IJCSR: The Indonesian Journal of Computer Science Research
E-ISSN 2963-9174 **DOI prefix** 10.37905
Volume 5, No. 1 Januari 2026
https://subset.id/index.php/IJCSR

Klasifikasi Bakteri E. Coli Menggunakan Algoritma Decission Tree dan Random Forest

Khoirul Riyadi[1, a)] Dina Maulina[2, b)] Melany Mustika Dewi [3, c)] Ninik Trihartanti [4,d)]

[1)]*Program Studi Informatika,*
[2,3)] *Program Studi D3 Manajemen Informatika,*
[4)]*Program Studi Sistem Informasi*
Fakultas Ilmu Komputer
Universitas Amikom Yogyakarta, Jl. Padjajaran, Ring Road Utara, Yogyakarta 55281, Indonesia

Author Emails
[b)] Corresponding author: dina.m@amikom.ac.id
[a)] khorulriyadi@students.amikom.ac.id
[c)] melany@amikom.ac.id
[d)] ninik.t@amikom.ac.id

Abstract. Escherichia coli *(E. coli) is a commensal micro-organism commonly found in the digestive tracts of humans and animals. However, it encompasses a diverse range of strains with varying characteristics, from neutral to pathogenic ones capable of causing serious illness. Consequently, accurate and rapid classification of different E. coli types is crucial, particularly in clinical diagnostics, food safety, and microbiological research. This study aims to develop and evaluate classification models to distinguish E. coli bacterial strains by utilizing two machine learning algorithms: Decision Tree and Random Forest. The dataset used in this research was sourced from the UCI Machine Learning Repository, containing a series of microbiological features. Model performance was evaluated using several standard metrics, namely accuracy, precision, recall, and F1-score, to provide a comprehensive analysis. The experimental results indicate that the Random Forest model significantly outperforms the Decision Tree. Random Forest achieved an accuracy of 88.24%, while Decision Tree reached 79.41%. Random Forest's superiority was also consistently reflected in higher precision, recall, and F1-score values, demonstrating a better ability to capture complex patterns in the data and reduce overfitting. Therefore, Random Forest is proven more effective and reliable for this classification task. The findings of this study offer a potential contribution as a foundation for developing automated decision-support systems that can assist laboratory personnel in pathogen identification, thereby enhancing diagnostic efficiency in the biomedical field.*

Keywords **:**
E Coli, classification, decission tree, random forest

Abstraksi. Bakteri *Escherichia coli* (E. coli) merupakan mikroorganisme komensal yang umum ditemukan dalam saluran pencernaan manusia dan hewan, namun mencakup beragam strain dengan karakter berbeda, dari yang netral hingga patogen yang dapat menyebabkan penyakit serius. Akibatnya, klasifikasi yang akurat dan cepat terhadap berbagai jenis E. coli menjadi hal yang sangat penting, terutama dalam konteks diagnostik klinis, keamanan pangan, dan penelitian mikrobiologi. Penelitian ini bertujuan mengembangkan dan mengevaluasi model klasifikasi untuk membedakan strain bakteri E. coli dengan memanfaatkan dua algoritma *machine learning*, yaitu Decision Tree dan Random Forest. Dataset yang digunakan dalam penelitian ini bersumber dari UCI Machine Learning Repository, yang mencakup serangkaian fitur mikrobiologis. Evaluasi performa model dilakukan menggunakan beberapa metrik standar, yaitu akurasi, *precision*, *recall*, dan *F1-score* untuk memberikan analisis yang komprehensif. Hasil eksperimen menunjukkan bahwa model Random Forest secara signifikan mengungguli Decision Tree. Random Forest mencapai akurasi sebesar

IJCSR: The Indonesian Journal of Computer Science Research
E-ISSN 2963-9174 **DOI prefix** 10.37905
Volume 5, No. 1 Januari 2026
https://subset.id/index.php/IJCSR

88.24%, sedangkan Decision Tree mencapai 79.41%. Keunggulan Random Forest juga konsisten terlihat pada nilai *precision*, *recall*, dan *F1-score* yang lebih tinggi, menunjukkan kemampuan yang lebih baik dalam menangkap pola kompleks dalam data dan mengurangi *overfitting*. Dengan demikian, Random Forest terbukti lebih efektif dan andal untuk tugas klasifikasi ini. Temuan penelitian ini memberikan kontribusi potensial sebagai dasar pengembangan sistem pendukung keputusan otomatis yang dapat membantu tenaga laboratorium dalam identifikasi patogen, sehingga berdampak pada peningkatan efisiensi diagnostik di bidang biomedis.

Kata Kunci:
E Coli, kalsifikasi, decission tree, random forest

PENDAHULUAN

Escherichia coli (E. coli) merupakan bakteri gram negatif berbentuk batang yang secara alami menghuni saluran pencernaan manusia maupun hewan berdarah panas. Dalam keadaan normal, sebagian besar strain E. coli termasuk ke dalam flora usus yang tidak berbahaya dan justru memberikan kontribusi penting dalam proses pencernaan, seperti membantu pemecahan makanan dan sintesis vitamin. Namun demikian, terdapat beberapa strain patogenik yang memiliki kemampuan untuk menyebabkan berbagai gangguan kesehatan serius. Infeksi oleh strain patogen tersebut dapat menimbulkan penyakit mulai dari diare berat, infeksi saluran kemih (ISK), hingga komplikasi yang lebih parah seperti sindrom uremik hemolitik (HUS) yang berpotensi mengancam jiwa. Penularan E. coli ini umumnya terjadi melalui konsumsi makanan atau air yang terkontaminasi bakteri, serta melalui kontak langsung dengan hewan atau lingkungan yang tercemar oleh kotoran. Kondisi sanitasi yang buruk dan kurangnya pengelolaan lingkungan yang higienis memperbesar risiko penyebaran bakteri ini di masyarakat. Hal tersebut sejalan dengan hasil penelitian yang dilakukan di Rumah Sakit Umum Pusat Sanglah Bali, yang mengidentifikasi E. coli sebagai penyebab utama infeksi saluran kemih dengan prevalensi mencapai 39,76%[1]. Studi ini juga menemukan bahwa infeksi E. coli lebih sering terjadi pada wanita dan kelompok usia lanjut, menegaskan bahwa faktor biologis dan demografis turut berperan dalam kerentanan terhadap penyakit ini. Selain di fasilitas kesehatan, berbagai penelitian di Indonesia juga mengungkap tingginya tingkat kontaminasi E. coli pada makanan dan minuman yang beredar di masyarakat. Penelitian [2] di Kendari Beach menemukan bahwa 81,2% minuman tradisional terkontaminasi E. coli, sementara penelitian [3] di Semarang menunjukkan sepertiga jajanan sekolah dasar positif mengandung bakteri ini, Kasus serupa juga terjadi di penelitian [4]pada minuman kekinian di Mamuju, dengan tingkat kontaminasi mencapai lebih dari 90%, serta pada cincau hitam di pasar tradisional Semarang, di mana seluruh sampel yang diuji 100% positif mengandung E. coli.

Identifikasi dan klasifikasi E. coli menjadi sangat penting untuk pencegahan penyebaran infeksi serta mendukung diagnosa medis yang cepat dan tepat. Metode identifikasi konvensional seperti kultur bakteri, uji biokimia, dan polymerase chain reaction (PCR) memang akurat, tetapi membutuhkan waktu lama, biaya tinggi, serta fasilitas laboratorium yang memadai. Oleh karena itu, dibutuhkan pendekatan alternatif yang lebih cepat, efisien, dan tetap akurat. Teknologi machine learning hadir sebagai solusi inovatif untuk mengatasi keterbatasan metode konvensional. Dengan kemampuannya dalam mengenali pola kompleks secara otomatis, metode ini berpotensi merevolusi proses identifikasi bakteri. Penelitian sebelumnya menunjukkan bahwa penerapan metode ensemble mampu meningkatkan akurasi klasifikasi bakteri E. coli secara signifikan dibandingkan metode tunggal [5].

Dalam bidang klasifikasi, algoritma Decision Tree banyak digunakan karena sederhana, mudah dipahami, dan dapat divisualisasikan dengan jelas [6]. Untuk mengatasi kelemahan tersebut, dikembangkan algoritma Random Forest sebagai metode ensemble yang menggabungkan banyak pohon keputusan sekaligus, sehingga menghasilkan prediksi yang lebih stabil, akurat, dan tahan terhadap overfitting. Kedua algoritma ini terbukti efektif dalam berbagai penelitian klasifikasi, termasuk pada data biologis dengan atribut kompleks seperti bakteri E. coli

TINJAUAN PUSTAKA

Penelitian membandingkan kinerja algoritma Random Forest, Decision Tree, dan Gradient Boosting dalam memprediksi kualitas air [3]. Hasil penelitian menunjukkan bahwa Random Forest memiliki akurasi tertinggi sebesar 88,33%, dibandingkan dengan Decision Tree yang hanya mencapai 80,83%. Penelitian ini menegaskan bahwa RF

IJCSR: The Indonesian Journal of Computer Science Research
E-ISSN 2963-9174 **DOI prefix** 10.37905
Volume 5, No. 1 Januari 2026
https://subset.id/index.php/IJCSR

cenderung lebih stabil dan unggul dalam mengklasifikasikan data lingkungan yang memiliki banyak atribut numerik, serta mampu menangani data yang tidak seimbang lebih baik daripada DT.

Selanjutnya, penelitian berfokus pada analisis dan klasifikasi data teks dalam domain kesehatan, dengan objek kajian berupa artikel-artikel yang membahas topik seputar penyakit kanker [2]. Tujuan utama dari studi ini adalah untuk mengevaluasi dan membandingkan kinerja dua algoritma klasifikasi populer, yaitu Decision Tree dan Random Forest, dalam mengelompokkan artikel berdasarkan kategorinya. Hasil eksperimen menunjukkan bahwa algoritma Random Forest secara konsisten memberikan performa yang jauh lebih unggul, dengan rata-rata akurasi mencapai 99,56%. Sebaliknya, Decision Tree hanya mampu memperoleh akurasi sebesar 75,24%, yang menunjukkan adanya perbedaan signifikan dalam kemampuan prediksi kedua metode tersebut. Temuan ini semakin menguatkan reputasi Random Forest sebagai model yang andal, tidak hanya dalam konteks data numerik, tetapi juga saat diterapkan pada data berbasis teks. Selain itu, penelitian ini menegaskan bahwa Random Forest mampu memberikan hasil prediksi yang stabil dan akurat dalam berbagai skenario pembagian data latih dan uji, sehingga menjadikannya pilihan yang lebih efektif dalam klasifikasi dokumen teks di bidang kesehatan.

Penelitian juga mendukung temuan serupa. Dalam sebuah studi klasifikasi bakteri E. coli, dilakukan perbandingan antara Naive Bayes, Decision Tree, dan Random Forest menggunakan teknik oversampling untuk menyeimbangkan distribusi kelas [4]. Hasil menunjukkan bahwa Random Forest memperoleh akurasi terbaik sebesar 84%, dibandingkan dengan Decision Tree dan Naive Bayes. Penelitian ini menunjukkan bahwa model ensemble seperti Random Forest mampu memberikan generalisasi yang lebih baik pada data tidak seimbang.

Penelitian membandingkan tujuh metode ensemble berbasis Decision Tree, termasuk Random Forest, dalam mengklasifikasikan E. coli [7]. Meski XGBoost menjadi algoritma dengan akurasi tertinggi (88%), Random Forest tetap menunjukkan kinerja yang stabil dan akurat, menjadikannya pilihan yang layak sebagai baseline dalam model pohon keputusan.

Penelitian yang turut membahas perbandingan performa beberapa algoritma klasifikasi, yakni Decision Tree (DT), Random Forest (RF), dan K-Nearest Neighbors (KNN), dalam konteks klasifikasi data promosi karyawan [8]. Dataset yang digunakan dalam penelitian ini memiliki karakteristik tidak seimbang, di mana jumlah data antar kelas berbeda secara signifikan. Untuk mengatasi permasalahan tersebut, peneliti menerapkan metode *Synthetic Minority Over-sampling Technique* (SMOTE), sebuah teknik yang umum digunakan dalam domain data mining untuk menyeimbangkan distribusi kelas dan meningkatkan performa algoritma pada kelas minoritas. Setelah proses balancing dilakukan, algoritma Random Forest berhasil mencatatkan akurasi tertinggi, yakni sebesar 86,37%, melampaui hasil yang dicapai oleh *Decision Tree*. Hasil ini menunjukkan keunggulan *Random Forest* dalam menangani data yang telah diseimbangkan, bahkan dalam domain yang berbeda dari penelitian ini. Walaupun fokus utama dari studi tersebut berada di ranah manajemen sumber daya manusia, pendekatan evaluasi yang digunakan tetap relevan untuk diterapkan dalam penelitian ini, terutama dalam konteks klasifikasi berbasis teks. Metode evaluasi yang digunakan, seperti confusion matrix, precision, recall, dan F1-score, terbukti efektif dalam mengukur performa model secara menyeluruh, tidak hanya dari sisi akurasi saja. Oleh karena itu, meskipun topik dan jenis data yang digunakan berbeda, strategi evaluasi dan teknik penyeimbangan data dari penelitian ini memberikan referensi penting dan relevan untuk mendukung validitas metode yang digunakan dalam skripsi ini.

Penelitian mengkaji perbandingan performa antara algoritma Decision Tree (DT) dan Random Forest (RF) dalam melakukan klasifikasi terhadap data diagnosis penyakit hepatitis [1]. Studi ini menggunakan data medis yang bersifat sensitif dan memerlukan tingkat akurasi yang tinggi dalam proses klasifikasinya, mengingat implikasi langsungnya terhadap pengambilan keputusan di bidang kesehatan. Berdasarkan hasil pengujian, algoritma Random Forest menunjukkan performa yang lebih unggul dengan capaian akurasi sebesar 90,32%, precision sempurna sebesar 100%, dan recall sebesar 89,28%. Sebaliknya, algoritma Decision Tree hanya mampu memperoleh akurasi sebesar 80,65%, dengan performa yang relatif lebih rendah dalam metrik lainnya. Perbedaan performa ini mengindikasikan bahwa Random Forest tidak hanya mampu menangkap pola data dengan lebih baik, tetapi juga lebih andal dalam menghindari kesalahan klasifikasi, terutama pada data medis yang memiliki konsekuensi tinggi. Oleh karena itu, temuan dari studi ini semakin memperkuat bukti bahwa Random Forest merupakan pilihan yang lebih efektif dibandingkan Decision Tree dalam konteks klasifikasi data kesehatan yang kompleks dan kritis.

Secara keseluruhan, berbagai studi tersebut menunjukkan bahwa algoritma Random Forest memiliki keunggulan dalam hal akurasi, stabilitas, dan kemampuan menangani data kompleks dibandingkan Decision Tree. Meski demikian, Decision Tree tetap relevan dalam situasi di mana interpretabilitas dan kejelasan aturan pengambilan keputusan menjadi prioritas utama. Dengan demikian, pemilihan algoritma bergantung pada konteks dan kebutuhan spesifik dari kasus klasifikasi yang dihadapi, termasuk dalam hal ini klasifikasi bakteri E. coli. Dengan memahami

IJCSR: The Indonesian Journal of Computer Science Research
E-ISSN 2963-9174 **DOI prefix** 10.37905
Volume 5, No. 1 Januari 2026
https://subset.id/index.php/IJCSR

rumus di atas, data yang telah diperoleh dapat dimasukkan dan diproses menggunakan algoritma ini untuk proses pembuatan pohon keputusan.

Escherichia Coli

Escherichia coli (E. coli) merupakan bakteri gram negatif berbentuk batang yang secara alami ditemukan dalam saluran pencernaan manusia maupun hewan berdarah panas (Utama et al., 2024). Sebagian besar strain E. coli bersifat komensal dan berperan dalam mendukung proses pencernaan, namun terdapat pula strain patogen yang dapat menyebabkan berbagai gangguan kesehatan, seperti infeksi saluran kemih, diare akut, hingga sindrom uremik hemolitik. Oleh karena itu, identifikasi serta klasifikasi strain E. coli sangat diperlukan untuk menentukan langkah penanganan yang tepat, khususnya dalam ranah kesehatan masyarakat dan mikrobiologi

Decision Tree

Decision Tree adalah algoritma klasifikasi yang bekerja dengan membagi data berdasarkan atribut tertentu, membentuk struktur berbentuk pohon dengan cabang dan simpul. Algoritma ini bekerja dari atas ke bawah, dimulai dari akar pohon, dan membagi dataset ke dalam kelompok berdasarkan nilai atribut yang paling signifikan. Keunggulan utama Decision Tree adalah kemudahannya dalam interpretasi, sebab hasil model dapat divisualisasikan dalam bentuk diagram pohon. Namun, kelemahannya adalah rentan terhadap overfitting, terutama jika pohon tumbuh terlalu dalam tanpa pemangkasan [10]. Beberapa penelitian menunjukkan bahwa meskipun Decision Tree dapat digunakan untuk klasifikasi data nutrisi makanan cepat saji, model ini cenderung kurang akurat dibanding metode ensemble seperti Random Forest [11]. Pada algoritma ini, pohon dibangun dengan membagi dataset berdasarkan kriteria ketidakmurnian, seperti Entropi. Entropi digunakan untuk mengukur ketidakpastian atau ketidakmurnian dalam sebuah dataset.

Random forest

Random Forest merupakan pengembangan dari algoritma Decision Tree dengan pendekatan ensemble learning. Algoritma ini bekerja dengan membangun sejumlah pohon keputusan secara acak, kemudian menggabungkan hasil prediksi dari setiap pohon melalui metode voting (untuk klasifikasi) atau rata-rata (untuk regresi) guna memperoleh hasil akhir yang lebih akurat. Secara matematis, rumus prediksi akhir Random Forest untuk klasifikasi dapat dinyatakan sebagai berikut:

$$y = mode(h_1(x), (h_2(x), \dots\dots, (h_k(x)) \quad (2)$$

Dimana:
y = hasil prediksi akhir dari random forest,
$h_i(x)$ = prediksi dari pohon keputusan ke-i terhadap input x,
k = jumlah total pohon dalam ensemble
mode(.) = fungsi yang memilih kelas terbanyak (mayoritas voting) dari semua prediksi pohon.

Salah satu keunggulan utama Random Forest adalah kemampuannya dalam mengurangi risiko overfitting yang sering terjadi pada algoritma Decision Tree tunggal. Selain itu, Random Forest juga memiliki ketahanan terhadap outlier, mampu menangani data dalam jumlah besar dan berdimensi tinggi, serta tetap menunjukkan performa yang baik meskipun tanpa proses tuning parameter yang kompleks [12]. Menariknya, performa Random Forest cenderung stabil meskipun parameter model tidak disesuaikan secara spesifik untuk masing-masing kasus. Artinya, meskipun komponen setiap pohon dibentuk secara acak dan tidak dioptimalkan secara eksplisit terhadap karakteristik data tertentu, model secara keseluruhan tetap mampu memberikan hasil prediksi yang baik.

Evaluasi kinerja machine learning

Untuk memperoleh nilai dari berbagai metrik evaluasi, digunakan *confusion matrix*, yang diukur berdasarkan nilai-nilai yang diperoleh dari confusion matrix seperti *true positive, false positive, true negative,* dan *false negative*.

IJCSR: The Indonesian Journal of Computer Science Research
E-ISSN 2963-9174 **DOI prefix** 10.37905
Volume 5, No. 1 Januari 2026
https://subset.id/index.php/IJCSR

TABEL 1 .Tabel Confusion Matrix

	Prediksi Positif	***Prediksi Negatif***
Data Aktual Positif	True Positive (TP)	False Negative (FN)
Data Aktual Negatif	False Positive (FP)	True Negative (TN)

ConfusionMatrix

Confusion matrix merupakan alat evaluasi dalam klasifikasi yang terdiri dari empat elemen utama untuk menggambarkan kinerja model. *True positive* menunjukkan jumlah kasus yang benar-benar termasuk dalam kelas positif dan berhasil diprediksi secara tepat oleh model. *False positive* adalah jumlah data yang sebenarnya berasal dari kelas negatif, namun keliru diklasifikasikan sebagai positif. *False negative* merujuk pada jumlah data yang sebenarnya termasuk kelas positif, tetapi salah diprediksi sebagai negatif. Sementara itu, *true negative* mencerminkan jumlah data yang benar-benar negatif dan juga berhasil dikenali dengan benar oleh model. Berdasarkan *confusion matrix* tersebut, berikut rumus perhitungan untuk masing-masing metrik:

a. Akurasi

Mengukur seberapa besar jumlah prediksi yang benar terhadap keseluruhan data:

$$Akurasi = \frac{TP+TN}{TP+TN+FP+FN} (3)$$

b. Presisi

Menilai sejauh mana model mampu melakukan prediksi secara benar terhadap kelas positif.

$$Presisi = \frac{TP}{TP+FP} (4)$$

c. Recall

Menilai sejauh mana model mampu mengenali seluruh data yang termasuk kategori positif.

$$Recall = \frac{TP}{TP+FN} (5)$$

d. F1-score

Yaitu nilai rata-rata harmonik antara presisi dan recall.

$$F1 - score = 2x\frac{Presisi\ x\ recall}{presisi+recall} \quad (6)$$

METODE PENELITIAN

Metode penelitian ini disusun secara sistematis melalui beberapa tahapan utama, yaitu pengumpulan data, pra-pemrosesan data, pembangunan model klasifikasi, evaluasi performa, serta analisis hasil. Data penelitian diperoleh dari UCI Machine Learning Repository berupa dataset E. coli. Dataset ini terdiri dari 336 entri data yang merepresentasikan berbagai strain dengan karakteristik biologis berbeda. Setiap entri data memiliki tujuh atribut numerik (mcg, gvh, lip, chg, aac, alm1, alm2) serta satu atribut target class yang menunjukkan lokasi subseluler protein ke dalam delapan kelas berbeda. Dataset ini dipilih karena telah banyak digunakan pada penelitian terdahulu dan dianggap representatif untuk tugas klasifikasi biologis.

$$\frac{d[F1]}{dw_2} = SAm_2 \cos w, \frac{d[F1]}{dw_3} = SAm_2 \cos w\ (7)$$

Berikut isi dataset lima baris pertama:

TABEL 2 . *Isi Dataset Lima Baris Pertama*

mcg	***gvh***	***lip***	***chg***	***aac***	***alm1***	***alm2***	***class***
0.49	0.29	0.48	0.5	0.56	0.24	0.35	cp
0.07	0.4	0.48	0.5	0.54	0.35	0.44	cp

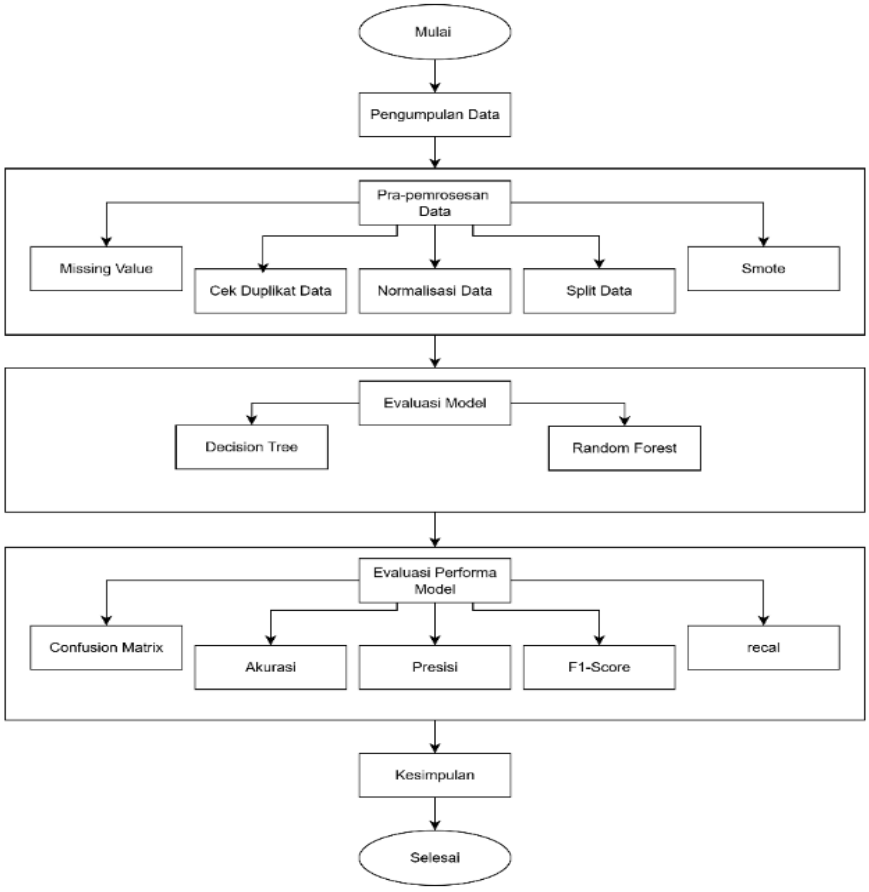
IJCSR: The Indonesian Journal of Computer Science Research
E-ISSN 2963-9174 **DOI prefix** 10.37905
Volume 5, No. 1 Januari 2026
https://subset.id/index.php/IJCSR

0.56	0.4	0.48	0.5	0.49	0.37	0.46	cp
0.59	0.49	0.48	0.5	0.52	0.45	0.36	cp

Setelah data diperoleh, dilakukan tahap pra pemrosesan untuk memastikan kualitas data sesuai kebutuhan algoritma machine learning. Tahap ini meliputi pembersihan data dari missing value dan duplikasi, normalisasi atribut numerik agar berada dalam skala seragam, serta pengkodean label kelas jika diperlukan. Dataset kemudian dibagi menjadi dua bagian, yaitu data latih dan data uji. Apabila terdapat ketidakseimbangan distribusi antar kelas, maka diterapkan teknik penyeimbangan seperti resampling atau Synthetic Minority Over-sampling Technique (SMOTE) guna meningkatkan representasi data dan mendukung kinerja model secara optimal [15]

Tahap berikutnya adalah pembangunan model klasifikasi dengan menggunakan algoritma Decision Tree dan Random Forest. Model dilatih menggunakan data latih, kemudian diuji menggunakan data uji untuk memprediksi kelas lokasi subseluler E. coli. Pemilihan kedua algoritma ini didasarkan pada kemampuannya dalam memberikan interpretasi terhadap pengaruh masing-masing atribut, di mana struktur pohon keputusan yang terbentuk dapat menunjukkan kontribusi setiap fitur terhadap hasil klasifikasi.

Selanjutnya, dilakukan evaluasi performa model dengan menggunakan metrik akurasi, presisi, recall, F1-score yang dihitung berdasarkan confusion matrix. Evaluasi ini bertujuan untuk menilai sejauh mana model mampu mengklasifikasikan data secara tepat, baik pada kelas mayoritas maupun minoritas. Pada tahap akhir, hasil evaluasi dianalisis untuk menentukan algoritma dengan performa terbaik, disertai pembahasan mengenai interpretasi hasil, kendala yang dihadapi selama penelitian, serta rekomendasi pengembangan lebih lanjut pada penelitian mendatang.



GAMBAR 1. Alur penelitian

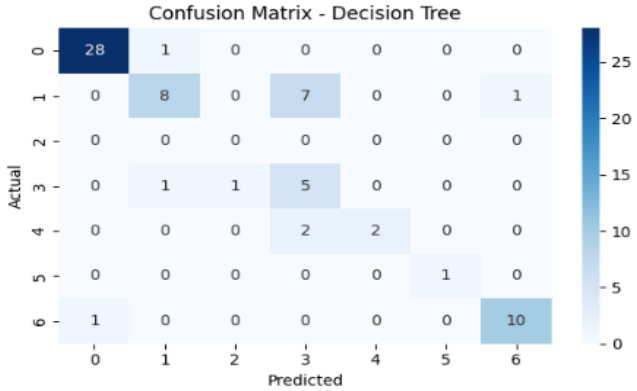
Dengan demikian, rangkaian metode penelitian yang telah dijelaskan dirancang secara sistematis agar mampu menghasilkan model klasifikasi yang akurat dan andal dalam memprediksi lokasi subseluler Escherichia coli.

IJCSR: The Indonesian Journal of Computer Science Research
E-ISSN 2963-9174 **DOI prefix** 10.37905
Volume 5, No. 1 Januari 2026
https://subset.id/index.php/IJCSR

Setiap tahapan, mulai dari pengumpulan data hingga evaluasi performa, saling berkaitan dan menjadi dasar penting dalam mencapai tujuan penelitian ini.

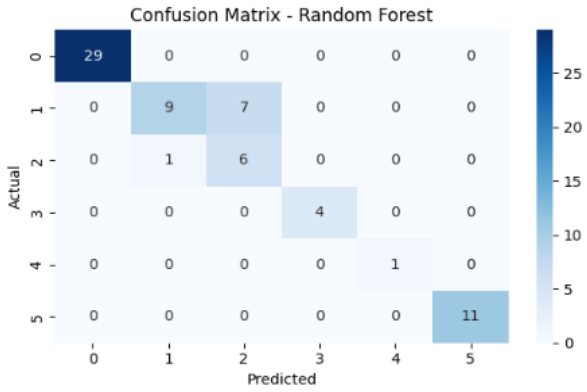
HASIL DAN PEMBAHASAN

Pada tahap ini dipaparkan hasil implementasi algoritma Decision Tree dan Random Forest dalam mengklasifikasikan bakteri Escherichia coli. Hasil evaluasi disajikan dalam bentuk metrik kinerja yang meliputi akurasi, precision, recall, F1-score. Evaluasi ini bertujuan untuk mengetahui sejauh mana model mampu mengklasifikasikan data dengan tepat dan untuk membandingkan efektivitas kedua algoritma dalam kasus yang sama.



GAMBAR 2. Confusion Matrix Decision Tree

Gambar confusion matrix di atas memperlihatkan kinerja model Decision Tree dalam mengklasifikasikan data uji. Hasilnya menunjukkan bahwa sebagian besar data berhasil diprediksi dengan benar, terlihat dari dominasi nilai pada diagonal utama. Meskipun demikian, masih terdapat sejumlah kesalahan klasifikasi, di mana beberapa sampel dari kelas tertentu salah teridentifikasi sebagai kelas lain. Hal ini menandakan bahwa Decision Tree memiliki keterbatasan dalam membedakan pola antar kelas yang kompleks, sehingga performanya cenderung kurang stabil dibandingkan algoritma ensemble seperti Random Forest.

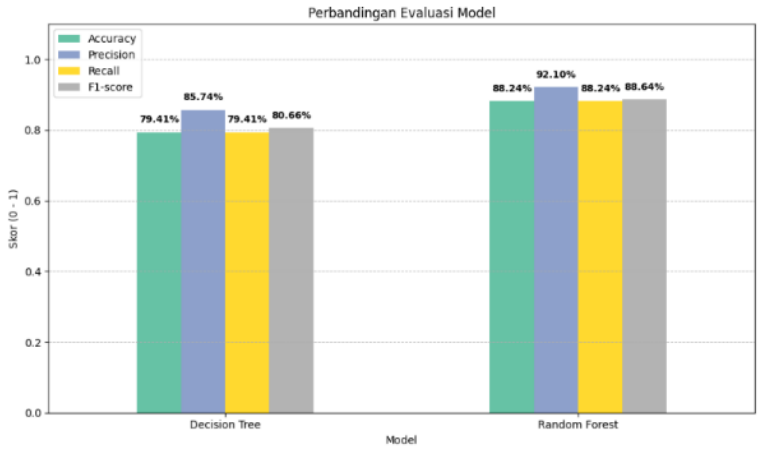


GAMBAR 3. Confusion Matrix Random Forest

Gambar confusion matrix di atas menunjukkan kinerja model Random Forest dalam melakukan klasifikasi data uji. Terlihat bahwa sebagian besar prediksi berada pada diagonal utama, yang menandakan tingkat akurasi tinggi karena banyak data berhasil dikenali sesuai dengan kelas aslinya. Hanya sedikit kesalahan klasifikasi yang terjadi, di mana sebagian sampel dari kelas tertentu masih teridentifikasi sebagai kelas lain. Secara keseluruhan, distribusi

IJCSR: The Indonesian Journal of Computer Science Research
E-ISSN 2963-9174 **DOI prefix** 10.37905
Volume 5, No. 1 Januari 2026
https://subset.id/index.php/IJCSR

prediksi pada confusion matrix ini mengindikasikan bahwa Random Forest lebih stabil dan andal dibandingkan Decision Tree, terutama dalam membedakan pola antar kelas yang kompleks, sehingga mampu menghasilkan hasil klasifikasi yang lebih konsisten.



GAMBAR 4. Perbandingan Evaluasi Model

Dari grafik terlihat bahwa Random Forest unggul di semua metrik evaluasi dibandingkan dengan Decision Tree. Model ini mampu memberikan prediksi yang lebih konsisten, baik dalam hasil ketepatan klasifikasi keseluruhan, kemampuan mengenali kelas positif, maupun keseimbangan antara precision dan recall. Sementara itu, Decision Tree masih menunjukkan performa yang lebih rendah pada seluruh metrik, meskipun memiliki keunggulan dari sisi waktu pelatihan yang lebih singkat.
Dengan demikian, dapat disimpulkan bahwa Random Forest lebih efektif dan stabil dalam mengklasifikasikan bakteri Escherichia coli, sedangkan Decision Tree lebih efisien secara komputasi namun kurang optimal dari sisi akurasi hasil klasifikasi.

TABEL 3. *Hasil Evaluasi Model*

Algoritma	*Accuracy*	*Precision(%)*	*Recall(%)*	*F1-Score(%)*	*Waktu Pelatihan (detik)*
Decision Tree	79.41	85.74	79.41	80.66	0.0077
Random Forest	88.24	92.10	88.24	88.64	0.2297

Hasil evaluasi performa model menunjukkan bahwa algoritma Random Forest memiliki kinerja yang lebih baik dibandingkan Decision Tree. Pada metrik akurasi, Random Forest mencapai 88,24%, lebih tinggi dibandingkan Decision Tree yang hanya memperoleh 79,41%. Hal yang sama juga terlihat pada metrik precision, recall, dan F1-score, di mana Random Forest masing-masing mencatat nilai 92,10%, 88,24%, dan 88,64%, sedangkan Decision Tree hanya mencapai 85,74%, 79,41%, dan 80,66%. Hasil ini mengindikasikan bahwa Random Forest lebih konsisten dalam menghasilkan prediksi yang tepat, baik dalam mengenali kelas positif maupun dalam menjaga keseimbangan antara precision dan recall.

Namun demikian, perbedaan kinerja ini juga diikuti oleh perbedaan pada waktu pelatihan model. Decision Tree membutuhkan waktu relatif singkat, yaitu sekitar 0,0077 detik, sedangkan Random Forest memerlukan waktu lebih lama yaitu 0,2297 detik. Hal ini disebabkan karena Random Forest membangun sejumlah pohon keputusan secara bersamaan dalam proses pelatihannya. Meskipun demikian, tambahan waktu komputasi ini masih tergolong sangat kecil dan dapat ditoleransi, mengingat peningkatan akurasi dan stabilitas yang diperoleh cukup signifikan. Dengan demikian, dapat disimpulkan bahwa Random Forest merupakan algoritma yang lebih unggul untuk klasifikasi bakteri Escherichia coli dibandingkan Decision Tree.

KESIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan mengenai klasifikasi bakteri Escherichia coli menggunakan algoritma Decision Tree dan Random Forest dengan dataset dari UCI Machine Learning Repository, dapat disimpulkan beberapa hal penting. Dataset yang digunakan terdiri dari 336 entri dengan delapan atribut, kemudian dibagi secara proporsional menjadi 80% data latih dan 20% data uji. Sebelum tahap pembangunan model, data melalui serangkaian proses pra-pemrosesan yang meliputi pemeriksaan kelengkapan data, normalisasi dengan Min-Max Scaler, konversi label target menggunakan label encoding, penyeimbangan kelas dengan SMOTE, serta deteksi dan penanganan outlier dengan metode LOF dan IQR. Hasil pra-pemrosesan menunjukkan bahwa dataset tidak memiliki missing value maupun data duplikat, sehingga kualitas data dapat dipastikan dalam kondisi baik untuk tahap selanjutnya. Pada tahap pembangunan model dan evaluasi, kedua algoritma dilatih dan diuji menggunakan data yang telah diproses. Evaluasi dilakukan menggunakan metrik akurasi, precision, recall, dan F1-score untuk memberikan gambaran yang lebih komprehensif terkait kinerja model. Dari hasil evaluasi diperoleh bahwa algoritma Random Forest secara konsisten menunjukkan performa yang lebih unggul dibandingkan Decision Tree, baik dari sisi akurasi keseluruhan maupun keseimbangan antara precision dan recall. Random Forest berhasil mencapai nilai akurasi dan F1-score yang lebih tinggi dibandingkan Decision Tree, yang menandakan bahwa model ini lebih efektif dalam mengklasifikasikan data multikelas.

TINJAUAN PUSTAKA

[1] Y. Kandarini, W. Mahadita, and P. A. Marciyasa, "Profil kuman pada infeksi saluran kemih di Rumah Sakit Umum Pusat Sanglah Bali tahun 2019-2020," *Jurnal Penyakit Dalam Udayana*, vol. 4, no. 2, pp. 45–52, Dec. 2020, doi: 10.36216/jpd.v4i2.140.

[2] F. Fithria, Y. Yasnani, and H. N. Alhajar, "HUBUNGAN HIGIENE SANITASI DENGAN KEBERADAAN E.coli PADA MINUMAN OLAHAN DI KENDARI BEACH," *IKESMA*, vol. 18, no. 3, p. 192, Sep. 2022, doi: 10.19184/ikesma.v18i3.30780.

[3] T. Astuti, "https://repository.uinjkt.ac.id/dspace/handle/123456789/65552," Fakultas Sains dan Teknologi Universitas Islam Negeri Syarif Hidayatullah Jakarta.

[4] F. Islam, H. Ahmad, and S. Saddania, "Kontaminasi Escherichia coli pada Minuman Kekinian di Pusat Kota Kabupaten Mamuju," *Jurnal kesehatan komunitas (Journal of community health)*, vol. 10, no. 2, pp. 222–228, Jun. 2024, doi: 10.25311/keskom.Vol10.Iss2.1664.

[5] A. R. Al Musyaffa, Y. Pristyanto, and N. Mauliza, "COMPARISON OF ENSEMBLE METHODS FOR DECISION TREE MODELS IN CLASSIFYING E. COLI BACTERIA," *JITK (Jurnal Ilmu Pengetahuan dan Teknologi Komputer)*, vol. 10, no. 3, pp. 514–522, Feb. 2025, doi: 10.33480/jitk.v10i3.5972.

[6] F. Baharuddin and A. Tjahyanto, "Peningkatan Performa Klasifikasi Machine Learning Melalui Perbandingan Metode Machine Learning dan Peningkatan Dataset," *Jurnal Sisfokom (Sistem Informasi dan Komputer)*, vol. 11, no. 1, pp. 25–31, Mar. 2022, doi: 10.32736/sisfokom.v11i1.1337.

[7] M. Amien, "Workshop Pengenalan Bahasa Pemrograman Python untuk Data Sains," *Dharma Nusantara: Jurnal Ilmiah Pemberdayaan dan Pengabdian kepada Masyarakat*, vol. 1, no. 2, pp. 39–42, Feb. 2024, doi: 10.32664/dharma.v1i2.1158.

[8] S. Japit, Y. Risyani, C. Bombongan, T. Selamat, and Y. Yuliana, "Pemanfaatan Pandas, NumPy, Spark, dan TableauPower BI dalam Pengolahan dan Visualisasi Big Data untuk Mendukung Pengambilan Keputusan," *Jurnal Minfo Polgan*, vol. 14, no. 1, pp. 452–456, May 2025, doi: 10.33395/jmp.v14i1.14762.

[9] H. U. Utama, A. Masruro, and A. Triyadi, "Kolaborasi Naïve Bayes dan AdaBoost dalam Klasifikasi Bakteri E.coli," *Jurnal Sistem Informasi, Manajemen dan Teknologi Informasi*, vol. 2, no. 2, pp. 68–82, Jul. 2024, doi: 10.33020/jsimtek.v2i2.756.

[10] A. Tharwat, "Classification assessment methods," *Applied Computing and Informatics*, vol. 17, no. 1, pp. 168–192, Jan. 2021, doi: 10.1016/j.aci.2018.08.003.

[11] N. I. Yaman, A. R. Juwita, S. A. P. Lestari, and S. Faisal, "Perbandingan Kinerja Algoritma Decision Tree dan Random Forest untuk Klasifikasi Nutrisi pada Makanan Cepat Saji," *Jurnal Algoritma*, vol. 21, no. 2, pp. 184–196, Dec. 2024, doi: 10.33364/algoritma/v.21-2.1649.

[12] N. Maulidah, M. Maulidah, R. Supriyadi, H. Nalatissifa, S. Diantika, and A. Fauzi, "PREDIKSI KUALITAS AIR MENGGUNAKAN METODE RANDOM FOREST, DECISION TREE, DAN GRADIENT BOOSTING," *Jurnal Khatulistiwa Informatika*, vol. 12, no. 1, pp. 1–6, Jul. 2024, doi: 10.31294/jki.v12i1.16004.

[13] A. E. Sawitri and A. Irhandayaningsih, "Analisis Penggunaan Aplikasi Tata Surat Dalam Pengelolaan Surat Masuk Dan Surat Keluar Pada Dinas Kesehatan Provinsi Jawa Tengah," *Jurnal Ilmu Perpustakaan*, vol. 6, no. 3, pp. 411–420, 2017, [Online]. Available: https://ejournal3.undip.ac.id/index.php/jip/article/view/23173

[14] A. Pannadhitthana Candra, "Analisis Data Menggunakan Python: Memperkenalkan Pandas dan NumPy," *Journal of Information System and Education Development*, vol. 3, no. 1, pp. 11–16, Mar. 2025, doi: 10.62386/jised.v3i1.118.

[15] LOUIS MADAERDO SOTARJUA and DIAN BUDHI SANTOSO, "PERBANDINGAN ALGORITMA KNN, DECISION TREE,*DAN RANDOM*FOREST PADA DATA IMBALANCED CLASS UNTUK KLASIFIKASI PROMOSI KARYAWAN," *Jurnal INSTEK (Informatika Sains dan Teknologi)*, vol. 7, no. 2, pp. 192–200, Aug. 2022, doi: 10.24252/instek.v7i2.31385.