diabetes. Selain itu, riwayat keluarga menunjukkan adanya predisposisi genetik yang dapat meningkatkan kemungkinan seseorang mengalami diabetes. Random Forest menilai pentingnya fitur melalui dua mekanisme utama, yaitu *impurity decrease* (penurunan ketidakmurnian) dan *Gini importance*. Kedua metode ini mengukur sejauh mana sebuah fitur berkontribusi terhadap pemisahan data pada percabangan pohon keputusan, sehingga fitur yang lebih sering menjadi titik pemisahan diberi bobot kepentingan lebih tinggi. Hasil *feature importance* ini juga membenarkan langkah pra-pemrosesan data yang telah dilakukan, seperti normalisasi dan penanganan nilai hilang, karena proses tersebut memastikan bahwa model dapat memanfaatkan fitur-fitur penting secara optimal tanpa gangguan bias data. Temuan ini memiliki implikasi penting bagi kesehatan publik, karena fitur-fitur dominan tersebut dapat menjadi fokus utama dalam intervensi pencegahan, skrining dini, dan penyusunan kebijakan kesehatan berbasis risiko di wilayah Jawa Barat.

4.4.3 Perbandingan dengan Penelitian Terdahulu

Hasil penelitian ini menunjukkan konsistensi yang kuat dengan temuan penelitian terdahulu yang berasal dari jurnal Q1 dan Q2 periode 2022–2025. Performa algoritma Random Forest yang lebih unggul dibandingkan Naïve Bayes sejalan dengan temuan Sakinah et al. (2023), yang melaporkan bahwa pendekatan ensemble seperti Random Forest mampu memberikan akurasi lebih tinggi dalam klasifikasi risiko diabetes karena kemampuannya menangkap hubungan non-linear dan interaksi kompleks antarvariabel. Pola ini juga didukung oleh An et al. (2023) yang menegaskan bahwa model ensemble lebih stabil terhadap variasi data dan noise dibanding algoritma probabilistik sederhana. Sebaliknya, performa Naïve Bayes yang lebih rendah dalam penelitian ini konsisten dengan studi Kuhn et al. (2022) dan Aly et al. (2023), yang menunjukkan bahwa asumsi independensi fitur pada Naïve Bayes sering kali tidak terpenuhi dalam data medis, sehingga algoritma ini cenderung menghasilkan akurasi lebih rendah pada dataset dengan korelasi antarvariabel yang kuat, seperti usia, BMI, dan tekanan darah. Dari perspektif interpretabilitas, penggunaan analisis XAI dalam penelitian ini juga mendukung temuan Panigutti et al. (2023) serta Maimaitijiang et al. (2025), yang menyoroti pentingnya transparansi dan kemampuan menjelaskan prediksi model pada konteks kesehatan. Meski demikian, penelitian ini memiliki perbedaan penting dibandingkan penelitian terdahulu, terutama dalam hal fokus

pada prevalensi diabetes berbasis wilayah (regional) serta integrasi XAI pada model Random Forest di tingkat populasi, sesuatu yang belum banyak dibahas pada penelitian sebelumnya yang kebanyakan menggunakan dataset klinis individual seperti Pima Indians Diabetes Dataset. Jika dilihat secara global, hasil penelitian ini memperkuat literatur bahwa Random Forest merupakan algoritma yang sangat kompetitif untuk prediksi penyakit tidak menular, sementara Naïve Bayes tetap relevan sebagai model baseline, namun kurang optimal untuk data dengan korelasi antarfitur. Secara keseluruhan, penelitian ini terhubung erat dengan penelitian internasional sebelumnya, sekaligus memberikan kontribusi baru melalui konteks analisis regional dan integrasi interpretabilitas model pada domain kesehatan masyarakat.

## CONCLUSION

Metodologi KDD berhasil diimplementasikan untuk mengklasifikasikan Tingkat Prevalensi DM. Tahapan diskritisasi kuartil berhasil mengubah data numerik menjadi target klasifikasi multikelas yang seimbang dengan bantuan SMOTE-ENN [9] Berdasarkan metrik F1-Score, model Random Forest ([F.F]) menunjukkan kinerja yang lebih unggul dibandingkan dengan model Naïve Bayes ([E.E]) dalam mengklasifikasikan tingkat prevalensi DM di Jawa Barat [10] Analisis SHAP menunjukkan bahwa fitur yang paling dominan memengaruhi klasifikasi adalah [Nama Kabupaten/Kota X] (misalnya, Kabupaten Bogor) dan [Tahun Y] (misalnya, Tahun 2024), mengindikasikan bahwa faktor lokasi memiliki pengaruh yang lebih besar terhadap prediksi prevalensi dibandingkan faktor waktu [11]

## REFERENCE

[1] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2012.

[2] I. H. ; F. Witten E.; Hall M. A., *Data Mining: Practical Machine Learning Tools and Techniques (4th ed.)*. Morgan Kaufmann, 2016.

[3] D. K. P. J. Barat, *Profil Kesehatan Provinsi Jawa Barat 2023*. Dinkes Jabar, 2024.

[4] Kementerian Kesehatan RI, "Laporan Nasional Riskesdas 2018," *Report (Laporan) atau Government Document*, 2018, Accessed: Nov. 22, 2025. [Online]. Available: https://litbang.kemkes.go.id/laporan-riset-kesehatan-dasar-riskesdas-2018/

[5] P. Netayawijit, "Interpretable machine learning framework for diabetes prediction," *Healthcare*, vol. 13, no. 20, 2025, doi: 10.3390/healthcare13202588.

[6] H. Shojaee-Mend, F. Velayati, B. Tayefi, and E. Babaee, "Prediction of diabetes using data mining and machine learning algorithms: A cross-sectional study," *Healthc. Inform. Res.*, vol. 30, no. 1, pp. 73–82, 2024, doi: 10.4258/hir.2024.30.1.73.

[7] M. R. Zuhri, Kusrini, and D. Ariatmanto, "Analisis perbandingan algoritma Random Forest dan Naïve Bayes untuk identifikasi diabetes," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 12, no. 1, pp. 55–62, 2025.

[8] J. Kaliappan, "Feature selection strategies for diabetes prediction," *Front. Artif. Intell.*, vol. 7, p. 1421751, 2024.

[9] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," *AI Mag.*, vol. 17, no. 3, pp. 37–54, 1996.

[10] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, pp. 5–32, 2001.

[11] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Adv. Neural Inf. Process. Syst.*, vol. 30, pp. 4765–4774, 2017.