JURNAL FASILKOM
Volume 15 No. 2 | Agustus 2025: 430-435

P-ISSN : 2089-3353
E-ISSN : 2808-9162

Perbandingan Algoritma Random Forest Dan Xgboost Untuk Klasifikasi Penyakit Jantung Berdasarkan Data Medis

Celvin Arafat[1], Muhammad Cavin Ramadhan[2], Fikri Abdul jafar[3], Muhammad Rayenra Azhi Pramudya[4], Edi Ismanto[5]
[1,2,3,4,5]Teknik Informatika, Fakultas Ilmu Komputer, Universitas Muhammadiyah Riau
[1]celvinarafat@gmail.com, [2]2304011003@student.umri.ac.id, [3]230401003@student.umri.ac.id, [4]rayenraazthi@gmail.com, [5]edi.ismanto@umri.ac.id

Abstract

Heart disease remains one of the leading causes of death worldwide, making early detection a critical step in reducing fatal risks. With the advancement of technology, machine learning has emerged as a promising approach to support medical diagnosis based on clinical data. This study aims to compare the performance of two widely used machine learning algorithms, namely Random Forest and XGBoost, in classifying heart disease. The dataset employed is the UCI Heart Disease dataset, which contains 303 patient records with 13 clinical attributes, including age, blood pressure, cholesterol levels, maximum heart rate, electrocardiogram (ECG) results, and signs of thalassemia. The research methodology began with data preprocessing, which involved handling missing values, outlier treatment, categorical feature encoding, and splitting the dataset into training and testing subsets with an 80:20 ratio using stratified sampling. Exploratory Data Analysis (EDA) was conducted to examine feature distribution and correlation patterns, providing insights for model development. Both models were trained using carefully selected hyperparameters and evaluated based on accuracy, precision, recall, F1-score, and ROC AUC metrics. The evaluation results indicate that Random Forest achieved higher performance with an accuracy of 90.16% and recall of 0.96, compared to XGBoost with an accuracy of 85.25% and recall of 0.86. These findings suggest that Random Forest demonstrates greater sensitivity and reliability in identifying patients with heart disease while minimizing false negatives. Consequently, ensemble methods based on decision trees, such as Random Forest, can serve as an effective and dependable solution for heart disease risk prediction systems utilizing medical data..

Keywords: Heart Disease, Classification, Random Forest, XGBoost, Machine Learning, Medical Data.

Abstrak

Penyakit jantung merupakan salah satu penyebab utama kematian di dunia, sehingga upaya deteksi dini sangat penting untuk menurunkan risiko fatal yang ditimbulkan. Seiring berkembangnya teknologi, machine learning menjadi salah satu pendekatan yang potensial dalam mendukung diagnosis penyakit berbasis data medis. Penelitian ini bertujuan untuk melakukan perbandingan kinerja dua algoritma machine learning, yaitu Random Forest dan XGBoost, dalam klasifikasi penyakit jantung. Dataset yang digunakan berasal dari UCI Heart Disease yang terdiri dari 303 data pasien dengan 13 atribut medis, di antaranya usia, tekanan darah, kolesterol, detak jantung maksimum, hasil elektrokardiogram (EKG), serta indikasi thalassemia. Metode penelitian diawali dengan tahapan pra-pemrosesan data, termasuk pembersihan data dari nilai kosong, penanganan outlier, transformasi fitur kategorikal menjadi numerik, serta pembagian data menjadi latih dan uji dengan perbandingan 80:20 menggunakan teknik stratified split. Proses eksplorasi data (EDA) dilakukan untuk memahami pola distribusi dan korelasi antar variabel, yang selanjutnya digunakan dalam pemodelan. Model dilatih dengan menggunakan parameter yang telah ditentukan, dan dievaluasi berdasarkan metrik akurasi, presisi, recall, F1-score, serta ROC AUC. Hasil penelitian menunjukkan bahwa algoritma Random Forest memperoleh akurasi sebesar 90,16% dengan nilai recall 0,96, sedangkan XGBoost menghasilkan akurasi 85,25% dengan recall 0,86. Hal ini mengindikasikan bahwa Random Forest lebih unggul dalam mendeteksi pasien yang benar-benar menderita penyakit jantung dengan tingkat kesalahan klasifikasi yang lebih rendah. Temuan ini membuktikan bahwa metode ensemble berbasis decision tree seperti Random Forest dapat menjadi solusi yang efektif dan andal dalam sistem prediksi risiko penyakit jantung berbasis data medis.

Kata kunci: Klasifikasi, Penyakit Jantung, Random Forest, Xgboost, Machine Learning, Data Medis.

1. Pendahuluan

Kesehatan merupakan kebutuhan dasar manusia yang harus dijaga dan dipantau secara berkala. Berdasarkan data World Health Organization (WHO) tahun 2024, penyakit jantung coroner menyebabkan lebih dari 17,9 juta kematian global. Di Indonesia, Kementerian Kesehatan Republik Indonesia melaporkan prevelensi penyakit jantung mencapai 1,5% dari populasi, dengan tren peningkatan yang signifikan seiring perubahan gaya hidup modern yang cenderung minim aktifitas fisik, pola konsumsi tinggi lemak, dan Tingkat stress yang tinggi. Dampak ekonomi yang ditimbulkan juga sangat besar, meliputi biaya perawatan yang tinggi dan penurunan produktivitas akibat morbiditas mau pun mortalitas. Oleh karena itu, deteksi dini penyakit jantung melalui teknologi berbasis machine learning

P-ISSN : 2089-3353
E-ISSN : 2808-9162

menjadi Solusi yang sangat relevan dalam mendukung diagnosis cepat dan akurat.

Penelitian sebelumnya, seperti yang di lakukan oleh [1], menunjukan bahwa model berbasis decision tree memiliki keunggulan dalam pengolahan data medis numerik, sedangkan metode boosting seperti XGBoost mampu memberikan peningkatan akurasi pada data yang bersifat kompleks. Dalam era digital dan big dataseperti sekarang, data medis bukan hanya sekedar catatan, tetapi juga sumber informasi berharga yang dapat diolah untuk mendeteksi dan mengklasifikasikan penyakit secara lebih akurat dan cepat. Salah satu pendekatan yang semakin banyak digunakan dalam pengolahan data medis adalah teknik klasifikasi menggunakan algoritma machine learning [1].

Klasifikasi adalah proses pemetaan data ke dalam kategori atau kelas tertentu berdasarkan pola dan karakteristik data tersebut. Dalam konteks medis, klasifikasi dapat digunakan untuk mendeteksi jenis penyakit berdasarkan gejala atau parameter medis pasien, seperti tekanan darah, kadar gula, usia, atau indeks massa tubuh. Salah satu dengan perkembangan teknologi informasi dan ketersediaan data medis yang semakin banyak, pendekatan machine learning menjadi solusi potensial untuk menganalisis data dan membuat prediksi klinis [2].

Dua algoritma yang banyak digunakan dalam klasifikasi adalah RandomForest dan XGBoost. Random Forest merupakan metode ensembel berbasis pohon keputusan, sedangkan XGBoost merupakan algoritma boosting yang terbukti efisien dan memiliki akurasi tinggi dalam banyak kompetisi data science [3].

2. Metode Penelitian

Penelitian ini dilakukan dengan menggunakan pendekatan kuantitatif melalui implementasi dan evaluasi dua algoritma machine learning, yaitu Random Forest dan XGBoost, dalam klasifikasi penyakit jantung berdasarkan data medis [4]. Data yang digunakan diperoleh dari dataset terbuka UCI Heart Disease yang memuat atribut-atribut medis seperti usia, jenis kelamin, tekanan darah, kolesterol, detak jantung maksimum, dan hasil elektrokardiogram [2].

Tahapan penelitian dimulai dengan pra-pemrosesan data, yang mencakup pembersihan data (*handling missing values and outlier*), transformasi numerik, serta encoding pada fitur kategorikal [5]. Data kemudian dibagi menjadi dua bagian, yaitu data latih dan data uji, dengan rasio 80:20. Selanjutnya dilakukan pelatihan model menggunakan algoritma Random Forest, yang menggabungkan beberapa pohon keputusan untuk meningkatkan akurasi melalui voting mayoritas, dan algoritma XGBoost, yang membangun model secara bertahap melalui teknik boosting dengan fokus pada perbaikan kesalahan model sebelumnya.

Evaluasi performa kedua model dilakukan menggunakan metrik evaluasi klasifikasi seperti akurasi, presisi, recall, F1-score, serta ROC AUC Score [6]. Selain itu, ditampilkan pula confusion matrix dan kurva ROC sebagai visualisasi kinerja model. Hasil evaluasi dibandingkan untuk mengetahui algoritma mana yang lebih efektif dalam mengklasifikasikan risiko penyakit jantung berdasarkan data medis [4]. Dataset yang digunakan merupakan data medis pasien dengan fitur- fitur klinis seperti tekanan darah, kadar kolesterol, usia, dan hasil pemeriksaan lain yang relevan untuk diagnosis penyakit jantung [7].

Proses pelatihan model dalam penelitian ini menggunakan pengaturan parameter (*hyperparameter*) yang dipilih secara hati-hati untuk mengoptimalkan kinerja masing-masing algoritma. Pada *Random Forest*, parameter yang digunakan antara lain n_estimators=100, max_depth=None, dan max_features='sqrt'. Sementara itu, pada *XGBoost* digunakan parameter n_estimators=100, learning_rate=0.1, max_depth=6, subsample=0.8, dan colsample_bytree=0.8.

Untuk memastikan distribusi kelas yang seimbang pada setiap pembagian data, digunakan teknik *stratified k-fold cross-validation* dengan nilai k = 5. Implementasi model dilakukan menggunakan bahasa pemrograman *Python*, dengan bantuan *library scikit-learn* untuk algoritma *Random Forest* dan *xgboost* untuk algoritma *XGBoost*.

2.1. *Preprocessing Data*

Tahap pra-pemrosesan data dilakukan untuk menjamin kualitas dan konsistensi dataset sebelum memasuki proses pelatihan model. Langkah pertama adalah pemeriksaan kelengkapan data dengan tujuan mengidentifikasi adanya nilai yang hilang (*missing values*) pada setiap atribut. Untuk atribut numerik, nilai yang hilang diestimasi menggunakan metode mean imputation, sedangkan untuk atribut kategorikal digunakan pengisian berdasarkan nilai modus [8].

Proses berikutnya adalah penanganan outlier yang dilakukan menggunakan metode *Interquartile Range* (IQR) untuk mendeteksi nilai ekstrem. *Outlier* yang tidak memiliki relevansi klinis dihapus dari dataset guna mencegah distorsi pada proses pembelajaran model. Sementara itu, *outlier* yang memiliki makna klinis dipertahankan untuk memastikan model tetap merepresentasikan karakteristik data medis secara akurat [9].

Tahap pengkodean variabel kategorikal dilakukan agar seluruh atribut dapat diproses oleh algoritma *machine learning*. Variabel dengan skala nominal dikonversi menggunakan metode *One-Hot Encoding*, sedangkan variabel dengan skala ordinal diubah melalui *Label Encoding* [10]. Selanjutnya, dilakukan normalisasi skala fitur numerik menggunakan metode *StandardScaler* yang tersedia pada pustaka *scikit-learn* [13]. Langkah ini bertujuan untuk menyeragamkan skala seluruh fitur, sehingga mencegah dominasi fitur dengan rentang nilai besar dan memastikan proses

P-ISSN : 2089-3353
E-ISSN : 2808-9162

pembelajaran, khususnya pada algoritma berbasis gradien seperti XGBoost, dapat berjalan secara optimal [12].

Dataset telah diperiksa dan tidak ditemukan nilai yang hilang (missing value). Selain itu, seluruh fitur sudah berupa data numerik, sehingga proses preprocessing difokuskan pada normalisasi dan scaling untuk memastikan data siap diproses oleh model. Preprocessing ini bertujuan untuk menjaga kualitas data sebelum dilakukan pelatihan model. Dataset dibagi menjadi data latih dan data uji dengan proporsi 80:20 menggunakan teknik stratified split agar distribusi kelas pada kedua subset tetap seimbang [8].

2.2. *Exploratory* Data *Analysis*

Exploratory Data Analysis (EDA) dilakukan untuk memahami struktur data dan karakteristik fitur-fitur medis yang digunakan dalam klasifikasi penyakit jantung [11]. Dataset yang digunakan terdiri dari 303 data pasien dengan 13 atribut input dan 1 atribut target yang merepresentasikan kondisi jantung pasien. Fitur-fitur dalam data mencakup informasi medis seperti usia, jenis kelamin, tekanan darah istirahat, kadar kolesterol, detak jantung maksimum, serta hasil pemeriksaan EKG dan thalassemia. EDA menunjukkan bahwa fitur-fitur seperti *cp, thalach, oldpeak, exang, ca*, dan *thal* memiliki pengaruh signifikan terhadap klasifikasi penyakit jantung [10]. Temuan ini menjadi dasar dalam proses pemodelan dan pemilihan fitur yang digunakan oleh algoritma *Random Forest* dan *XGBoost* untuk mengidentifikasi pasien yang berisiko.

2.3. Analisis EDA Lanjutan

Hasil analisis korelasi menunjukkan bahwa variabel cp (*chest pain type*) memiliki korelasi positif yang kuat terhadap variabel target, diikuti oleh *thalach* (*maximum heart rate achieved*) dan *oldpeak* (*ST depression induced by exercise*). Distribusi usia pasien menunjukkan mayoritas responden berada pada rentang usia 40–60 tahun, dengan proporsi pasien laki-laki yang lebih tinggi dibandingkan perempuan.

Analisis *boxplot* terhadap kadar kolesterol memperlihatkan adanya outlier dengan nilai di atas 350 mg/dL, yang mengindikasikan potensi risiko tinggi terhadap penyakit jantung. Heatmap korelasi antarvariabel digunakan untuk mendeteksi potensi *multicollinearity*, yang kemudian menjadi bahan pertimbangan dalam proses pemilihan fitur saat pelatihan model.

2.4. *Random Forest*

Random Forest merupakan algoritma *ensemble* yang membangun sejumlah pohon keputusan secara paralel, di mana setiap pohon dilatih menggunakan subset data yang dipilih secara acak melalui teknik *bootstrap*. Prediksi akhir dihasilkan melalui mekanisme voting mayoritas pada kasus klasifikasi atau nilai rata-rata pada regresi. Keunggulan utama algoritma ini meliputi ketahanan terhadap *overfitting* dan kemampuannya menangani data berdimensi tinggi dengan distribusi yang beragam.

Metode *Random Forest* adalah perkembangan dari metode Decision Tree. Pada algoritma ini setiap Decision Tree telah dilakukan proses training menggunakan sampel individu.[14] Ketika suatu data bertambah, maka tree akan bertambah atau berkembang. Proses prediksi Random Forest yaitu menggabungkan hasil dari setiap Decision Tree lalu dilakukan majority-voting untuk memperoleh hasil klasifikasi atau rata – rata regresi.

2.5. *XGBoost*

XGBoost (*Extreme Gradient Boosting*) adalah pengembangan dari algoritma *gradient boosting decision tree* yang membangun model secara bertahap, di mana setiap pohon baru difokuskan untuk memperbaiki kesalahan yang dihasilkan pohon sebelumnya. Algoritma ini dikenal memiliki optimisasi memori, kecepatan eksekusi yang tinggi, serta parameter regularisasi yang dapat mengontrol kompleksitas model untuk mencegah *overfitting*. Kedua algoritma ini banyak diaplikasikan dalam berbagai penelitian medis dan kompetisi data *science* karena kombinasi antara akurasi, efisiensi, dan kemampuan generalisasi yang dimilikinya.

XGBoost adalah versi dari algoritma pohon keputusan yang diperkuat secara bertingkat (gradient boosted decision trees). Algoritma ini membangun pohon keputusan secara berurutan. Semua variabel independen diberikan bobot tertentu, yang kemudian digunakan oleh pohon keputusan untuk membuat prediksi [15].

2.6. Evaluasi Model

a. Evaluasi Random Forest

Model Random Forest pada penelitian ini menunjukkan kinerja yang sangat baik dalam mendeteksi penyakit jantung. Hasil pengujian memberikan nilai akurasi sebesar 90,16%, yang mengindikasikan bahwa model mampu mengklasifikasikan pasien dengan dan tanpa penyakit jantung secara tepat pada lebih dari sembilan dari sepuluh kasus.

Berdasarkan analisis metrik evaluasi, nilai precision untuk kelas 0 (pasien sehat) mencapai 0,97, yang berarti 97% dari seluruh prediksi negatif yang dihasilkan model sesuai dengan kondisi aktual pasien yang tidak menderita penyakit jantung. Untuk kelas 1 (pasien sakit), nilai *precision* tercatat sebesar 0,84, yang menunjukkan bahwa 84% dari prediksi positif model benar-benar merupakan pasien yang menderita penyakit jantung.

Dari sisi *recall*, *Random Forest* memperoleh nilai sebesar 0,96 untuk kelas 1, menandakan bahwa 96% dari seluruh pasien yang benar-benar menderita penyakit jantung berhasil diidentifikasi dengan tepat

P-ISSN : 2089-3353
E-ISSN : 2808-9162

oleh model. Tingginya nilai recall ini sangat penting dalam konteks medis karena dapat meminimalkan risiko terjadinya *false negative*, yakni kondisi di mana pasien yang sebenarnya sakit tidak terdeteksi oleh sistem.

Nilai F1-score sebesar 0,9017 mengindikasikan keseimbangan yang baik antara *precision* dan *recall*, yang menjadikan Random Forest sebagai model yang andal untuk skrining awal penyakit jantung. Keunggulan ini dapat dikaitkan dengan mekanisme *bagging* pada Random Forest, di mana banyak pohon keputusan dibangun secara independen menggunakan subset data yang berbeda, kemudian hasilnya digabungkan melalui voting. Mekanisme ini efektif dalam mengurangi varians, mengatasi *overfitting*, serta menjaga stabilitas performa meskipun data memiliki variasi yang tinggi.

Dengan demikian, Random Forest dapat direkomendasikan sebagai model utama untuk aplikasi skrining awal di bidang kesehatan, terutama pada kasus yang memerlukan deteksi dini dengan sensitivitas tinggi.

Tabel 1. Hasil Evaluasi Random Forest

METRIK	NILAI
Accuracy	0.9016
Precison	0.9096
Recall	0.9016
Fi-Score	0.9017

b. Evaluasi XGBoost

Model XGBoost yang diimplementasikan pada penelitian ini menghasilkan tingkat akurasi sebesar 85,25%, yang menunjukkan kinerja klasifikasi yang cukup baik dalam membedakan antara pasien dengan dan tanpa penyakit jantung. Berdasarkan hasil pengujian, nilai precision untuk kelas 0 tercatat sebesar 0,88, yang mengindikasikan bahwa 88% dari seluruh prediksi negatif yang dihasilkan model sesuai dengan kondisi aktual pasien yang tidak menderita penyakit jantung. Sementara itu, nilai precision untuk kelas 1 sebesar 0,83 menunjukkan bahwa 83% dari seluruh prediksi positif yang dibuat model benar-benar merupakan pasien yang menderita penyakit jantung.

Pada metrik recall, model mencatatkan nilai sebesar 0,86 untuk kelas 1, yang berarti bahwa 86% dari seluruh pasien yang benar-benar menderita penyakit jantung berhasil diidentifikasi dengan tepat oleh model. Nilai ini mengindikasikan kemampuan XGBoost dalam meminimalkan terjadinya false negative, yang dalam konteks medis sangat penting untuk mencegah terlewatnya diagnosis terhadap pasien berisiko tinggi.

Meskipun kedua model yang dibandingkan menunjukkan performa yang kompetitif, hasil penelitian ini memperlihatkan bahwa Random Forest memiliki keunggulan pada nilai F1-score yang lebih seimbang serta recall yang lebih tinggi dibandingkan XGBoost. Perbedaan ini dapat dikaitkan dengan karakteristik XGBoost yang bersifat boosting, sehingga lebih sensitif terhadap distribusi data dan keberadaan outlier. Pada dataset dengan variasi yang cukup tinggi, metode bagging seperti Random Forest cenderung memberikan performa yang lebih stabil dan konsisten.

Dengan demikian, XGBoost tetap dapat dipertimbangkan sebagai alternatif yang efektif, terutama pada kondisi yang memerlukan waktu pelatihan lebih singkat dan efisiensi komputasi. Namun, untuk keperluan skrining awal penyakit jantung yang memprioritaskan sensitivitas, Random Forest menjadi pilihan yang lebih tepat karena kemampuannya dalam meminimalkan risiko kasus positif yang terlewat.

Tabel 2. Hasil Evaluasi XGBoost

METRIK	NILAI
Accuracy	0.8525
Precison	0.8532
Recall	0.8525
Fi-Score	0.8526

3. Hasil dan Pembahasan

Preprocessing data pada prediksi penyakit jantung dengan algoritma Random Forest dan XGBoost melibatkan beberapa tahapan penting, yaitu pembersihan data dari nilai kosong, pengubahan fitur kategorikal menjadi numerik menggunakan one-hot encoding, pemisahan antara fitur dan target, serta pembagian data menjadi 80% untuk pelatihan dan 20% untuk pengujian secara stratifikasi guna menjaga proporsi kelas yang seimbang. Meskipun tidak dilakukan normalisasi karena kedua algoritma berbasis pohon keputusan, hasil evaluasi dari confusion matrix menunjukkan bahwa preprocessing telah dilakukan dengan baik, terbukti dari akurasi Random Forest sebesar 90,16% dan XGBoost sebesar 85,25%, serta distribusi prediksi yang relatif seimbang antara kelas positif dan negatif, yang mencerminkan bahwa data yang digunakan bersih, terstruktur, dan siap digunakan untuk pelatihan model machine learning secara efektif.

3.1. Evaluasi

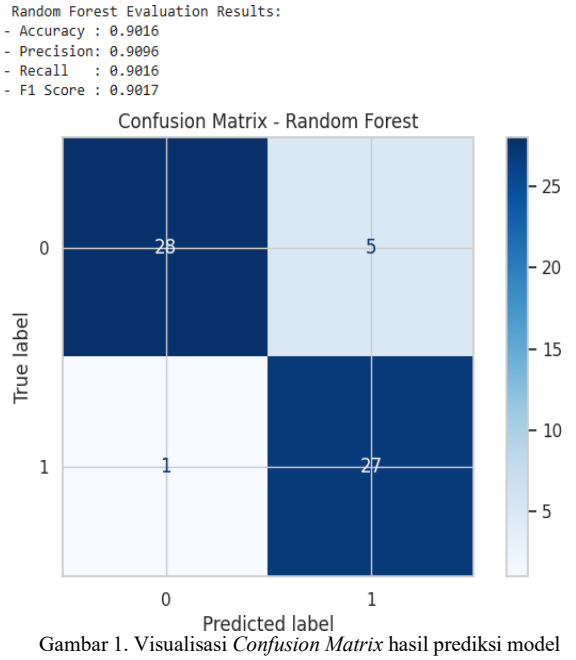
Model diuji pada data uji dan menghasilkan metrik evaluasi sebagai berikut:

a. Hasil Confusion Matrix

Evaluasi performa model klasifikasi sangat penting dalam menentukan efektivitas suatu algoritma dalam mengenali pola, terutama dalam konteks medis seperti deteksi penyakit jantung. Pada penelitian ini, dilakukan analisis perbandingan kinerja dua algoritma pembelajaran mesin, yaitu Random Forest dan XGBoost, menggunakan metrik evaluasi berbasis confusion matrix. Tujuan utama dari analisis ini adalah untuk menilai seberapa baik masing-masing model dalam membedakan antara pasien yang terindikasi

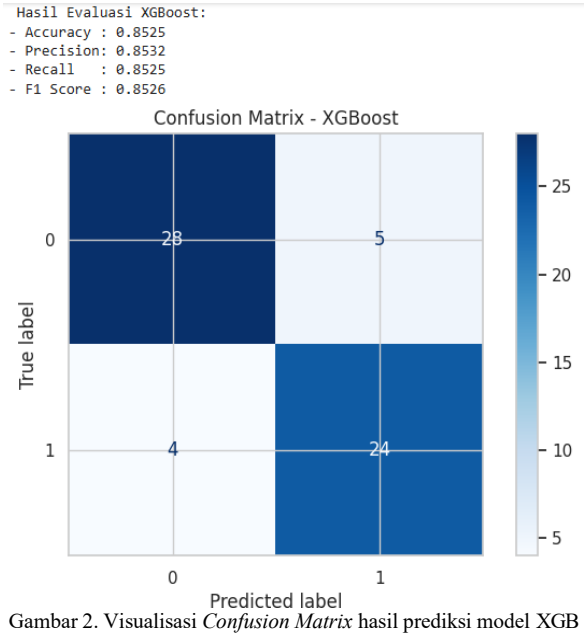
P-ISSN : 2089-3353
E-ISSN : 2808-9162

memiliki penyakit jantung (kelas positif) dan yang tidak (kelas negatif).



Gambar 1. Visualisasi *Confusion Matrix* hasil prediksi model Random Forest

Hasil evaluasi menunjukkan bahwa model Random Forest mampu mengklasifikasikan 28 data negatif dengan benar sebagai negatif (True Negative/TN) dan 27 data positif dengan benar sebagai positif (True Positive/TP). Namun, model ini juga menghasilkan 5 kesalahan klasifikasi data negatif sebagai positif (False Positive/FP) dan 1 kesalahan klasifikasi data positif sebagai negatif (False Negative/FN). Sebaliknya, model XGBoost menunjukkan hasil yang identik dalam hal True Negative dan False Positive, yakni masing-masing sebesar 28 dan 5. Namun, jumlah True Positive yang berhasil dikenali lebih rendah, yaitu sebanyak 24, serta jumlah False Negative lebih tinggi, sebanyak 4. Perbandingan ini mengindikasikan bahwa meskipun kedua model memiliki kemampuan yang serupa dalam mengidentifikasi data negatif, model Random Forest unggul secara signifikan dalam mengenali data positif, yang ditunjukkan oleh jumlah False Negative yang lebih rendah. Dalam konteks deteksi penyakit jantung, hal ini sangat penting, karena False Negative dapat berakibat fatal—pasien yang sebenarnya memiliki penyakit jantung dapat lolos dari diagnosis dan tidak memperoleh penanganan yang tepat waktu. Dengan demikian, dapat disimpulkan bahwa Random Forest memiliki performa yang lebih baik dibandingkan XGBoost dalam mengidentifikasi pasien yang benar-benar berisiko terhadap penyakit jantung, menjadikannya model yang lebih direkomendasikan untuk diterapkan dalam sistem klasifikasi medis yang menuntut tingkat sensitivitas tinggi.



Gambar 2. Visualisasi *Confusion Matrix* hasil prediksi model XGB

Hasil evaluasi model XGBoost menunjukkan kinerja yang cukup baik dengan akurasi sebesar 85,25%, precision 85,32%, recall 85,25%, dan F1-score 85,26%. Nilai metrik yang seimbang antara precision dan recall mengindikasikan bahwa model mampu mengklasifikasikan data secara konsisten tanpa bias yang signifikan terhadap salah satu kelas. Kinerja ini menandakan bahwa model tidak hanya mampu mengidentifikasi data positif dengan baik, tetapi juga cukup efektif dalam meminimalkan kesalahan klasifikasi.

Berdasarkan confusion matrix, model berhasil memprediksi 28 data kelas negatif dan 24 data kelas positif secara benar. Namun, terdapat 5 kasus false positive dan 4 kasus false negative yang menunjukkan adanya ruang untuk peningkatan, misalnya dengan tuning hyperparameter atau menambah variasi fitur. Secara keseluruhan, hasil ini menegaskan bahwa algoritma XGBoost memiliki performa yang stabil dan dapat diandalkan dalam mendukung proses klasifikasi pada penelitian ini.

6. Kesimpulan

Penelitian ini menyimpulkan bahwa algoritma Random Forest lebih unggul dibandingkan XGBoost dalam melakukan klasifikasi risiko penyakit jantung berdasarkan data medis numerik. Hasil ini menunjukkan potensi Random Forest untuk diterapkan dalam sistem pendukung keputusan klinis. Sebagai saran untuk penelitian selanjutnya, disarankan untuk menggunakan dataset yang lebih besar dan beragam. Dataset dengan jumlah sampel lebih banyak dan berasal dari populasi yang berbeda dapat membantu meningkatkan generalisasi model dan mengurangi overfitting. Mencoba fitur tambahan seperti data non-numerik dan gejala klinis lainnya seperti riwayat keluarga, kebiasaan hidup, dan keluhan subyektif dari

P-ISSN : 2089-3353
E-ISSN : 2808-9162

pasien dapat memperkaya representasi fitur dan meningkatkan akurasi model membandingkan lebih banyak model seperti LightGBM, CatBoost, atau neural network. Model-model ini memiliki pendekatan yang berbeda dan terkadang menunjukkan performa lebih baik tergantung pada struktur data.

Ucapan Terimakasih

Kami mengucapkan terima kasih kepada Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Muhammadiyah Riau atas dukungan fasilitas akademik dan sumber daya yang diberikan selama penelitian ini berlangsung. Apresiasi yang sebesar-besarnya disampaikan kepada Assoc. Prof. Edi Ismanto, S.T., M.Kom., Ph.D. selaku dosen pengampu mata kuliah yang telah memberikan bimbingan ilmiah, arahan, dan masukan yang berharga dalam penyusunan penelitian ini. Penulis juga menghargai kontribusi dan kerja sama dari rekan-rekan mahasiswa yang terlibat dalam diskusi serta memberikan saran konstruktif selama proses penelitian. Penelitian ini tidak menerima pendanaan khusus dari lembaga pendanaan publik, komersial, maupun organisasi nirlaba.

Daftar Rujukan

[1] Anshori, M., Regasari, A., & Putri, M. D. (2018). Optimasi Nilai K pada Algoritma K- Nearest Neighbor untuk Klasifikasi Data. Jurnal Teknologi dan Sistem Komputer,6(3),123-130. https://doi.org/10.14710/jtsiskom.6.3.123-130

[2] Sumarlin, A. (2016). Penerapan Algoritma K-Nearest Neighbor dalam Seleksi Penerima Beasiswa. Jurnal Sistem Informasi,12(2),45-52. http://jurnal.universitas.ac.id/index.php/jsi/article/view/1234

[3] Tang, J., Jing, L., Li, J., & Atkinson, P. M. (2016). A review of K-nearest neighbor algorithm and its applications. Neurocomputing,241,171-185. https://doi.org/10.1016/j.neucom.2017.01.042

[4] Tharwat, A., Gaber, T., Ibrahim, A., & Hassanien, A. E. (2018). Linear Discriminant Analysis: A Detailed Tutorial. AI Communications, 30(2), 169-190.https://doi.org/10.3233/AIC-170631

[5] Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13(1), 21-27. https://doi.org/10.1109/TIT.1967.1053964

[6] Jain, A. K., Duin, R. P. W., & Mao, J. (2000). Statistical pattern recognition: A review. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(1),437. https://doi.org/10.1109/34.824819

[7] Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI), 2, 1137-1145. https://www.ijcai.org/Proceedings/95-2/Papers/205.pdf

[8] Powers, D. M. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. Journal of Machine Learning Technologies, 2(1), 37-63. https://arxiv.org/abs/2010.16061

[9] Saito, T., & Rehmsmeier, M. (2015). The precision- recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. https://doi.org/10.1371/journal.pone.0118432

[10] Zhang, C. (2019). One-hot encoding for categorical data. Data Science Journal, 18(1), 1-9.https://datascience.codata.org/articles/10.5334/dsj-2019- 001/

[11] Firmansyah, I., Samudra, J. T., Pardede, D., & Situmorang, Z. (2022). Komparasi Random Forest Dan Logistic Regression Dalam Klasifikasi Penderita Covid-19 Berdasarkan Gejalanya. Journal of Science and Social Research, 5(3), 595. https://doi.org/10.54314/jssr.v5i3.994

[12] Inayah, K., & Ramli, K. (2024). Analisis Kinerja Intrusion Detection System Berbasis Algoritma Random Forest Menggunakan Dataset Unbalanced Honeynet BSSN. Jurnal Teknologi Informasi Dan Ilmu Komputer, 11(4), 867–876. https://doi.org/10.25126/jtiik.1148911

[13] Nur Azizah, A., Falach Asy'ari, M., Wisma Dwi Prastya, I., & Purwitasari, D. (2023). Easy Data Augmentation untuk Data yang Imbalance pada Konsultasi Kesehatan Daring. Jurnal Teknologi Informasi Dan Ilmu Komputer, 10(5), 1095–104. https://doi.org/10.25126/jtiik.2023107082

[14] Salmon, S., Azahari, A., & Ekawati, H. (2024). Perbandingan Kinerja Algoritma K-Nearest Neighbor dan Algoritma Random Forest Untuk Klasifikasi Data Mining Pada Penyakit Gagal Ginjal. Building of Informatics, Technology and Science (BITS), 6(3),1943–1953. https://doi.org/10.47065/bits.v6i3.6476

[15] Saputro, E., & Rosiyadi, D. (2022). Penerapan Metode Random Over-Under Sampling Pada Algoritma Klasifikasi Penentuan Penyakit Diabetes. Bianglala Informatika, 10(1), 42–47. https://doi.org/10.31294/bi.v10i1.11739