OPEN ACCESS

Volume 10 (3), July 2026, 1250-1258

E-ISSN:2580-1643

Jurnal JTIK (Jurnal Teknologi Informasi dan Komunikasi)

DOI: https://doi.org/10.35870/jtik.v10i3.6379

Pemodelan *Hybrid* untuk Prediksi Risiko Keparahan Penyakit Tuberkulosis Menggunakan Algoritma *K-Means* dan *Random Forest*

Hasan Ibrohim [1*], Harminto Mulyo [2], Gentur Wahyu Nyipto Wibowo [3]

[1*,2,3] *Universitas Islam Nahdlatul Ulama Jepara, Kabupaten Jepara, Jawa Tengah, Indonesia.*

a r t i c l e i n f o

Article history:
Received 20 December 2025
Received in revised form
20 January 2026
Accepted 1 February 2026
Available online July 2026.

Keywords:
Tuberculosis; K-Means; Random Forest; Hybrid Model.

Kata Kunci:
Tuberkulosis; K-Means; Random Forest; Hybrid Model.

a b s t r a c t

Tuberculosis (TB) remains a major infectious disease in Indonesia, while the identification of patient severity levels in healthcare facilities is often time-consuming due to manual assessment of medical records. At Puskesmas Bonang 1, TB cases increased from 41 in 2023 to 57 in 2024, yet no data-driven analytical system is available to support rapid and objective risk evaluation. This study utilizes 2,546 TB patient medical records from 2023–2024 and applies preprocessing, normalization, encoding, clustering using K-Means, and the development of both baseline and hybrid models. The evaluation results indicate that the Hybrid K-Means + Random Forest model with hyperparameter tuning outperforms the standalone Random Forest model. The baseline Random Forest achieved an accuracy of 81.72% with an F1-Score of 80.98%, while the Hybrid + Tuning model obtained an accuracy of 82.51% and an F1-Score of 81.34%. This improvement demonstrates that cluster-based features extracted using K-Means successfully enhance data representation and improve the predictive performance of Tuberculosis severity risk classification.

a b s t r a k

Tuberkulosis (TB) masih menjadi penyakit menular dengan angka kasus tinggi di Indonesia, sementara proses identifikasi risiko keparahan pasien di fasilitas kesehatan sering berlangsung lama karena masih dilakukan secara manual. Di Puskesmas Bonang 1, jumlah pasien TB meningkat dari 41 kasus pada 2023 menjadi 57 kasus pada 2024, namun belum tersedia sistem analitik berbasis data untuk membantu penilaian risiko secara cepat dan objektif. Penelitian ini memanfaatkan 2.546 rekam medis pasien TB tahun 2023–2024 dan menerapkan tahapan preprocessing, normalisasi, encoding, clustering menggunakan K-Means, serta pelatihan model baseline dan model hybrid. Hasil evaluasi menunjukkan bahwa model Hybrid K-Means + Random Forest dengan hyperparameter tuning memperoleh kinerja yang lebih baik dibandingkan model Random Forest tunggal. Model baseline Random Forest mencapai akurasi sebesar 81,72% dengan nilai F1-Score 80,98%, sedangkan model Hybrid + Tuning mencapai akurasi 82,51% dan F1-Score 81,34%. Peningkatan performa ini menunjukkan bahwa penambahan fitur klaster hasil K-Means mampu memperkaya representasi data klinis dan meningkatkan kemampuan model dalam memprediksi risiko keparahan Tuberkulosis secara lebih optimal.

Corresponding Author. Email: hasanassafa@gmail.com [1].

1. Pendahuluan

Tuberkulosis (TB) tetap menjadi beban penyakit menular utama di tingkat global. Indonesia secara konsisten berada pada jajaran negara dengan prevalensi TB tertinggi di dunia. Fenomena lonjakan kasus serta variabilitas derajat keparahan pasien menuntut mekanisme penilaian klinis yang presisi demi menentukan prioritas intervensi secara akurat. Namun, penentuan risiko keparahan di fasilitas kesehatan primer, khususnya Puskesmas, masih didominasi oleh evaluasi klinis konvensional. Prosedur manual tersebut sangat bergantung pada subjektivitas serta kapasitas sumber daya manusia di setiap unit layanan, yang berisiko memicu disparitas penilaian antar petugas dan menghambat operasional manajemen program TB (Juliasih *et al.*, 2020). Kondisi tersebut teramati di Puskesmas Bonang 1, yang mencatat kenaikan jumlah pasien dari 41 jiwa pada 2023 menjadi 57 jiwa pada 2024. Meskipun beban pelayanan meningkat, unit kesehatan tersebut belum memiliki instrumen analitik untuk mendukung penilaian risiko secara objektif. Ketiadaan sistem pendukung keputusan berpotensi mengakibatkan keterlambatan penanganan pada pasien berisiko tinggi serta menurunkan efektivitas pengelolaan kasus secara keseluruhan.

Kemajuan *machine learning* menawarkan peluang transformasi dalam analisis klinis melalui arsitektur prediktif yang mampu mengolah data masif secara efisien. Penggunaan algoritma Random Forest telah terbukti efektif dalam klasifikasi derajat keparahan penyakit melalui rekam medis (Nasution & Juledi, 2025). Kendati demikian, implementasi model tunggal sering kali menemui kendala dalam mengidentifikasi struktur laten pada data klinis yang bersifat heterogen. Di sisi lain, algoritma K-Means lazim diterapkan untuk pengelompokan pasien berdasarkan karakteristik medis tertentu, namun sinergi langsungnya sebagai penambah atribut dalam model prediksi masih jarang dipraktikkan (Laksono *et al.*, 2024). Berangkat dari kesenjangan penelitian tersebut, studi ini mengusulkan arsitektur *hybrid* yang mengintegrasikan K-Means sebagai metode ekstraksi fitur dan Random Forest sebagai klasifikator. Sinergi kedua metode bertujuan memanfaatkan pola tersembunyi hasil klasterisasi untuk memperkuat representasi data, sehingga diperoleh hasil prediksi yang lebih presisi. Penelitian ini mengolah 2.546 rekam medis pasien TB dari Puskesmas Bonang 1 serta membandingkan performa model *hybrid* terhadap model standar Random Forest. Temuan dari studi ini diharapkan dapat memperkuat pengembangan sistem pendukung keputusan klinis di pusat kesehatan masyarakat.

2. Metodologi Penelitian

Studi menggunakan pendekatan kuantitatif dengan mengolah data rekam medis pasien Tuberkulosis (TB) yang bersumber dari Puskesmas Bonang 1. Arsitektur penelitian dirancang untuk mengonstruksi serta menguji performa model prediksi risiko keparahan berbasis *machine learning* melalui sinergi teknik ekstraksi fitur dan klasifikasi. Alur kerja sistematis dimulai dari akuisisi data, prapemrosesan, pembentukan klaster menggunakan K-Means, pengembangan model Random Forest, hingga tahap pengujian kinerja. Integrasi antara K-Means dan Random Forest diadopsi berdasarkan efektivitas metode tersebut pada pengolahan data kompleks, yang terbukti mampu meningkatkan presisi melalui pemanfaatan pola tersembunyi hasil klasterisasi dalam arsitektur pembelajaran berbasis *ensemble* (Rahman *et al.*, 2025).

Gambar 1. Skema Penelitian

Data yang dianalisis bersumber dari rekam medis digital pasien TB di Puskesmas Bonang 1 periode 2023–2024, mencakup total 2.546 entri. Setiap rekaman memuat atribut klinis numerik maupun kategorikal, meliputi usia, tekanan darah, berat badan, tinggi badan, hasil anamnesis, riwayat diagnosa, serta regimen pengobatan. Pemilihan variabel tersebut didasarkan pada relevansinya sebagai indikator kondisi klinis dalam menentukan derajat keparahan penyakit. Label target diklasifikasikan ke dalam tiga kategori risiko rendah, sedang, dan tinggi berdasarkan standar evaluasi medis pada fasilitas kesehatan terkait. Cakupan populasi selama dua tahun memberikan gambaran variasi kondisi klinis yang cukup representatif di lapangan. Namun, karakteristik data medis mentah (*raw data*) menuntut rangkaian prosedur

prapemrosesan guna meningkatkan kualitas serta validitas struktur data sebelum diintegrasikan ke dalam arsitektur pemodelan.

Prapemrosesan Data

Mekanisme prapemrosesan diterapkan untuk memastikan dataset berada pada kondisi optimal. Langkah sistematis yang meliputi pembersihan, transformasi, normalisasi, hingga penyandian variabel menjadi syarat mutlak agar data dapat diproses secara akurat oleh algoritma (Ramadhani *et al*., 2024). Prosedur ini krusial untuk mengeliminasi duplikasi, menangani nilai yang hilang, serta menyesuaikan format data dengan spesifikasi model prediktif (Mutawali *et al*., 2022). Identifikasi nilai kosong pada atribut numerik diatasi melalui teknik imputasi median. Metode tersebut dipilih karena ketangguhannya (*robustness*) terhadap pencilan (*outlier*) yang sering muncul pada variabel klinis dengan distribusi tidak normal. Imputasi median menjamin kelengkapan data tanpa harus mereduksi sampel secara masif, sehingga model tetap dilatih menggunakan dataset yang berkualitas (Airlangga, 2024). Selanjutnya, atribut kategorikal ditransformasikan melalui proses *encoding* ke dalam format numerik agar kompatibel dengan logika komputasi Random Forest dan K-Means. Standardisasi fitur numerik dilakukan menggunakan *StandardScaler* untuk menghasilkan distribusi dengan rata-rata nol dan deviasi standar satu. Langkah tersebut bertujuan menyetarakan skala antar atribut sehingga mencegah dominasi fitur tertentu dalam kalkulasi jarak, mengingat sensitivitas algoritma K-Means terhadap perbedaan skala data (Arif *et al*., 2025). Dataset kemudian dipisahkan menjadi data latih dan data uji dengan rasio 80:20 melalui metode *stratified split*. Teknik ini menjamin proporsi kelas target tetap konsisten pada kedua subset, sehingga mencegah bias terhadap kelas mayoritas dan memastikan evaluasi performa bersifat representatif (Umar *et al*., 2025). Seluruh alur prapemrosesan dikelola dalam *pipeline* terintegrasi; parameter hanya dipelajari dari data latih untuk kemudian diaplikasikan pada data uji guna menghindari kebocoran data (*data leakage*).

Ekstraksi Fitur

Pasca prapemrosesan, data diolah melalui tahap ekstraksi fitur menggunakan algoritma K-Means untuk mengidentifikasi struktur atau pola laten yang tidak terakomodasi oleh fitur asli. K-Means dipilih karena efisiensinya dalam mengelompokkan sampel berdasarkan kedekatan karakteristik numerik. Informasi hasil klasterisasi kemudian dikonversi menjadi fitur tambahan guna memperkaya representasi data klinis (Renny Afriany *et al*., 2025). Model K-Means dilatih secara eksklusif menggunakan data latih yang telah terstandarisasi. Label klaster yang dihasilkan diintegrasikan ke dalam *pipeline* klasifikasi untuk membantu model prediktif menangkap fenomena medis yang mungkin tersembunyi. Penggunaan teknik klasterisasi sebelum tahap klasifikasi sejalan dengan praktik *data mining* di sektor kesehatan yang menunjukkan bahwa pengelompokan data klinis mampu memberikan dukungan analitik yang lebih tajam (Pratiwi *et al*., 2025). Label klaster merepresentasikan segmentasi alami berdasarkan kemiripan profil pasien. Meskipun beroperasi secara tidak terawasi (*unsupervised*), atribut tambahan ini berfungsi memperluas dimensi input model. Dengan representasi data yang lebih kaya, arsitektur klasifikasi diharapkan mampu mempelajari pola kompleks secara lebih stabil dan menghasilkan akurasi yang lebih optimal (Bagus Pratama & Setiawan, 2024).

Pembangunan Model Baseline dan Hybrid

Pasca tahap ekstraksi fitur, penelitian dilanjutkan dengan mengonstruksi arsitektur prediksi melalui dua skema utama: model *baseline* dan model *hybrid*. Skema *baseline* hanya melibatkan atribut klinis asli, sementara skema *hybrid* mengintegrasikan label klaster K-Means guna memperkuat representasi data. Penerapan kedua pendekatan secara paralel bertujuan mengukur pengaruh penambahan informasi klaster terhadap performa klasifikasi secara sistematis (Kevin & Wijaya, 2025). Perbandingan tersebut krusial untuk mengevaluasi sejauh mana integrasi struktur laten mampu mengoptimalkan akurasi serta reliabilitas model prediktif. Model *baseline* dikembangkan menggunakan algoritma Random Forest yang mengolah fitur klinis hasil prapemrosesan tanpa melibatkan hasil klasterisasi. Algoritma ini dipilih karena ketangguhannya dalam mengelola data heterogen, toleransi terhadap pencilan (*outlier*), serta resistansi terhadap *overfitting* yang umum ditemukan pada dataset klinis kompleks. Keandalan Random Forest telah divalidasi dalam berbagai studi klasifikasi risiko penyakit kronis di Indonesia, di mana algoritma

tersebut mampu mengidentifikasi kelas target melalui atribut medis yang relevan secara presisi (Setiawan *et al.*, 2024). Pada penelitian ini, model *baseline* berfungsi sebagai standar pembanding untuk mengukur signifikansi peningkatan performa setelah fitur tambahan diintegrasikan. Arsitektur *hybrid* dibangun dengan basis algoritma Random Forest yang identik, namun diperkaya dengan atribut klaster hasil K-Means. Penambahan dimensi informasi ini memungkinkan model menangkap karakteristik pasien yang tidak terakomodasi sepenuhnya oleh fitur klinis konvensional.

Pendekatan serupa sering diaplikasikan pada diagnosa penyakit kritis untuk menghasilkan prediksi yang lebih kuat dan stabil di berbagai kelas (Riansyah *et al.*, 2025). Dengan mempertahankan algoritma yang sama, variasi performa yang muncul dapat diatribusikan secara langsung pada kontribusi fitur klaster dalam memperkaya proses pembelajaran model. Guna memitigasi ketidakseimbangan distribusi kelas target, parameter class_weight='balanced' diterapkan pada kedua model. Pengaturan tersebut memberikan bobot lebih tinggi pada kelas minoritas demi menjamin objektivitas prediksi. Selain itu, optimasi model *hybrid* dilakukan melalui *hyperparameter tuning* menggunakan *GridSearchCV* untuk menemukan konfigurasi parameter paling efisien (Agusyul & Firmansyah, 2023). Seluruh tahapan pelatihan dan pengujian dijalankan pada subset data yang sama untuk menjaga validitas metodologis.

Evaluasi Model

Kinerja klasifikasi diukur menggunakan empat metrik standar: *Accuracy*, *Precision*, *Recall*, dan *F1-Score*. Instrumen evaluasi tersebut dipilih karena mampu memotret efektivitas model secara menyeluruh, terutama pada dataset dengan sebaran kelas yang tidak merata. Analisis melalui metrik jamak memastikan bahwa penilaian tidak hanya terpaku pada ketepatan prediksi global, melainkan juga pada kemampuan model dalam mendeteksi kelas positif serta menjaga rasio kesalahan prediksi secara seimbang. Metodologi evaluasi ini selaras dengan praktik penelitian *machine learning* pada data kesehatan dengan tingkat kompleksitas tinggi (Nugroho *et al.*, 2025; Susanto *et al.*, 2025). Adapun formulasi metrik yang digunakan adalah sebagai berikut:

1) Accuracy
   Mengukur rasio prediksi benar terhadap keseluruhan populasi data (Helmiyah *et al.*, 2025).

$$accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)}$$

2) Precision
   Menilai tingkat akurasi model dalam mengidentifikasi sampel positif (Helmiyah *et al.*, 2025).

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

3) Recall
   Mengukur kemampuan arsitektur dalam menjaring seluruh kasus positif yang tersedia pada data aktual (Helmiyah *et al.*, 2025).

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

4) F1-Score
   Rata-rata harmonik antara *Precision* dan *Recall* yang memberikan nilai tunggal untuk menilai stabilitas performa model (Helmiyah *et al.*, 2025).

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

3. Hasil dan Pembahasan

Hasil

Karakteristik Dataset

Data penelitian mencakup 2.546 rekam medis pasien Tuberkulosis dari Puskesmas Bonang 1 periode 2023–2024. Terdapat 11 atribut yang memetakan profil klinis subjek, mulai dari informasi demografis, tekanan darah, data antropometri, hingga detail regimen pengobatan dan catatan anamnesis. Secara teknis, dataset mencerminkan kompleksitas data medis pada umumnya, ditandai dengan variabilitas skala antar atribut, adanya nilai kosong, serta format data yang heterogen. Karakteristik tersebut mengharuskan dilakukannya prosedur prapemrosesan sebelum tahap pemodelan dijalankan. Rangkaian prapemrosesan yang diimplementasikan meliputi pembersihan data, mitigasi nilai hilang, transformasi

kategorikal melalui penyandian, serta standarisasi fitur numerik. Langkah sistematis tersebut menjamin seluruh atribut berada pada kondisi optimal guna mendukung efektivitas ekstraksi fitur melalui K-Means maupun akurasi klasifikasi pada model utama. Penggunaan dataset yang berasal dari layanan kesehatan primer memberikan nilai relevansi tinggi karena merefleksikan profil pasien TB secara aktual. Hal tersebut krusial bagi pengembangan model prediksi yang ditujukan untuk implementasi klinis di lapangan. Kelengkapan variabel pendukung seperti data fisiologis dan riwayat terapi memungkinkan pengujian efektivitas arsitektur *hybrid* dalam mengidentifikasi pola laten sekaligus mengoptimalkan penggolongan risiko keparahan penyakit secara objektif.

Tabel 1. Atribut Dataset Tuberkolosis

No	Atribut	Deskripsi
1	Tanggal	Tanggal kunjungan pasien ke Puskesmas.
2	Nama_lengkap	Nama pasien (digunakan untuk identifikasi internal, tidak dianalisis secara langsung)
3	Alamat	Alamat domisili pasien.
4	Umur	Usia pasien pada saat kunjungan.
5	Systole	Tekanan darah bagian atas (mmHg).
6	Diastole	Tekanan darah bagian bawah (mmHg)
7	Berat	Berat badan (kg)
8	Tinggi	Tingi badan pasien (cm).
9	Diagnosa	Hasil diagnose penyakit pasien
10	Obat	Jenis obat yang diberikan kepada pasien
11	Amnesa	Catatan Riwayat keluhan atau kondisi klinis pasien

Prapemrosesan Data

Rangkaian prapemrosesan menghasilkan dataset yang terstruktur dan kompatibel dengan algoritma pembelajaran mesin. Setiap atribut diperiksa secara saksama guna mengidentifikasi nilai kosong, inkonsistensi format, maupun entri yang tidak valid. Baris data dengan kekosongan pada atribut numerik krusial seperti tekanan darah, berat badan, dan tinggi badan dieliminasi karena keterbatasan akurasi jika dilakukan rekonstruksi manual. Sebaliknya, kekosongan pada atribut kategorikal seperti jenis obat dan catatan anamnesis diatasi dengan imputasi modus. Langkah tersebut diambil guna menjaga stabilitas distribusi kategori sekaligus mencegah bias selama fase pelabelan dan klasifikasi. Pasca pembersihan, atribut kategorikal ditransformasikan melalui mekanisme penyandian (*encoding*), sementara nilai hilang pada fitur numerik lainnya diolah menggunakan imputasi median. Selanjutnya, standarisasi fitur numerik diterapkan untuk menyetarakan skala antar variabel. Prosedur tersebut krusial guna menghindari dominasi atribut tertentu saat perhitungan jarak pada algoritma K-Means. Seluruh tahapan prapemrosesan dieksekusi melalui *pipeline* terintegrasi yang dilatih secara eksklusif pada data latih guna menjamin validitas pengujian dan mencegah kebocoran informasi (*data leakage*).

Hasil Ekstraksi Fitur (K-Means)

Ekstraksi fitur melalui algoritma K-Means bertujuan mengidentifikasi struktur laten berdasarkan kemiripan karakteristik klinis pasien. Dimensi data yang digunakan dalam pengelompokan tidak hanya mencakup variabel numerik standar (usia, tekanan darah, berat badan, dan tinggi badan), tetapi juga diperluas dengan fitur representasi tekstual berupa panjang karakter pada kolom anamnesis dan obat. Integrasi fitur tambahan tersebut berfungsi untuk memetakan kompleksitas kondisi medis pasien yang sering kali tidak terakomodasi secara optimal oleh variabel numerik murni. Melalui pendekatan ini, model mampu mengonversi pola kemiripan antar pasien menjadi atribut baru yang memperkaya dataset. Atribut hasil klasterisasi berperan sebagai prediktor tambahan yang memberikan informasi mengenai posisi relatif seorang pasien dalam kelompok risiko tertentu, sehingga meningkatkan kapasitas diskriminasi model pada tahap klasifikasi akhir.

Tabel 2. Kategorisasi Klaster

No	Klaster	Kategorisasi Risiko
1	Klaster 0	Pasien dengan tekanan darah dan usia relatif rendah, serta panjang anamnesa dan terapi obat yang singkat, menunjukkan kondisi klinis yang relatif ringan
2	Klaster 1	Pasien dengan usia menengah, tekanan darah moderat, serta panjang anamnesa sedang.
3	Klaster 2	Pasien dengan tekanan darah cenderung tinggi dan panjang anamnesa lebih kompleks.

Interpretasi Kategorisasi Klaster

Hasil pengelompokan melalui algoritma K-Means memetakan pasien ke dalam segmen-segmen dengan karakteristik klinis yang homogen. Setiap klaster mencerminkan kesamaan profil fisiologis serta tingkat kompleksitas pada catatan anamnesis dan regimen terapi. Kategorisasi tersebut berfungsi sebagai instrumen interpretasi data sekaligus menjadi basis augmentasi fitur dalam arsitektur model *hybrid*. Temuan klasterisasi mengonfirmasi adanya variasi kondisi klinis yang nyata pada pasien TB. Perbedaan karakteristik antar kelompok ini memvalidasi bahwa data medis memiliki struktur yang dapat diekstraksi untuk memperkuat variabel prediktor dalam pemodelan risiko keparahan.

Komparasi Model Baseline dan Hybrid

Tahap pengembangan model melibatkan dua skema klasifikasi untuk menguji efektivitas integrasi fitur. Model *baseline* dibangun menggunakan algoritma Random Forest dengan mengandalkan fitur klinis asli, yang berperan sebagai representasi standar dalam prosedur klasifikasi konvensional. Sebagai perbandingan, model *hybrid* mengintegrasikan label klaster hasil K-Means ke dalam arsitektur Random Forest. Penambahan atribut tersebut dimaksudkan untuk memperkuat representasi data melalui identifikasi pola laten yang tidak terakomodasi oleh variabel klinis primer. Sinergi antara pembelajaran tak terawasi (*unsupervised*) dan terawasi (*supervised*) ini diharapkan mampu menghasilkan keputusan klasifikasi yang lebih presisi. Kedua arsitektur dievaluasi menggunakan pembagian dataset yang identik. Konsistensi pengujian dijamin melalui penerapan seluruh rangkaian prapemrosesan, ekstraksi fitur, dan pelatihan model di dalam satu *pipeline* terintegrasi. Dengan membatasi pembelajaran parameter hanya pada subset data latih, validitas hasil tetap terjaga dari risiko kebocoran data (*data leakage*). Guna memitigasi ketidakseimbangan distribusi kelas pada variabel target, parameter class_weight='balanced' diaktifkan pada kedua model. Pengaturan tersebut berfungsi mereduksi bias terhadap kelas mayoritas sekaligus mempertajam sensitivitas model dalam mengidentifikasi kelas minoritas. Detail konfigurasi parameter yang membedakan kedalaman dan kompleksitas antara model *baseline* dan *hybrid* dirangkum dalam tabel berikut.

Tabel 3. Konfigurasi Model

Parameter	Baseline	Hybrid Model
n_estimator	300	300
max_depth	default	20
Min_Sampel_Split	default	5

Evaluasi Model

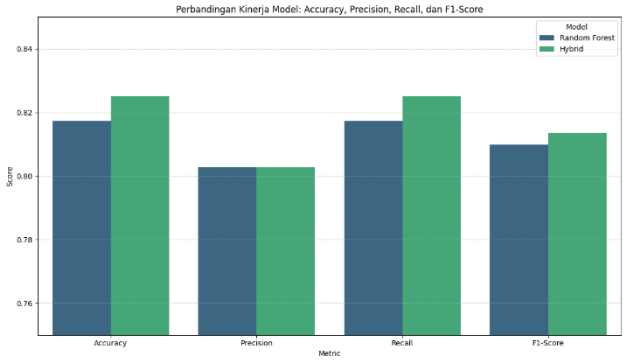
Evaluasi performa model dilakukan menggunakan empat metrik utama, yaitu accuracy, precision, recall, dan F1-score. Metrik-metrik ini dipilih karena mampu memberikan gambaran yang komprehensif mengenai kemampuan model dalam melakukan klasifikasi secara keseluruhan serta pada masing-masing kelas.

Tabel 4. Hasil Evaluasi Model

No	Model	Accuracy	Precission	Recall	F1-score
1	Baseline	0.8173	0.8027	0.8173	0.8098
2	Hybryd	0.8251	0.8027	0.8251	0.8134

Berdasarkan hasil evaluasi, model Hybrid K-Means + Random Forest menunjukkan performa yang lebih baik dibandingkan model Random Forest tunggal pada seluruh metrik evaluasi. Model baseline memperoleh akurasi sebesar 81,72%, sedangkan model hybrid mencapai akurasi 82,51%. Peningkatan

juga terlihat pada nilai F1-Score yang naik dari 80,98% menjadi 81,34%. Meskipun peningkatan performa relatif moderat, hasil ini menunjukkan bahwa fitur klaster hasil K-Means mampu memberikan informasi tambahan yang relevan dalam membedakan tingkat risiko keparahan pasien Tuberkulosis, terutama pada kasus-kasus dengan karakteristik klinis yang saling tumpang tindih.



Gambar 2. Perbandingan Kinerja Model

Visualisasi pada Gambar 2 mengonfirmasi keunggulan model Hybrid K-Means + Random Forest dibandingkan arsitektur Random Forest tunggal di mayoritas metrik evaluasi. Capaian paling signifikan teramati pada skor *Accuracy* dan *Recall*, yang mengalami kenaikan dari 0,817 menjadi 0,825. Peningkatan pada aspek *Recall* mengindikasikan bahwa integrasi fitur klaster memperkuat kemampuan model dalam mendeteksi kasus berisiko tinggi secara lebih akurat, yang merupakan faktor krusial dalam triase klinis. Stabilitas nilai *Precision* pada kedua model menandakan bahwa augmentasi fitur klaster tidak memicu peningkatan kesalahan prediksi positif (*false positive*). Namun, kenaikan skor *F1-Score* pada arsitektur hybrid membuktikan terciptanya keseimbangan yang lebih baik antara presisi dan sensitivitas. Hasil tersebut menegaskan bahwa model hybrid memiliki performa yang lebih optimal dan tangguh dalam memprediksi derajat keparahan Tuberkulosis dibandingkan pendekatan klasifikasi tunggal.

Pembahasan

Hasil eksperimen mengonfirmasi bahwa integrasi pola laten melalui teknik klasterisasi memberikan dampak positif terhadap akurasi prediksi risiko Tuberkulosis. Keunggulan model Hybrid K-Means + Random Forest, yang mencapai akurasi 82,51% dan *F1-Score* 81,34%, membuktikan bahwa atribut tambahan berupa label klaster mampu menutupi keterbatasan fitur klinis mentah dalam merepresentasikan kompleksitas kondisi pasien. Peningkatan performa ini terjadi karena algoritma K-Means berhasil mengelompokkan karakteristik pasien yang memiliki kemiripan struktur medis, sehingga Random Forest dapat memetakan batas keputusan (*decision boundary*) secara lebih presisi. Temuan tersebut sejalan dengan penelitian Rahman *et al.* (2025), yang menyatakan bahwa pemanfaatan informasi pola tersembunyi dari K-Means dalam model *ensemble* seperti Random Forest secara signifikan mampu meningkatkan kapabilitas klasifikasi pada data yang bersifat heterogen. Keberhasilan tahap prapemrosesan dalam menjaga integritas data juga menjadi faktor penentu stabilitas model. Penggunaan imputasi median dan standardisasi fitur memastikan bahwa algoritma tidak terdistorsi oleh pencilan atau dominasi skala variabel tertentu.

Pendekatan sistematis ini selaras dengan studi Ramadhani *et al.* (2024) dan Mutawali *et al.* (2022), yang menekankan bahwa kualitas data medis sangat bergantung pada eliminasi kebisingan (*noise*) dan penyesuaian format yang akurat. Selain itu, peningkatan nilai *Recall* pada model hybrid menunjukkan sensitivitas yang lebih tinggi dalam mendeteksi kelas risiko, yang merupakan aspek krusial dalam layanan kesehatan primer untuk menghindari keterlambatan penanganan medis. Efektivitas Random Forest dalam menangani variabel klinis yang kompleks ini juga didukung oleh temuan Setiawan *et al.* (2024) serta Riansyah *et al.* (2025), yang memvalidasi ketangguhan algoritma tersebut dalam menghasilkan prediksi yang stabil dan adil melalui penerapan bobot kelas yang seimbang. Secara keseluruhan, sinergi antara pembelajaran tak terawasi (*unsupervised*) dan terawasi (*supervised*) dalam penelitian ini membuktikan bahwa pengayaan fitur berbasis data (*data-driven feature augmentation*) lebih efektif dibandingkan hanya mengandalkan model klasifikasi tunggal. Penggunaan label klaster sebagai fitur tambahan memungkinkan model untuk mempelajari keterkaitan antar variabel klinis secara lebih mendalam. Prinsip penguatan representasi input ini konsisten dengan kajian Bagus Pratama & Setiawan (2024) serta Pratiwi *et al.* (2025), yang melaporkan bahwa segmentasi data melalui K-Means efektif dalam

membentuk profil risiko klinis yang lebih tajam, sehingga mendukung pengambilan keputusan medis yang lebih objektif dan terukur di fasilitas pelayanan kesehatan.

4. Kesimpulan dan Saran

Hasil penelitian menunjukkan bahwa integrasi K-Means sebagai metode feature extraction dengan Random Forest sebagai algoritma klasifikasi mampu meningkatkan performa prediksi risiko keparahan Tuberkulosis. Model Hybrid + Tuning memperoleh akurasi 82,51% dan F1-Score 81,34%, lebih tinggi dibandingkan model Random Forest tunggal yang mencapai akurasi 81,72% dan F1-Score 80,98%. Peningkatan ini menegaskan bahwa informasi klaster yang dihasilkan dari K-Means berhasil memperkaya representasi data klinis, sehingga model dapat mengenali pola laten yang tidak sepenuhnya tertangkap oleh fitur konvensional.

5. Ucapan Terima Kasih

Penulis menyampaikan apresiasi kepada Universitas Islam Nahdlatul Ulama Jepara, dosen pembimbing, serta rekan-rekan yang telah memberikan dukungan, arahan, dan motivasi sehingga penelitian ini dapat diselesaikan dengan baik.

6. Daftar Pustaka

Agusyul, A. Y., & Firmansyah, F. (2023). Prediksi Penyakit Jantung Menggunakan Algoritma Random Forest. *Jurnal Minfo Polgan*, *12*(2). https://doi.org/10.33395/jmp.v12i2.13214.

Airlangga, G. (2024). Leveraging Machine Learning for Accurate Anemia Diagnosis Using Complete Blood Count Data. *Indonesian Journal of Artificial Intelligence and Data Mining*, *7*(2), 318. https://doi.org/10.24014/ijaidm.v7i2.29869.

Arif, M., Setiawan, M., Dwi Hartono, A., Arif Ma, M., & Setiawan, R. (2025). Menggunakan Metode Machine Learning Untuk Memprediksi Nilai Mahasiswa Dengan Model Prediksi Multiclass. *Jurnal Informatika: Jurnal Pengembangan IT*, *10*(1).

Bagus Pratama, Y., & Setiawan, A. (2024). RESOLUSI: Rekayasa Teknik Informatika dan Informasi Implementasi Machine Learning Menggunakan Algoritma K-Means Untuk Klasifikasi Sekolah Dasar. *Media Online*, *4*(3).

Helmiyah, S., Pramestiawan, R., Studi Pendidikan Informatika, P., & Tinggi Keguruan dan Ilmu Pendidikan Rosalia Lampung, S. (2025). Analisis Komparatif Algoritma Machine Learning dengan Metrik Akurasi, Presisi, Recall, dan F1-Score pada Dataset Kacang Kering. *IKOMTI*, *6*(3), 152–159. https://doi.org/10.35960/ikomti.v6i3.2031.

Juliasih, N. N., Soedarsono, & Sari, R. M. (2020). Analysis of tuberculosis program management in primary health care. *Infectious Disease Reports*, *12*. https://doi.org/10.4081/idr.2020.8728.

Kevin, J., & Wijaya, B. A. (2025). Implementasi Algoritma Clustering dan Classification dalam Data Mining: Systematic Literature Review terhadap Tren dan Tantangan Terkini. *Jurnal Publikasi Sistem Informasi Dan Manajemen Bisnis*, *4*(3), 372–386. https://doi.org/10.55606/jupsim.v4i3.5240.

Laksono, B., Syahidin, Y., & Yunengsih, Y. (2024). Implementasi Data Mining Klasterisasi Data Pasien Rawat Inap dengan Algoritma K-Means Clustering. *Jurnal Teknologi Sistem Informasi Dan Aplikasi*, *7*(2), 621–627. https://doi.org/10.32493/jtsi.v7i2.39354.

Mutawali, L., Murniati, W., & Kunci, K. (2022). Penerapan KNNImputer dalam mengolah data missing value untuk membantu meningkatkan akurasi Support Vector Machine klasifikasi penyakit tiroid. *JINTEKS*, *4*(4).

Nasution, F. A., & Juledi, A. P. (2025). Penerapan Algoritma Random Forest untuk Klasifikasi Tingkat Keparahan Penyakit pada Data Rekam Medis. *Journal of Computer Science and Information Systems*, 371–378.

Nugroho, A. K., Hayati, L. N., & Jabir, S. R. (2025). Analisis Perbandingan Metode Naive Bayes dan Random Forest pada Klasifikasi Sentimen Publik terhadap Aplikasi Identitas Kependudukan Digital (IKD). *Jurnal Algoritma*, *22*(2). https://doi.org/10.33364/algoritma/v.22-2.2729.

Pratiwi, A., Manurung, H., Pita Uli Sitompul, M., Informasi, S., & Kaputama, S. (2025). Implementasi metode K-Means Clustering untuk mengklasifikasi status gizi balita berdasarkan wilayah kerja Puskesmas Karang Rejo. *Great Journal*.

Rahman, F. I., Lukman, & Hildayanti. (2025). Arus Jurnal Sains dan Teknologi (AJST) Sistem Klasifikasi Kerusakan Jalan Metode Machine Learning dengan Algoritma K-Means dan Random Forest. *AJST*, *3*(1).

Ramadhani, F., Septiana, D., Amalia, S. N., Fadilah, P. M., & Satria, A. (2024). Klasifikasi risiko gizi buruk pada ibu hamil menggunakan metode Random Forest. *Jurnal Teknologi Informasi*, *5*(2). https://doi.org/10.46576/djtechno.

Renny Afriany, Rudolf Sinaga, & Samsinar Samsinar. (2025). Segmentasi Pasien Berbasis K-Means dari Tanda-tanda Vital dan Demografi: Pendekatan Unsupervised Learning untuk Profil Risiko Klinis. *Jurnal Ilmiah Sistem Informasi Dan Ilmu Komputer*, *5*(3), 638–650. https://doi.org/10.55606/juisik.v5i3.1811.

Riansyah, M., Bahri, S., & Fiqna, H. P. (2025). Perbandingan Akurasi Algoritma Decision Tree dengan Random Forest dalam Diagnosa Penyakit Hepatitis. *Indonesian Journal of Science*, *2*(3).

Setiawan, A., Hadryan Nst, Z., Khairi, Z., & Efrizoni, L. (2024). Klasifikasi tingkat risiko diabetes menggunakan algoritma Random Forest. *Jurnal Informatika & Rekayasa Elektronika (JIRE)*, *7*(2).

Susanto, E. R., Inzaghi, M. R., Amarudin, A., & Neneng, N. (2025). Evaluasi Kinerja Model Random Forest Dalam Memprediksi Diabetes Berdasarkan Dataset Kesehatan di Indonesia. *Jurnal Pendidikan Dan Teknologi Indonesia*, *5*(7), 1857–1866. https://doi.org/10.52436/1.jpti.871.

Umar, M., Adytia, P., Yusnita, A., Widya Cipta Dharma, S., Yamin No, J. M., Kelua, G., Samarinda Ulu, K., Samarinda, K., & Timur, K. (2025). Penerapan Algoritma Support Vector Machine (SVM) dalam Analisis Sentimen Mahasiswa Terhadap Sistem Layanan KPST STMIK Widya Cipta Dharma. *Jurnal Nasional Komputasi Dan Teknologi Informasi (JNKTI)*, *8*(3).