Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2582−2593
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.7029

Deteksi Dini Risiko Penyakit Jantung Koroner Menggunakan Algoritma Decision Tree dan Random Forest

Slamet Hudha Nurrohman*, Defri Kurniawan

Fakultas Ilmu Komputer, Program Studi Teknik Informatika, Universitas Dian Nuswantoro, Semarang, Indonesia
Email: [1,*]111202113261@mhs.dinus.ac.id, [2]defri.kurniawan@dsn.dinus.ac.id
Email Penulis Korespondensi: 111202113261@mhs.dinus.ac.id
Submitted: **25/02/2025**; Accepted: **10/03/2025**; Published**: 16/03/2025**

Abstrak−Penyakit jantung koroner merupakan penyebab utama kematian global, dengan angka mortalitas mencapai 17,9 juta jiwa per tahun. Oleh karena itu, deteksi dini menjadi langkah krusial dalam mitigasi risiko guna mencegah komplikasi lebih lanjut. Namun, metode konvensional metode diagnosis tradisional atau evaluasi medis manual sering kali kurang efisien dalam menangani volume data medis yang besar, sehingga diperlukan pendekatan yang lebih optimal. Dalam upaya meningkatkan efisiensi, penelitian ini menerapkan pembelajaran mesin untuk membangun model klasifikasi risiko penyakit jantung koroner menggunakan algoritma Decision Tree dan Random Forest. Kedua model tersebut kemudian dibandingkan untuk menentukan model yang paling optimal. Model ini dibangun menggunakan Framingham Heart Study Dataset, yang terdiri dari 4.240 record dengan 15 variabel fitur yang relevan. Mengingat adanya ketidakseimbangan dalam distribusi kelas target, metode Random Over-Sampling diterapkan untuk meningkatkan performa klasifikasi. Selanjutnya, evaluasi model dilakukan menggunakan confusion matrix untuk membandingkan kinerja kedua algoritma. Hasil analisis menunjukkan bahwa Random Forest memiliki performa yang lebih unggul dengan akurasi 97,64%, presisi 96,02%, recall 99,29%, dan F1-score 97,63%. Sementara itu, Decision Tree menunjukkan akurasi 91,04%, presisi 84,76%, recall 99,57%, dan F1-score 91,57%. Penelitian ini mengindikasikan bahwa Random Forest lebih optimal dalam deteksi dini penyakit jantung koroner dibandingkan Decision Tree. Dengan demikian, model berbasis Random Forest berpotensi untuk diterapkan dalam sistem prediksi klinis, meskipun optimalisasi lebih lanjut masih diperlukan guna meningkatkan akurasi dan keandalan prediksi.

Kata Kunci: Penyakit Jantung Koroner; Pembelajaran Mesin; Decision Tree; Random Forest; Prediksi

Abstract−Coronary heart disease is the leading cause of global mortality, accounting for 17.9 million deaths annually. Early detection is crucial in mitigating risks and preventing further complications. However, conventional diagnostic methods, such as traditional medical evaluations, often struggle to efficiently process large volumes of medical data, necessitating a more optimal approach. To enhance efficiency, this study employs machine learning to develop a classification model for coronary heart disease risk using Decision Tree and Random Forest algorithms. These models are then compared to determine the most optimal approach. The model is built using the Framingham Heart Study Dataset, consisting of 4,240 records with 15 relevant features. Due to class imbalance in the target variable, the Random Over-Sampling method is applied to improve classification performance. Model evaluation is conducted using a confusion matrix to compare the performance of both algorithms. The results indicate that Random Forest outperforms Decision Tree, achieving an accuracy of 97.64%, precision of 96.02%, recall of 99.29%, and F1-score of 97.63%. In contrast, Decision Tree yields an accuracy of 91.04%, precision of 84.76%, recall of 99.57%, and F1-score of 91.57%. This study suggests that Random Forest is more effective for early detection of coronary heart disease. Therefore, Random Forest-based models hold potential for clinical prediction systems, though further optimization is needed to enhance accuracy and reliability.

Keywords: Coronary Heart Disease; Machine Learning; Decision Tree; Random Forest; Prediction

1. PENDAHULUAN

Kesehatan adalah aspek fundamental yang berperan penting dalam kehidupan manusia, sehingga upaya perlindungan dan pencegahan terhadap gangguan serta penyakit berbahaya menjadi fokus utama. Sektor ini terus mengalami peningkatan dan pengembangan yang bertujuan untuk memastikan kesejahteraan masyarakat secara berkelanjutan [1]. Meskipun demikian, dalam beberapa tahun terakhir, prevalensi penyakit kronis mengalami peningkatan yang signifikan. Salah satu penyakit kronis yang menunjukkan tren peningkatan kasus secara global dan memiliki dampak besar terhadap kesehatan masyarakat adalah *cardiovascular disease* (CVD) atau penyakit kardiovaskular [2].

Cardiovascular disease (CVD) adalah sekelompok kelainan yang berdampak pada jantung dan sistem peredaran darah, mencakup penyakit arteri koroner, stroke, serta berbagai gangguan pembuluh darah lainnya. Penyakit ini merupakan penyebab utama kematian di seluruh dunia dengan menyumbang sekitar 32% dari total kematian global, dengan 85% di antaranya disebabkan oleh penyakit jantung dan stroke [3], [4]. Berdasarkan data dari *American Heart Association* (AHA), CVD merupakan penyebab utama kematian di Amerika Serikat, mencakup 43,8% dari total 836.456 kematian yang terkait dengan penyakit jantung [5]. Di Asia, angka kematian akibat CVD pada tahun 2019 tercatat mencapai 10,8 juta jiwa, lebih tinggi dibandingkan dengan angka kematian di Amerika Serikat dan Eropa, dengan penyakit jantung koroner sebagai penyumbang utama [6]. Pernyataan ini selaras dengan temuan *World Health Organization* (WHO), yang mengungkapkan bahwa penyakit jantung koroner terbukti sebagai salah satu penyebab kematian utama di dunia, dengan angka mortalitas mencapai 17,9 juta jiwa setiap tahunnya [3]. Seiring dengan meningkatnya angka mortalitas akibat penyakit jantung, diperlukan langkah-langkah preventif yang lebih terarah dan efektif untuk memitigasi dampaknya. Salah satu pendekatan efektif adalah melalui deteksi dini, yang memungkinkan identifikasi melalui faktor risiko utama seperti hipertensi, diabetes, dan obesitas, serta pemberian intervensi medis yang sesuai dan tepat waktu [5], [7]. Pendekatan ini berpotensi secara signifikan mengurangi risiko berkembangnya

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2582−2593
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.7029

komplikasi sekaligus menurunkan angka mortalitas akibat penyakit jantung. Namun, metode diagnosis konvensional sering kali memerlukan waktu lebih lama dan memiliki potensi kesalahan yang lebih tinggi, terutama saat menangani volume data medis dalam jumlah yang besar [8].

Dalam sistem layanan kesehatan modern yang semakin dipengaruhi oleh data medis yang besar, pendekatan manual ini menjadi kurang efisien dan berpotensi menghasilkan diagnosis yang tidak akurat [8]. Ketidakakuratan tersebut dapat berdampak signifikan pada pengelolaan penyakit jantung serta meningkatkan risiko komplikasi yang lebih serius [9]. Untuk mengatasi tantangan ini, pembelajaran mesin (*machine learning*) telah muncul sebagai pendekatan inovatif yang dapat meningkatkan akurasi dan efisiensi dalam diagnosis penyakit jantung. Sebagai cabang dari kecerdasan buatan, pembelajaran mesin memanfaatkan kemampuan untuk menganalisis data yang kompleks serta dapat mengenali pola yang tidak mudah diidentifikasi secara langsung. Teknologi tersebut sudah banyak digunakan untuk analisis pencitraan medis dan pengelolaan data pasien [10], [11]. Penelitian menunjukkan bahwa model berbasis pembelajaran mesin mampu memberikan hasil diagnosis yang lebih cepat dan akurat dibandingkan metode konvensional, sehingga berkontribusi pada pengelolaan penyakit jantung yang lebih efektif [12]. Pembelajaran mesin menggunakan berbagai algoritma yang dikelompokkan ke dalam dua kategori utama, yaitu *supervised* dan *unsupervised*, untuk mendukung proses prediksi serta pengambilan keputusan secara otomatis berdasarkan pada data yang tersedia [13]. Pada *supervised learning*, algoritma memanfaatkan data berlabel untuk mempelajari pola yang kemudian digunakan dalam proses prediksi atau klasifikasi [14]. Sebaliknya, *unsupervised learning* menggunakan algoritma untuk mengidentifikasi struktur tersembunyi atau pola pada data yang tidak memiliki label [14].

Penelitian ini menggunakan dataset berlabel dalam pengembangan model. Oleh karena itu, algoritma pembelajaran *supervised learning* diterapkan dalam proses pembuatan model. Dataset yang digunakan pada penelitian ini berasal dari *National Heart, Lung, and Blood Institute* (NHLBI) yaitu *"Framingham heart study dataset"* [15]. *Framingham heart study dataset* mencakup 4.240 *record* yang mewakili individu-individu peserta studi, dengan 1 variabel target dan 15 variabel fitur yang relevan dengan risiko penyakit jantung. Variabel target dalam dataset ini adalah *TenYearCHD*, yang menunjukkan risiko seseorang mengalami penyakit jantung koroner dalam periode waktu 10 tahun dengan nilai 1 untuk risiko positif dan 0 untuk risiko negatif. Pada dataset tersebut, terdapat ketidakseimbangan distribusi pada variabel target, dimana jumlah data dengan nilai 1 (positif) hanya mencakup 664 *record*, sedangkan nilai 0 (negatif) mencakup 3.596 *record*. Ketidakseimbangan tersebut dapat mempengaruhi hasil klasifikasi dengan menurunkan sensitivitas dan *Area Under Curve* (AUC), karena distribusi data yang tidak merata dapat menyebabkan model lebih cenderung mengidentifikasi kelas mayoritas dan mengabaikan kelas minoritas [16]. Oleh karena itu, untuk mengatasi ketidakseimbangan pada dataset, penelitian ini menerapkan metode *Random Over-Sampling*. Metode ini memperkuat representasi kelas minoritas dengan menambahkan sampel secara acak, sehingga distribusi target dalam dataset menjadi seimbang.

Penelitian sebelumnya telah banyak membahas prediksi risiko penyakit jantung koroner menggunakan metode machine learning. Salah satu studi yang membandingkan dua metode statistik, yaitu *decision tree* dan *multinomial naïve bayes*, menunjukkan bahwa metode *decision tree* memiliki kinerja yang lebih baik, dengan tingkat akurasi mencapai 99,63% [17]. Penelitian lain yang mengevaluasi berbagai algoritma untuk memprediksi penyakit jantung koroner menemukan bahwa model *decision tree* mencapai performa optimal dengan penerapan teknik *feature importance*, dimana pemilihan *max_feature* sebanyak 6 berhasil meningkatkan akurasi model hingga 92,64% [18]. Selain itu, penelitian yang dilakukan oleh Rian Oktafiani, Arief Hermawan, dan Donny Avianto pada tahun 2024 dengan tujuan untuk mengetahui pengaruh parameter *max_depth* pada algoritma berbasis pohon keputusan menunjukkan bahwa algoritma *random forest* mendapatkan nilai akurasi tertinggi sebesar 99,29%, sedangkan *decision tree* memperoleh akurasi sebesar 98,05% [19]. Penelitian lain yang membandingkan berbagai algoritma *machine learning* untuk prediksi penyakit kardiovaskular, termasuk *decision tree, naïve bayes, KNN, SVM, random forest, neural network, logistic regression*, dan *gradient boosting*, menunjukkan bahwa algoritma *random forest* mencapai akurasi tertinggi dengan nilai sebesar 85% [20]. Penelitian lainnya menunjukkan bahwa *decision tree* dapat mencapai akurasi tertinggi sebesar 88,8% dalam memprediksi penyakit kardiovaskular dengan variasi ukuran data latih [21].

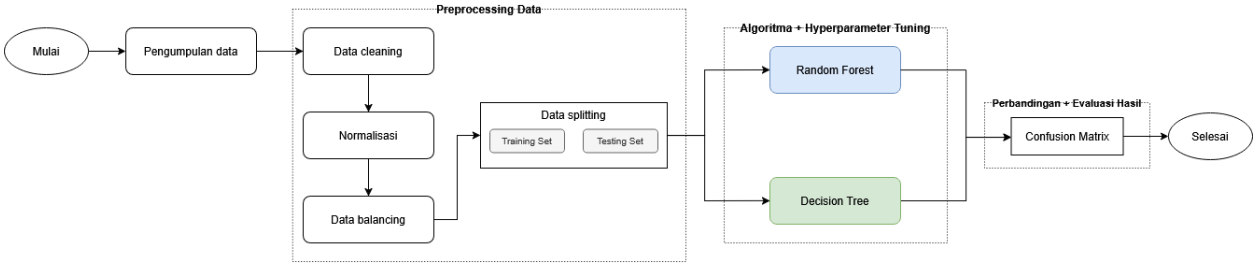
Berdasarkan keunggulan tersebut, algoritma *decision tree* diusulkan sebagai metode dalam membangun model klasifikasi. Namun, model *decision tree* yang dibangun dengan tingkat kompleksitas tinggi memiliki risiko signifikan untuk mengalami *overfitting* [22]. Hal ini sejalan dengan hasil analisis pada penelitain terdahulu yang menunjukkan bahwa performa model *decision tree* cenderung menurun seiring dengan meningkatnya kompleksitasnya [21]. Oleh karena itu, berdasarkan keunggulannya pada penelitian [20], penggunaan algoritma *random forest* juga disarankan dalam penelitian ini agar kinerjanya dapat dibandingkan dengan *decision tree*, sehingga dapat menentukan metode yang lebih optimal dalam prediksi penyakit jantung koroner. Selain itu, kedua algoritma tersebut memiliki kesamaan yaitu berbasis pada pohon keputusan atau *decision tree*, dimana *decision tree* hanya membentuk satu pohon keputusan tunggal. Akan tetapi, *random forest* menggunakan pendekatan *esemble* dengan membangun lebih dari satu pohon keputusan dan akan memilih subset variabel untuk setiap pohon yang dibuat secara acak [23]. Oleh karena itu, *random forest* dipilih sebagai algoritma pembanding untuk *decision tree*. Penelitian ini bertujuan untuk mendeteksi risiko penyakit jantung koroner dalam sepuluh tahun ke depan dengan mengembangkan model berbasis *decision tree*. Model ini dibangun menggunakan data yang tersedia untuk mengidentifikasi pola risiko penyakit. Selain itu, model *random forest* juga dikembangkan dan dibandingkan dengan *decision tree* guna memperoleh model dengan performa terbaik. Sebagai upaya optimalisasi tambahan, *hyperparameter tuning* diterapkan untuk menyesuaikan parameter model agar menghasilkan kinerja yang lebih optimal. Model dengan performa terbaik akan dijadikan sebagai hasil akhir penelitian

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2582−2593
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.7029

ini, dengan harapan dapat memberikan kontribusi signifikan dalam deteksi dini penyakit jantung koroner, serta mendukung upaya pencegahan dan pengambilan keputusan medis yang lebih akurat.

2. METODOLOGI PENELITIAN

Penelitian ini menerapkan dua algoritma berbasis pohon keputusan, yaitu *random forest* dan *decision tree*, untuk mengklasifikasikan risiko penyakit jantung koroner yang mungkin dialami seseorang dalam 10 tahun ke depan berdasarkan fitur dalam dataset yang digunakan. Bagian metode penelitian menjelaskan proses penelitian secara menyeluruh, termasuk pengumpulan data, pengolahan data, serta algoritma yang digunakan dalam membangun model. Proses metode penelitian dalam analisis perbandingan algoritma klasifikasi penyakit kardiovaskular dapat dilihat pada Gambar 1.



Gambar 1. Alur Penelitian

Gambar 1 menggambarkan alur penelitian yang dilakukan dalam penelitian ini. Proses diawali dengan pengumpulan data, kemudian dilanjutkan dengan tahap *preprocessing*, yang meliputi *data cleaning*, *normalisasi*, penyeimbangan distribusi data (*balancing*), serta pembagian dataset menjadi *training set* dan *testing set*. Setelah tahap ini, data siap digunakan untuk pelatihan dan pengujian model. Selanjutnya, algoritma *decision tree* dan *random forest* dengan parameter terbaik diterapkan untuk melakukan klasifikasi. Untuk tahap akhir pada penelitian ini adalah evaluasi model menggunakan *confusion matrix* dan *classification report* guna mengukur kinerja algoritma yang diterapkan.

2.1 Pengumpulan Data

Proses pengumpulan data merupakan tahap awal dalam siklus analisis data [24]. Dalam penelitian ini, dataset diperoleh dari situs penyedia dataset online, yaitu Kaggle. Dataset tersebut berasal dari *National Heart, Lung, and Blood Institute* (NHLBI) yang bekerja sama dengan *Boston University*, dikenal sebagai *Framingham Heart Study Dataset* [15]. Dataset ini terdiri dari 4.240 *records* yang merepresentasikan individu peserta studi, dengan satu variabel target dan 15 variabel fitur yang berkaitan dengan risiko penyakit jantung. Variabel target dalam dataset ini adalah *TenYearCHD*, yang menunjukkan kemungkinan seseorang mengalami penyakit jantung koroner dalam periode waktu 10 tahun. Nilai variabel ini diklasifikasikan dengan 1 sebagai risiko positif dan 0 sebagai risiko negatif.

2.2 *Preprocessing*

Data preprocessing mencakup serangkaian langkah untuk membersihkan, mengorganisir, dan mentransformasikan data mentah agar lebih sesuai untuk analisis. Tahap ini berperan penting dalam memastikan bahwa data berada dalam kondisi optimal, terbebas dari kesalahan, serta siap digunakan dalam pemodelan dan pengambilan keputusan [24].

2.2.1 *Data Cleaning*

Data cleaning adalah tahap awal *preprocessing* dengan tujuan untuk memastikan bahwa dataset yang digunakan bebas dari kesalahan atau ketidaksesuaian data. Tahap ini menjadi langkah krusial dalam proses analisis data, mencakup identifikasi, penanganan, serta penghapusan nilai yang hilang (*missing values*), pencilan (*outliers*), dan anomali dalam dataset [24]. Pada penelitian ini, proses *data cleaning* diawali dengan pengecekan nilai yang hilang atau kosong (*missing values*) dalam dataset yang digunakan. Hasil pengecekan menunjukkan adanya *missing values* pada beberapa fitur, sehingga diperlukan metode penanganan untuk mengatasi permasalahan tersebut. Dalam penelitian ini, metode yang diterapkan untuk menangani *missing values* adalah menggantinya dengan nilai rata-rata (*mean imputation*). Metode ini dipilih karena kesederhanaannya dan kemampuannya dalam menjaga ukuran sampel, serta mengurangi potensi bias yang mungkin terjadi akibat penghapusan data. Namun, penggunaan imputasi nilai rata-rata dapat mengurangi variabilitas data, sehingga diperlukan evaluasi model lebih lanjut untuk memastikan bahwa performa tetap optimal dan distribusi data tetap representatif.

2.2.2 Normalisasi

Normalisasi data pada penelitian ini dilakukan dengan menggunakan *Z-Score normalization* untuk melakukan rescaling, yang bertujuan untuk mengurangi dampak outlier dan memastikan bahwa semua variabel memiliki skala yang konsisten, sehingga dapat meningkatkan akurasi model klasifikasi [25]. *Z-Score normalization* dapat

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2582−2593
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.7029

mempertahankan distribusi data asli dan lebih efektif dalam menangani outlier dibandingkan dengan *Min-Max scaling*, yang membatasi nilai dalam rentang tertentu [26]. Oleh karena itu, pendekatan tersebut dipilih untuk memastikan bahwa skala variabel menjadi seragam tanpa menghilangkan informasi penting yang diperlukan dalam proses klasifikasi. Persamaan *Z-Score normalization* dapat dilihat pada persamaan (1).

$$x' = \frac{x_i - mean(x)}{std(x)} \quad (1)$$

Pada persamaan (1), x' adalah nilai hasil normalisasi, sedangkan x_i merepresentasikan nilai yang akan dinormalisasi atau nilai asli sebelum dinormalisasi, $mean(x)$ untuk nilai rata-rata dari sebuah atribut, dan $std(x)$ untuk nilai standar deviasi dari sebuah atribut.

2.2.3 *Resampling*

Resampling adalah metode yang diterapkan untuk menangani ketidakseimbangan kelas dengan menggunakan pendekatan oversampling dan undersampling. [27]. *Oversampling* bertujuan untuk menambah jumlah data pada kelas minoritas dengan menduplikasi data atau menambah informasi baru antar data dalam kelas tersebut [28]. Sebaliknya, pada *undersampling* jumlah data pada kelas mayoritas akan dikurangi dengan menghapus atau mengganti beberapa data, sehingga distribusi data menjadi lebih seimbang [29]. Untuk mengoptimalkan penanganan ketidakseimbangan kelas, penelitian ini merekomendasikan penerapan metode *Random Over-Sampling*. Teknik ini meningkatkan atau menambahkan jumlah data pada kelas minoritas dengan cara mereplikasi *instance* secara acak, sehingga menghasilkan distribusi data yang lebih proporsional dan lebih representatif terhadap kelas minoritas.

2.2.4 Pembagian Data

Data yang telah melalui serangkaian proses *preprocessing* kemudian memasuki tahap pembagian data. Pembagian data menjadi set pelatihan (*training set*) dan set pengujian (*testing set*) merupakan langkah krusial dalam pembangunan model pada pembelajaran mesin. Tahap ini bertujuan untuk menilai sejauh mana model yang dikembangkan mampu menggeneralisasi terhadap data yang belum pernah ditemui sebelumnya [24]. Dalam penelitian ini, dataset dibagi dengan ukuran 80:20, dengan 80% (3.392 data) digunakan untuk *training set* dan 20% (848 data) digunakan untuk *testing set*.

2.3 Implementasi Algoritma

Pada penelitian ini, proses implementasi algoritma dilakukan dengan menggunakan dua algortima klasifikasi, yaitu algoritma decision tree dan algoritma random forest. Tujuan utama dari penelitian ini adalah untuk membandingkan kinerja kedua algoritma tersebut dalam rangka menentukan model dengan performa terbaik. Evaluasi dilakukan berdasarkan metrik tertentu untuk mengukur efektivitas masing-masing algoritma dalam melakukan klasifikasi.

2.3.1 *Decision Tree*

Decision tree merupakan sebuah algoritma klasifikasi yang sederhana, populer, dan mudah diinterpretasikan. Algoritma ini bertujuan untuk menciptakan model yang dapat memprediksi kelas atau nilai variabel target dengan mempelajari aturan keputusan yang terdapat dalam data pelatihan [30]. *Decision tree* membangun model dalam bentuk struktur pohon, dimana setiap *node* merepresentasikan kondisi pada suatu fitur, setiap *branch* menunjukkan hasil dari kondisi tersebut, dan setiap *leaf* menggambarkan hasil akhir klasifikasi [31], [32].

$$Entropy(S) = -\sum_{i=1}^{n} Pi\ log_2\ Pi \quad (2)$$

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{S_i}{S} * Entropy(Si) \quad (3)$$

Langkah awal dalam algoritma *decision tree* adalah menghitung *entropy* untuk mengukur tingkat ketidakpastian dalam dataset. Pendefinisian *entropy* dapat dilihat pada persamaan (2). Dimana S adalah himpunan data pada simpul yang sedang dievaluasi, n adalah jumlah kelas dalam dataset, dan Pi adalah probabilitas kemunculan kelas ke-i. Setelah menghitung *entropy*, langkah berikutnya adalah menentukan *information gain*, yang mengukur sejauh mana ketidakpastian dalam dataset berkurang setelah pemisahan berdasarkan suatu atribut. *Information gain* dihitung menggunakan rumus pada persamaan (3). Dimana $Entropy(S)$ adalah *entropy* awal dataset sebelum pemisahan, n adalah jumlah subset yang terbentuk dari pembagian dataset berdasarkan atribut A, $|\ S_i\ |$ adalah jumlah sampel dalam subset S_i, $|\ S\ |$ adalah jumlah total sampel dalam dataset, dan $Entropy(S_i)$ adalah *entropy* dari subset S_i.

2.3.2 *Random Forest*

Random forest adalah metode klasifikasi yang mengombinasikan hasil dari beberapa pohon keputusan untuk meningkatkan akurasi dan mengurangi kemungkinan terjadinya *overfitting*. Algoritma ini menggunakan teknik *bootstrap sampling* untuk membangun beberapa pohon dari subset acak dataset, dimana setiap pohon dibangun secara independen dengan pemilihan fitur yang bervariasi [33]. Hasil prediksi akhir ditentukan melalui *voting* mayoritas dari semua pohon dalam *ensemble*.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2582−2593
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.7029

2.4 *Hyperparameter Tuning*

Dalam penelitian ini, dilakukan *hyperparameter tuning* untuk mengoptimalkan performa algoritma. Berbagai kombinasi parameter diuji pada dataset, kemudian dipilih parameter dengan hasil terbaik untuk setiap dataset. Pendekatan ini bertujuan untuk meningkatkan akurasi model dan memastikan kinerja yang optimal dalam proses klasifikasi. Proses *tuning* diawali dengan mendefinisikan himpunan nilai untuk setiap *hyperparameter*. Untuk algoritma *decision tree*, parameter yang disesuaikan meliputi *max_depth* (kedalaman maksimum pohon), *min_samples_split* (jumlah minimal sample yang dibutuhkan untuk membagi *node*), dan *min_samples_leaf* (jumlah minimum sample pada *node* daun). Sementara itu, untuk *random forest*, selain parameter yang sama dengan *decision tree*, juga dilakukan tuning pada *n_estimators*, yaitu jumlah pohon dalam *ensemble*.

Tabel 1. *Parameter Tuning*

Algoritma	*max_depth*	*min_samples_split*	*min_samples_leaf*	*n_estimators*
Decision Tree	[None, 5, 10, 15]	[2, 5, 10]	[1, 2, 4]	-
Random Forest	[None, 5, 10, 15]	[2, 5, 10]	[1, 2, 4]	[150, 300]

Tabel 1 menunjukkan nilai setiap parameter yang digunakan dalam penelitian tersebut. Dengan menyesuaikan parameter-parameter ini, model diharapkan mampu menghasilkan prediksi yang lebih akurat serta menghindari permasalahan *overfitting* atau *underfitting* dalam proses klasifikasi.

2.5 Evaluasi Hasil

Confusion matrix merupakan sebuah alat evaluasi yang digunakan untuk mengukur tingkat kinerja dari suatu algoritma klasifikasi. Alat ini memberikan wawasan yang lebih mendalam mengenai prediksi yang benar dan salah, serta menyajikan rincian performa model klasifikasi dengan menampilkan jumlah prediksi yang benar dan salah pada setiap kelas [34].

Tabel 2. *Confusion Matrix*

	Positive class predicted	*Negative class predicted*
Positive class actual	*True positive*	*False negative*
Negative class actual	*False positive*	*True negative*

Rincian *confusion matrix* dapat dilihat pada Tabel 2. Selain itu, analisis menggunakan *confusion matrix* memungkinkan perhitungan berbagai *evaluation matrix* yang penting, seperti *accuracy*, *precision*, *recall*, dan *F1-score.* Persamaan (4) menyajikan rumus untuk menghitung *accuracy*.

$$x_i^{t+1} = x_i^t + v_i^{t+1} \tag{4}$$

Dari Persamaan (4), *accuracy* didefinisikan dari rasio sampel penyakit jantung koroner yang terklasifikasi dengan benar terhadap total sampel penyakit jantung koroner. Persamaan (5) menampilkan *F1-score*, yang dihitung sebagai rata-rata aritmatika antara *precision* dan *recall*, dimana *precision* didefinisikan pada Persamaan (6), dan *recall* didefinisikan pada Persamaan (7).

$$F1 - score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{5}$$

Persamaan (5) menampilkan *F1-score*, yang dihitung sebagai rata-rata dari *precision* dan *recall*. *Precision* dapat dilihat pada Persamaan (6), dimana perhitungan dilakukan dengan mengukur proporsi prediksi positif yang benar dibandingkan dengan seluruh prediksi positif yang dibuat. Sementara itu, *recall*, yang didefinisikan dalam Persamaan (7), mengukur sejauh mana model mampu mendeteksi seluruh sampel positif yang ada dalam dataset. *Recall* menjadi penting dalam situasi di mana kesalahan dalam mendeteksi sampel positif memiliki konsekuensi yang signifikan, seperti dalam diagnosis penyakit.

$$Precision = \frac{TP}{TP+FP} \tag{6}$$

$$Recall = \frac{TP}{TP+FN} \tag{7}$$

3. HASIL DAN PEMBAHASAN

3.1 Pengumpulan Data

Penelitian ini menggunakan dataset yang berasal dari situs penyedia dataset online, yaitu Kaggle, yang menyajikan data dari *National Heart, Lung, and Blood Institute* (NHLBI) yang bekerja sama dengan *Boston University*. Dataset

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2582−2593
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.7029

ini dikenal sebagai *Framingham Heart Study Dataset*, sebuah studi *longitudinal* yang bertujuan untuk menganalisis faktor-faktor yang berhubungan dengan penyakit jantung. Dataset ini terdiri dari 4.240 *records* yang merepresentasikan individu peserta studi, masing-masing dengan satu variabel target dan 15 variabel fitur yang berkaitan dengan risiko penyakit jantung. Variabel target dalam dataset ini adalah *TenYearCHD*, yang menunjukkan kemungkinan seseorang untuk mengalami penyakit jantung koroner dalam periode 10 tahun ke depan. Sedangkan, 15 variabel fitur yang ada mencakup informasi demografis, gaya hidup, serta faktor kesehatan yang berperan dalam mempengaruhi risiko penyakit jantung, seperti usia, jenis kelamin, tekanan darah, kadar kolesterol, kebiasaan merokok, dan lain sebagainya. Untuk pemahaman lebih lanjut mengenai setiap fitur yang ada, dapat dilihat pada Tabel 3, yang menyediakan penjelasan rinci mengenai definisi dan satuan pengukuran masing-masing variabel dalam dataset ini.

Tabel 3. *Attributes of Framingham dataset*

No	*Attributes*	*Description*
1.	*Sex*	*0 : Male, 1 : Female*
2.	*age*	*Age of the patient in years*
3.	*currentSmoker*	*0 : If the patient not a current smoker, 1 : If current smoker*
4.	*cigsPerDay*	*Number of cigarettes the patient smokes per day*
5.	*BPMeds*	*0 : Not on BP medication, 1 : On BP medication*
6.	*prevalentStroke*	*0 : No previous stroke, 1 : Previous stroke*
7.	*prevalentHyp*	*0 : Not hypertensive, 1 : Hypertensive*
8.	*totChol*	*Total cholesterol level*
9.	*sysBP*	*Systolic blood pressure*
10.	*diaBP*	*Diastolic blood pressure*
11.	*diabetes*	*0 : Non-diabetic, 1 : Diabetic*
12.	*BMI*	*Body Mass Index*
13.	*heartRate*	*Heart rate*
14.	*glucose*	*Glucose level*
15.	*education*	*Education level of the person*
16.	*TenYearCHD*	*0 : No risk of CHD, 1 : Risk of CHD over 10 years*

Tabel 3 menjelaskan 16 atribut yang digunakan untuk memprediksi risiko penyakit jantung coroner dalam 10 tahun. Atribut terdiri dari faktor demografis (*sex*, *age*, *education*), gaya hidup (*currentSmoker*, *cigsPerDay*), serta indikator medis seperti tekanan darah (*sysBP*, *diaBP*), kadar kolesterol (*totChol*), diabetes, BMI, detak jantung, dan glukosa. Variabel target *TenYearCHD* menunjukkan adanya risiko CHD. Dataset Framingham ini memiliki potensi yang sangat besar dalam pengembangan model prediksi risiko penyakit jantung, serta dapat digunakan untuk analisis faktor-faktor risiko yang lebih dalam.

3.2 *Preprocessing*

Pada penelitian ini, tahap *preprocessing* mencakup berbagai langkah, yaitu menangani nilai yang hilang (*missing values*), melakukan normalisasi atau standardisasi fitur, menyeimbangkan distribusi kelas, serta membagi data untuk analisis lebih lanjut. Dengan melakukan *preprocessing* secara tepat, kualitas data dapat ditingkatkan sehingga model yang dibangun memiliki performa yang lebih optimal dan hasil yang lebih akurat.

3.2.1 *Data Cleaning*

Langkah pertama dalam tahap *preprocessing* adalah pembersihan data (*data cleaning*). Pada tahap ini, penting untuk mengidentifikasi dan menangani atribut atau fitur yang mengandung *missing value* atau nilai yang hilang, karena nilai yang hilang dapat memengaruhi kualitas analisis dan pemodelan. Berdasarkan analisis yang dapat dilihat pada Gambar 2, terdapat beberapa fitur atau atribut dalam dataset yang mengandung *missing value*, di antaranya adalah *glucose*, *education*, *BPMeds*, *totChol*, *cigsPerDay*, *BMI*, dan *heartRate*. Dari semua atribut tersebut, *glucose* memiliki jumlah *missing value* terbanyak dibandingkan atribut lainnya. Atribut-atribut ini perlu ditangani secara hati-hati untuk memastikan bahwa data yang digunakan tetap memiliki kualitas yang baik dan terjaga. Rincian iInformasi tersebut dapat dilihat pada gambar 2.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2582−2593
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.7029

glucose	388
education	105
BPMeds	53
totChol	50
cigsPerDay	29
BMI	19
heartRate	1

Gambar 2. Atribut Dengan *Missing Value*

Missing value yang terdapat pada atribut-atribut dalam dataset akan diatasi dengan teknik imputasi berbasis rata-rata yang dibedakan berdasarkan dua kelompok data, yaitu kelompok yang terindikasi berisiko terkena penyakit jantung koroner (*TenYearCHD* = 1) dan kelompok yang tidak berisiko (*TenYearCHD* = 0). Pendekatan ini dipilih untuk mempertahankan informasi yang ada dalam dataset, sekaligus menghindari kehilangan data yang dapat terjadi jika entri dengan nilai kosong dihapus. Dengan melakukan imputasi berdasarkan rata-rata masing-masing kelompok, distribusi data tetap representatif dan tidak menggeser pola hubungan antara variabel. Pendekatan ini juga memastikan bahwa nilai yang diisi lebih relevan dengan karakteristik kelompoknya, sehingga membantu model dalam memahami hubungan antar variabel dengan lebih baik. Selain itu, teknik ini dapat mengurangi potensi bias yang mungkin timbul jika semua nilai kosong diisi dengan rata-rata keseluruhan dataset tanpa mempertimbangkan perbedaan antara kelompok berisiko dan tidak berisiko. Tujuan utama dari proses imputasi ini adalah untuk memastikan bahwa model tidak kehilangan potensi informasi penting dari atribut yang mengalami *missing value*. Dengan demikian, model dapat tetap bekerja dengan data yang lebih lengkap tanpa mengorbankan akurasi prediksi. Hasil dari proses imputasi ini dapat dilihat pada Gambar 3, dimana atribut-atribut yang sebelumnya memiliki *missing value* telah terisi sepenuhnya, memungkinkan analisis lebih lanjut dilakukan dengan data yang lebih terstruktur dan informatif.

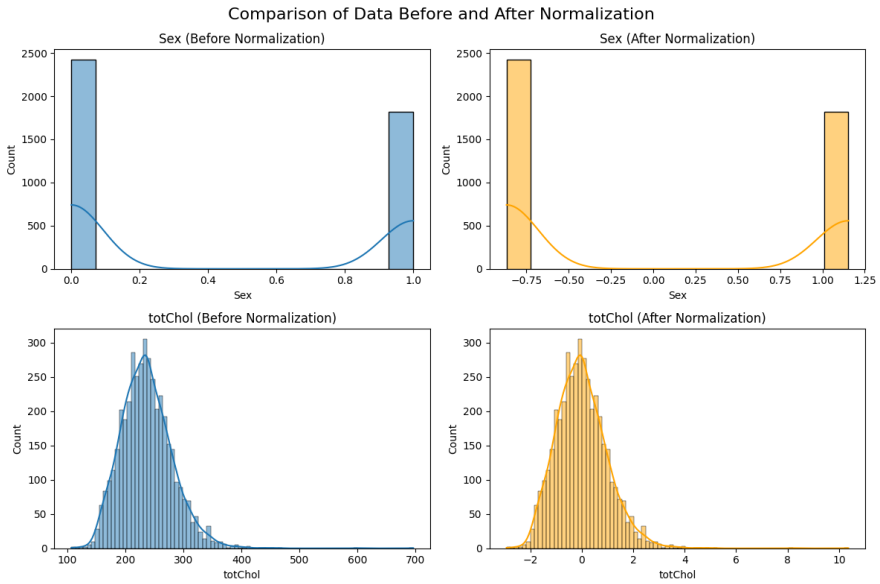
education	0
cigsPerDay	0
BPMeds	0
totChol	0
BMI	0
heartRate	0
glucose	0

Gambar 3. Hasil Penanganan *Missing Value*

3.2.2 Normalisasi

Penelitian ini menerapkan *Z-Score normalization* untuk menstandarisasi setiap fitur berdasarkan rata-rata ($mean(x)$) dan standar deviasi ($std(x)$). Metode ini mengubah distribusi data sehingga memiliki rata-rata nol dan standar deviasi satu, memastikan bahwa setiap fitur berada dalam skala yang sebanding. Dataset yang digunakan dalam penelitian ini memiliki fitur dengan rentang nilai yang bervariasi, seperti yang terlihat pada Gambar 5, dimana beberapa fitur memiliki skala yang jauh lebih besar dibandingkan dengan yang lainnya. Perbedaan skala ini dapat menyebabkan fitur dengan nilai yang lebih besar mendominasi proses pembelajaran model, yang dapat mengarah pada bias dan menurunnya kinerja model. Untuk mengatasi permasalahan ini, dilakukan normalisasi data agar setiap fitur memiliki skala yang lebih seimbang dalam proses klasifikasi. Dengan normalisasi, kontribusi setiap fitur menjadi lebih proporsional, yang pada gilirannya meningkatkan akurasi model dan mengoptimalkan hasil prediksi. Hasil dari proses normalisasi dapat diamati dalam Gambar 4, yang menampilkan perbandingan antara data sebelum dan sesudah normalisasi pada beberapa fitur. Dari gambar tersebut, terlihat bahwa sebelum normalisasi, distribusi nilai dari beberapa fitur sangat bervariasi dengan rentang yang lebar. Namun, setelah dilakukan normalisasi, semua fitur memiliki distribusi yang lebih seimbang, yang memungkinkan model untuk belajar dengan lebih efektif. Dengan demikian, penerapan *Z-Score normalization* dalam penelitian ini tidak hanya berfungsi untuk menstandarisasi data tetapi juga berperan penting dalam meningkatkan kinerja model pembelajaran mesin, sehingga menghasilkan prediksi yang lebih optimal dan akurat.

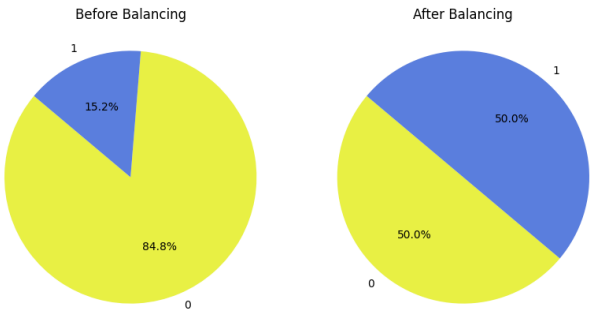
Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2582−2593
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.7029



Gambar 4. Hasil Normalisasi Pada Atribut *Sex* dan *totChol*

3.2.3 *Resampling (Balancing)* dan Pembagian Data

Pada tahap ini proses penyeimbangan distribusi kelas dilakukan untuk mengatasi masalah ketidakseimbangan distribusi kelas yang ada pada dataset. Hal tersebut terjadi ketika jumlah data pada satu kategori lebih banyak atau lebih sedikit dibandingkan kategori lainnya. Ketidakseimbangan ini dapat mengarah pada prediksi yang tidak akurat, dimana model lebih cenderung memprioritaskan kelas dengan jumlah data lebih besar. Untuk mengatasi masalah ini, diterapkan metode *Random Over-Sampling*, yang bertujuan untuk meningkatkan jumlah data pada kelas minoritas dengan menggandakan atau menambah sampel secara acak, sehingga proporsi antar kelas menjadi lebih seimbang. Proses penyeimbangan data diawali dengan pemilihan sampel dari kelas minoritas secara acak dengan penggantian, yang memungkinkan satu sampel dipilih lebih dari satu kali. Selanjutnya, sampel hasil duplikasi dari kelas minoritas ditambahkan kembali ke dalam dataset tanpa mengubah informasi aslinya. Dengan cara ini, distribusi data menjadi lebih seimbang, sehingga model tidak condong terhadap kelas mayoritas dan dapat melakukan prediksi dengan lebih akurat.



Gambar 5. Hasil *Class Balancing* Pada Target

Gambar 5 menampilkan distribusi kelas dalam dataset sebelum dan sesudah proses penyeimbangan data menggunakan teknik *Random Over-Sampling*. Pada grafik pertama (kiri), terlihat bahwa distribusi kelas awal tidak seimbang, dimana jumlah sampel pada kelas 0 mencapai 84,8% (3.596 data), sementara kelas 1 hanya sebesar 15,2% (644 data). Ketidakseimbangan ini dapat menyebabkan model cenderung lebih berpihak pada kelas mayoritas, yang berpotensi mengurangi akurasi prediksi pada kelas minoritas. Setelah dilakukan proses penyeimbangan data menggunakan *Random Over-Sampling*, distribusi kelas menjadi lebih proporsional, seperti yang ditunjukkan pada grafik kedua (kanan). Teknik ini bekerja dengan menggandakan sampel dari kelas minoritas secara acak hingga jumlahnya setara dengan kelas mayoritas. Dengan demikian, jumlah sampel antara kelas 0 dan kelas 1 menjadi seimbang, yang dapat meningkatkan kinerja model dalam proses klasifikasi karena model tidak lagi terpengaruh oleh distribusi data yang tidak merata.

Tahap selanjutnya setelah proses penyeimbangan data adalah pembagian dataset untuk keperluan pelatihan dan pengujian model. Dalam penelitian ini, dataset dibagi dengan proporsi 80:20, dimana 80% data (3.392 sampel) digunakan sebagai *training set* untuk melatih model, sementara 20% sisanya (848 sampel) digunakan sebagai *testing set* untuk mengevaluasi performa model. Pembagian ini bertujuan untuk memastikan bahwa model mendapatkan

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2582−2593
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.7029

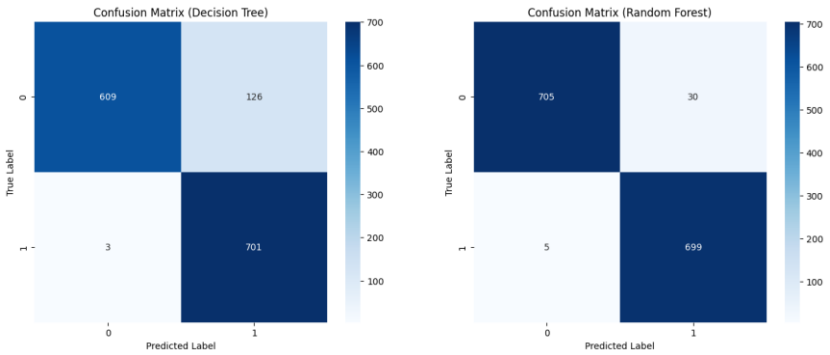
cukup data untuk belajar sekaligus diuji pada data yang belum pernah dilihat sebelumnya. Dengan demikian, hasil evaluasi model menjadi lebih objektif dan dapat mencerminkan kinerjanya dalam kondisi nyata.

3.3 Implementasi Algoritma

Setelah seluruh tahapan *preprocessing* data selesai, langkah berikutnya adalah membangun model klasifikasi menggunakan algoritma yang telah ditentukan, yaitu *Decision Tree* dan *Random Forest*. Model dikembangkan dengan melakukan optimasi *hyperparameter* guna meningkatkan performa klasifikasi. *Decision Tree* membangun struktur pohon berdasarkan fitur dengan informasi terbaik, dengan parameter utama yaitu *max_depth*, *min_samples_split*, dan *min_samples_leaf*. *Hyperparameter tuning* dilakukan untuk memperoleh kombinasi parameter yang optimal. Sementara itu, *Random Forest* sebagai model *ensemble* terdiri dari beberapa pohon keputusan yang bekerja bersama untuk meningkatkan akurasi prediksi. Parameter yang dioptimalkan mencakup *max_depth*, *min_samples_split*, *min_samples_leaf*, dan *n_estimators*.

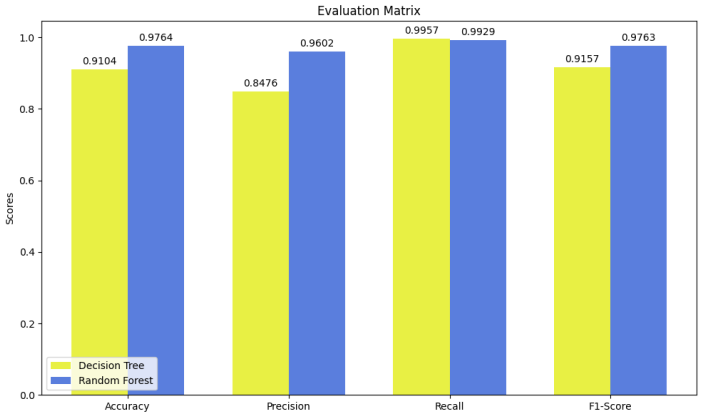
3.4 Evaluasi Hasil Pengujian

Setelah seluruh tahapan sebelumnya selesai, langkah selanjutnya adalah mengevaluasi model yang telah dibangun menggunakan *confusion matrix*. Evaluasi ini bertujuan untuk menilai sejauh mana model mampu mengklasifikasikan data dengan benar setelah melalui proses pelatihan. Dalam penelitian ini, model yang diuji terdiri dari dua algoritma utama, yaitu *Decision Tree* dan *Random Forest*. *Confusion matrix* memberikan gambaran mengenai jumlah prediksi yang benar dan salah dalam setiap kelas, yang kemudian digunakan untuk menghitung berbagai *evaluation matrix* seperti *accuracy*, *precision*, *recall*, dan *F1-score*. Metrik ini penting untuk menilai efektivitas model dalam melakukan klasifikasi, terutama dalam kasus dimana keseimbangan antara kelas positif dan negatif sangat berpengaruh terhadap hasil prediksi. Hasil evaluasi menggunakan *confusion matrix* ditampilkan pada Gambar 6, dimana model *Decision Tree* memiliki 3 *false positive* dan 126 *false negative*. Sementara itu, model *Random Forest* menunjukkan hasil dengan 5 *false positive* dan 30 *false negative*. Dari perbandingan ini, dapat dilihat bahwa *Random Forest* memiliki jumlah *false negative* yang lebih rendah dibandingkan *Decision Tree*, yang menunjukkan bahwa model ini lebih baik dalam mengidentifikasi kelas positif.



Gambar 6. Hasil *Confusion Matrix*

Berdasarkan data *confusion matrix* pada Gambar 6, dilakukan perhitungan lebih lanjut untuk mengevaluasi kinerja model menggunakan metrik-metrik evaluasi yang telah disebutkan. Hasil perhitungan tersebut disajikan pada Gambar 7, yang memberikan gambaran menyeluruh mengenai kemampuan model dalam mengklasifikasikan kelas positif maupun negatif secara akurat dengan menampilkan berbagai *evaluation matrix* seperti *accuracy*, *precision*, *recall*, dan *F1-score*.



Gambar 7. Hasil *Evaluation Matrix*

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2582−2593
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.7029

Berdasarkan hasil evaluasi yang ditampilkan pada Gambar 7, model *random forest* menunjukkan performa yang lebih unggul dibandingkan dengan *decision tree* dalam tugas klasifikasi. Hal ini dapat dilihat dari *evaluation matrix* yang lebih tinggi pada model *random forest* dibandingkan dengan *decision tree*. Pada model *decision tree*, diperoleh hasil *accuracy* sebesar 0.9104, *precision* 0.8476, *recall* 0.9957, dan *F1-score* 0.9157, sementara model *random forest* menunjukkan peningkatan yang signifikan dengan *accuracy* 0.9764, *precision* 0.9602, *recall* 0.9929, dan *F1-score* 0.9763. Meskipun secara keseluruhan model *random forest* menunjukkan kinerja yang lebih baik dibandingkan *decision tree*, dengan peningkatan yang signifikan pada *accuracy*, *precision*, dan *F1-score*, terdapat sedikit penurunan pada *recall*-nya. *Recall* yang lebih tinggi menunjukkan bahwa model memiliki sensitivitas yang lebih besar dalam mendeteksi penderita penyakit jantung koroner, yang berarti model lebih efektif dalam mengurangi *false negatives* (kasus yang seharusnya teridentifikasi sebagai positif tetapi tidak terdeteksi). Meskipun model *random forest* lebih unggul dalam mengurangi kesalahan klasifikasi secara keseluruhan dan menghasilkan prediksi yang lebih akurat, model ini cenderung lebih selektif dalam mengklasifikasikan individu sebagai positif. Selektivitas ini dapat berdampak pada penurunan *recall*, meskipun pada saat yang sama meningkatkan *precision*, karena model lebih berhati-hati dalam memprediksi kasus positif untuk menghindari kesalahan klasifikasi positif palsu (*false positives*). Secara keseluruhan, peningkatan performa pada *random forest* mencerminkan kemampuan yang lebih baik dalam menangkap pola kompleks dalam data, memberikan hasil yang lebih stabil dan generalisasi yang lebih baik dibandingkan dengan model *decision tree*.

Keunggulan *random forest* ini dapat dikaitkan dengan sifatnya sebagai *ensemble learning model*, dimana keputusan akhir dibuat berdasarkan agregasi dari beberapa pohon keputusan, sehingga mengurangi risiko *overfitting* yang sering terjadi pada model *decision tree*. Selain itu, salah satu keuntungan utama dari *random forest* adalah kemampuannya dalam menangani variabel dengan skala yang berbeda serta kemampuannya untuk menangani data yang memiliki dimensi tinggi. Dengan fitur tersebut, *random forest* dapat digunakan untuk berbagai jenis data, termasuk data dengan banyak fitur yang saling berkorelasi. Hal ini membuat *random forest* menjadi model yang lebih fleksibel dan adaptif dalam berbagai skenario klasifikasi. Meskipun *decision tree* mempunyai keunggulan pada interpretabilitas serta kecepatan dalam melakukan inferensi, model ini cenderung mengalami *overfitting* ketika diterapkan pada data dengan kompleksitas tinggi. *Overfitting* terjadi saat model terlalu beradaptasi dengan data pelatihan, sehingga kemampuan untuk menggeneralisasi data baru menjadi berkurang. Sebaliknya, r*andom forest* dengan pendekatan *ensemble learning* mampu mengatasi permasalahan *overfitting* dengan menggabungkan prediksi dari banyak pohon keputusan, sehingga model yang dihasilkan menjadi lebih stabil dan akurat.

Pada penelitian ini, hasil evaluasi yang diperoleh menunjukkan bahwa pemilihan model klasifikasi yang tepat sangat penting untuk mencapai hasil yang optimal. *Decision tree* mungkin cocok untuk kasus dimana interpretabilitas model lebih diutamakan, tetapi jika tujuan utama adalah mendapatkan akurasi tinggi dan mengurangi kesalahan klasifikasi, maka *random forest* adalah pilihan yang lebih baik. Dengan mempertimbangkan hasil evaluasi ini, penggunaan *random forest* direkomendasikan untuk tugas klasifikasi serupa di masa depan.

4. KESIMPULAN

Evaluasi kinerja model *decision tree* dan *random forest* dilakukan dengan menggunakan berbagai *evaluation matrix*, yaitu *accuracy*, *recall*, *precision*, dan *F1-score*. Selain itu, analisis lebih lanjut dilakukan dengan memanfaatkan *confusion matrix* untuk mendapatkan pemahaman yang lebih mendalam mengenai performa masing-masing model. Hasil evaluasi menunjukkan bahwa model *random forest* memiliki kinerja yang lebih baik dibandingkan dengan *decision tree* dalam memprediksi risiko penyakit jantung koroner dalam periode 10 tahun. Model *random forest* mencapai *accuracy* sebesar 97,64%, sementara *decision tree* hanya memperoleh 91,04%. Dari segi *precision*, *random forest* juga menunjukkan keunggulan dengan nilai sebesar 96,02%, sedangkan *decision tree* hanya mencapai 84,76%. Meskipun terjadi sedikit penurunan pada *recall*, dimana nilai *recall* random forest sebesar 99,29%, sedikit lebih rendah dibandingkan dengan nilai *recall decision tree* yang mencapai 99,57%, perbedaan ini tidak signifikan dalam konteks keseluruhan. Keunggulan *random forest* lebih jelas terlihat dari nilai *F1-score* yang mencapai 97,63%, dibandingkan *decision tree* yang hanya memperoleh 91,57%. *Confusion matrix* menunjukkan bahwa *random forest* memiliki tingkat kesalahan dalam klasifikasi yang lebih rendah jika dibandingkan dengan *decision tree*, terutama dalam *precision* dan *F1-score*. Namun, *decision tree* sedikit lebih unggul dalam hal *recall*, menunjukkan kemampuannya yang sedikit lebih tinggi dalam mendeteksi semua kasus positif. Berdasarkan hasil evaluasi, dapat disimpulkan bahwa random forest merupakan model yang lebih optimal dalam memprediksi risiko penyakit jantung koroner dalam periode 10 tahun karena kemampuannya dalam menangani variasi data serta mengurangi risiko *overfitting* melalui kombinasi berbagai pohon keputusan dalam proses klasifikasinya. Namun, penting untuk dicatat bahwa kinerja model masih sangat bergantung pada dataset yang digunakan serta parameter yang diterapkan. Oleh karena itu, pada penelitian selanjutnya optimalisasi lebih lanjut perlu dilakukan, seperti penyesuaian *hyperparameter* yang lebih bervariasi serta penerapan teknik optimalisasi lainnya dapat dilakukan untuk meningkatkan performa model secara keseluruhan.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2582−2593
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.7029

REFERENCES

[1] F. Andika, N. Afriza, A. Husna, N. Rahmi, and F. Safitri, "Edukasi Tentang Isu Permasalahan Kesehatan Di Indonesia Bersama Calon Tenaga Kesehatan Masyarakat Provinsi Aceh," 2022.
[2] Institute for Health Metrics and Evaluation (IHME), "Global Burden of Disease 2021," 2021. Accessed: Mar. 05, 2025. [Online]. Available: https://www.healthdata.org/research-analysis/library/global-burden-disease-2021-findings-gbd-2021-study
[3] World Health Organization, "Cardiovascular diseases (CVDs)." Accessed: Jan. 19, 2025. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)
[4] G. A. Mensah *et al.*, "Global Burden of Cardiovascular Diseases and Risks, 1990-2022," *J Am Coll Cardiol*, vol. 82, no. 25, pp. 2350–2473, Dec. 2023, doi: 10.1016/j.jacc.2023.11.007.
[5] C. W. Tsao *et al.*, "Heart Disease and Stroke Statistics-2022 Update: A Report from the American Heart Association," Feb. 22, 2022, *Lippincott Williams and Wilkins*. doi: 10.1161/CIR.0000000000001052.
[6] M. Di Cesare *et al.*, "The Heart of the World," *Glob Heart*, vol. 19, no. 1, 2024, doi: 10.5334/gh.1288.
[7] International Diabetes Federation (IDF), "IDF Diabetes Atlas 10th edition.," 2021. Accessed: Mar. 05, 2025. [Online]. Available: https://diabetesatlas.org/atlas/tenth-edition/
[8] C. Krittanawong *et al.*, "Deep learning for cardiovascularmedicine: A practical primer," Jul. 01, 2019, *Oxford University Press*. doi: 10.1093/eurheartj/ehz056.
[9] V. D. Nagarajan, S. L. Lee, J. L. Robertus, C. A. Nienaber, N. A. Trayanova, and S. Ernst, "Artificial intelligence in the diagnosis and management of arrhythmias," Oct. 07, 2021, *Oxford University Press*. doi: 10.1093/eurheartj/ehab544.
[10] K. Seetharam *et al.*, "Applications of Machine Learning in Cardiology," Sep. 01, 2022, *Adis*. doi: 10.1007/s40119-022-00273-7.
[11] A. Roihan, P. Abas Sunarya, and A. S. Rafika, "Pemanfaatan Machine Learning dalam Berbagai Bidang: Review paper," *IJCIT (Indonesian Journal on Computer and Information Technology)*, vol. 5, no. 1, pp. 75–82, 2019.
[12] Z. I. Attia *et al.*, "An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction," *The Lancet*, vol. 394, no. 10201, pp. 861–867, Sep. 2019, doi: 10.1016/S0140-6736(19)31721-0.
[13] R. G. Wardhana, G. Wang, and F. Sibuea, "Penerapan Machine Learning dalam Prediksi Tingkat Kasus Penyakit Di Indonesia," *Journal of Information System Management (JOISM) e-ISSN*, vol. 5, no. 1, pp. 2715–3088, 2023.
[14] E. Retnoningsih and R. Pramudita, "Mengenal Machine Learning Dengan Teknik Supervised dan Unsupervised Learning Menggunakan Python," *BINA INSANI ICT JOURNAL*, vol. 7, no. 2, pp. 156–165, 2020, [Online]. Available: https://www.python.org/
[15] National Heart Lung and Blood Institute (NHLBI), "Framingham heart study dataset," Kaggle. Accessed: Nov. 11, 2024. [Online]. Available: https://www.kaggle.com/datasets/aasheesh200/framingham-heart-study-dataset
[16] A. Wulan, N. Dari, and I. N. Fajri, "Penerapan Algoritma K-Nearest Neighbor (KNN) Untuk Klasifikasi Resiko Penyakit Jantung," *Journal of Information System Research*, vol. 6, no. 1, pp. 428–436, 2024, doi: 10.47065/josh.v6i1.6038.
[17] E. S. Kresnawati, Y. Resti, B. Suprihatin, M. R. Kurniawan, and W. A. Amanda, "Coronary Artery Disease Prediction Using Decision Trees and Multinomial Naïve Bayes with k-Fold Cross Validation," *Inovasi Matematika (Inomatika)*, vol. 3, no. 2, pp. 172–187, 2021, doi: 10.35438/inomatika.
[18] P. Gupta and D. Seth, "Comparative analysis and feature importance of machine learning and deep learning for heart disease prediction," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 29, no. 1, pp. 451–459, Jan. 2023, doi: 10.11591/ijeecs.v29.i1.pp451-459.
[19] Rian Oktafiani, Arief Hermawan, and Donny Avianto, "Max Depth Impact on Heart Disease Classification: Decision Tree and Random Forest," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 8, no. 1, pp. 160–168, Feb. 2024, doi: 10.29207/resti.v8i1.5574.
[20] R. J. Suhatril, R. D. Syah, M. Hermita, B. Gunawan, and W. Silfianti, "Evaluation of Machine Learning Models for Predicting Cardiovascular Disease Based on Framingham Heart Study Data," *ILKOM Jurnal Ilmiah*, vol. 16, no. 1, pp. 68–75, Apr. 2024, doi: 10.33096/ilkom.v16i1.1952.68-75.
[21] T. A. Assegie, K. K. Napa, T. Thulasi, A. K. Kumar, M. J. T. V. Priya, and V. Dhamodaran, "Scalability and performance of decision tree for cardiovascular disease prediction," *IAES International Journal of Artificial Intelligence*, vol. 13, no. 3, pp. 2540–2545, Sep. 2024, doi: 10.11591/ijai.v13.i3.pp2540-2545.
[22] I. Roshanski, M. Kalech, and L. R. Ben, "Automatic Feature Engineering for Learning Compact Decision Trees," 2022. [Online]. Available: https://ssrn.com/abstract=4280154
[23] Yovita, "Algoritma Machine Learning yang Harus Kamu Pelajari di Tahun 2021," Dqlab.id. Accessed: Mar. 05, 2025. [Online]. Available: https://dqlab.id/algoritma-machine-learning-yang-perlu-dipelajari
[24] I. Gede, I. Sudipa, and M. Darmawiguna, *BUKU AJAR DATA MINING*. [Online]. Available: https://www.researchgate.net/publication/377415198
[25] I. Permana and F. Nur Salisah, "Pengaruh Normalisasi Data Terhadap Performa Hasil Klasifikasi Algoritma Backpropagation," IJIRSE: Indonesian Journal of Informatic Research and Software Engineering, vol 2, no 1, 2022, doi: https://doi.org/10.57152/ijirse.v2i1.311
[26] R. Ramadhan Laska and A. Mudya Yolanda, "A Comparative Study of Z-Score and Min-Max Normalization for Rainfall Classification in Pekanbaru," Journal of Data Science, 2024, doi: 10.61453/jods.v2024no04
[27] A. T. Akbar, R. Husaini, B. M. Akbar, and S. Saifullah, "A proposed method for handling an imbalance data in classification of blood type based on Myers-Briggs type indicator," *Jurnal Teknologi dan Sistem Komputer*, vol. 8, no. 4, pp. 276–283, Oct. 2020, doi: 10.14710/jtsiskom.2020.13625.
[28] N. Cahyana, S. Khomsah, and A. S. Aribowo, "Improving Imbalanced Dataset Classification Using Oversampling and Gradient Boosting," in *Proceeding - 2019 5th International Conference on Science in Information Technology: Embracing*

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2582−2593
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.7029

Industry 4.0: Towards Innovation in Cyber Physical System, ICSITech 2019, Institute of Electrical and Electronics Engineers Inc., Oct. 2019, pp. 217–222. doi: 10.1109/ICSITech46713.2019.8987499.

[29] M. Al Khaldy, "Resampling Imbalanced Class and the Effectiveness of Feature Selection Methods for Heart Failure Dataset," *International Robotics & Automation Journal*, vol. 4, no. 1, Feb. 2018, doi: 10.15406/iratj.2018.04.00090.

[30] N. L. W. S. R. Ginantra *et al.*, *FullBook Data Mining dan Penerapan Algoritma*. Yayasan Kita Menulis, 2021.

[31] A. Afifuddin and L. Hakim, "Deteksi Penyakit Diabetes Mellitus Menggunakan Algoritma Decision Tree Model Arsitektur C4.5," *Jurnal Krisnadana*, vol. 3, no. 1, Sep. 2023, Accessed: Nov. 11, 2024

[32] S. Sza *et al.*, "Penerapan Decision Tree dan Random Forest dalam Deteksi Tingkat Stres Manusia Berdasarkan Kondisi Tidur," *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK)*, vol. 11, no. 5, pp. 1043–1050, 2024, doi: 10.25126/jtiik.2024117993.

[33] R. Sheila, T. Rahmayani, and F. Budiman, "Analisa Optimasi Grid Search pada Algoritma Random Forest dan Decision Tree untuk Klasifikasi Stunting," *Technology and Science (BITS)*, vol. 6, no. 3, 2024, doi: 10.47065/bits.v6i3.6128.

[34] N. C. Sari and T. Linda Larasati, "Komparasi Algoritma Naïve Bayes dan Gradient Boosting untuk Prediksi Pasien Diabetes," *Jurnal Nasional Teknologi dan Sistem Informasi*, vol. 10, no. 2, pp. 118–125, Aug. 2024, doi: 10.25077/TEKNOSI.v10i2.2024.118-125.