

ANALISIS DAN OPTIMASI PREDIKSI RISIKO PENYAKIT JANTUNG MENGGUNAKAN RANDOM FOREST DAN XGBOOST BERBASIS FEATURE SELECTION

Herwin Simanjuntak^{1*}, Juni Santo Sihotang², Purwa Hasan Putra³

¹²Institut Agama Kristen Negeri Tarutung, Tarutung

³Politeknik Negeri Medan, Medan

e-mail: ¹herwinsimanjuntak@iakntarutung.ac.id, ²junisanto@iakntarutung.ac.id,

³pputra@polmed.ac.id

Abstract: This study aims to analyze and optimize heart disease risk prediction using Random Forest and XGBoost algorithms combined with feature selection. The dataset used is the Cleveland Heart Disease Dataset, comprising 303 patient records, 13 predictor features, and a target variable converted into a binary classification. The research stages include preprocessing, data splitting using a stratified train-test split, feature selection, hyperparameter tuning, and evaluation based on accuracy, precision, recall, F1-score, and ROC-AUC. Feature selection yielded eight selected features—thal, cp, thalach, ca, oldpeak, exang, chol, and slope—resulting in a 38.46% reduction in the total number of features. Test results indicate that the Optimized Random Forest model combined with feature selection achieved the best performance, with an accuracy of 91.80%, precision of 89.66%, recall of 92.86%, F1-score of 91.23%, and ROC-AUC of 95.67%.

Keywords: Heart Disease, Random Forest, XGBoost, Feature Selection, Hyperparameter Tuning.

Abstrak: Penelitian ini bertujuan menganalisis dan mengoptimalkan prediksi risiko penyakit jantung menggunakan algoritma Random Forest dan XGBoost berbasis feature selection. Dataset yang digunakan adalah Cleveland Heart Disease Dataset yang terdiri atas 303 data pasien, 13 fitur prediktor, dan satu target yang dikonversi menjadi klasifikasi biner. Tahapan penelitian meliputi preprocessing, pembagian data menggunakan stratified train-test split, feature selection, hyperparameter tuning, serta evaluasi menggunakan accuracy, precision, recall, F1-score, dan ROC-AUC. Feature selection menghasilkan 8 fitur terpilih, yaitu thal, cp, thalach, ca, oldpeak, exang, chol, dan slope, sehingga jumlah fitur berkurang sebesar 38,46%. Hasil pengujian menunjukkan bahwa Optimized Random Forest + Feature Selection memberikan performa terbaik dengan accuracy 91,80%, precision 89,66%, recall 92,86%, F1-score 91,23%, dan ROC-AUC 95,67%.

Kata Kunci: Penyakit Jantung, Random Forest, XGBoost, Feature Selection, Hyperparameter Tuning.

PENDAHULUAN

Risiko penyakit jantung dipengaruhi oleh berbagai faktor, baik yang berkaitan dengan karakteristik individu maupun kondisi klinis pasien. Beberapa indikator yang umum digunakan antara lain usia, jenis kelamin, tekanan darah, kadar kolesterol, kadar

gula darah, denyut jantung, tipe nyeri dada, hasil elektrokardiografi, dan kondisi klinis lainnya. Banyaknya faktor yang saling berhubungan menyebabkan identifikasi risiko penyakit jantung menjadi permasalahan yang kompleks. Kondisi tersebut mendorong pemanfaatan teknologi komputasi untuk membantu menganalisis data kesehatan sehingga

pola yang sulit diidentifikasi secara konvensional dapat ditemukan secara lebih sistematis [1], [2].

Machine learning memiliki kemampuan untuk mempelajari pola dari data historis dan menghasilkan prediksi berdasarkan pola tersebut. menjelaskan bahwa penerapan berbagai algoritma machine learning dan strategi feature selection dapat digunakan dalam prediksi penyakit jantung. Penelitian tersebut mengevaluasi sejumlah algoritma, termasuk Decision Tree, Random Forest, Support Vector Machine, dan metode lainnya, serta menekankan pentingnya penggunaan metrik seperti precision dan recall dalam mengevaluasi kemampuan prediksi penyakit jantung [3]–[5].

Sejalan dengan perkembangan tersebut, penelitian ini memanfaatkan dataset penyakit jantung yang tersedia secara publik melalui Kaggle. Dataset penyakit jantung berbasis Cleveland/UCI yang tersedia di Kaggle memiliki sejumlah atribut klinis yang dapat digunakan dalam proses klasifikasi [6], [7].

Meskipun machine learning mempunyai potensi dalam membantu proses prediksi, performa suatu model sangat bergantung pada algoritma, kualitas data, fitur yang digunakan, dan konfigurasi model. Penggunaan seluruh fitur dalam dataset tidak selalu menghasilkan performa yang lebih baik karena terdapat kemungkinan beberapa fitur mempunyai informasi yang rendah, redundan, atau kurang relevan terhadap variabel target. Fitur-fitur tersebut dapat meningkatkan kompleksitas komputasi dan berpotensi memengaruhi kemampuan generalisasi model. Oleh sebab itu, salah satu pendekatan yang dapat digunakan untuk mengoptimalkan model adalah feature selection [8], [9].

Selain pemilihan fitur, pemilihan algoritma klasifikasi juga memiliki peranan penting. Penelitian ini menggunakan Random Forest dan Extreme Gradient Boosting (XGBoost). Random Forest merupakan metode ensemble learning berbasis bagging yang

membangun banyak decision tree menggunakan sampel data dan fitur yang bervariasi, kemudian menggabungkan hasil prediksi dari masing-masing pohon untuk menentukan hasil akhir. Pendekatan tersebut membuat Random Forest memiliki kemampuan menangani hubungan nonlinier antarvariabel serta relatif tahan terhadap overfitting dibandingkan sebuah decision tree tunggal [10].

Sementara itu, XGBoost merupakan algoritma ensemble learning berbasis gradient boosting yang membangun sejumlah model secara bertahap. Setiap model baru diarahkan untuk memperbaiki kesalahan prediksi model sebelumnya. XGBoost juga memiliki mekanisme regularisasi dan optimasi yang membuatnya banyak digunakan pada permasalahan klasifikasi berbasis data terstruktur [11], [12].

Meskipun berbagai penelitian sebelumnya telah menghasilkan performa yang baik, masih terdapat ruang penelitian dalam mengevaluasi seberapa besar pengaruh feature selection terhadap dua algoritma ensemble dengan mekanisme yang berbeda, yaitu Random Forest dan XGBoost. Dengan kata lain, pengurangan jumlah fitur perlu dianalisis tidak hanya berdasarkan jumlah fitur yang berhasil dieliminasi, tetapi juga berdasarkan perubahan accuracy, precision, recall, F1-score, dan ROC-AUC model [13], [14].

Evaluasi model dilakukan menggunakan beberapa metrik, yaitu accuracy, precision, recall, F1-score, confusion matrix, dan ROC-AUC. Penggunaan beberapa metrik diperlukan karena accuracy saja belum cukup untuk menggambarkan kemampuan model klasifikasi dalam konteks kesehatan. Recall penting untuk mengetahui kemampuan model dalam mendeteksi kasus positif penyakit jantung, sedangkan precision menunjukkan tingkat ketepatan prediksi positif. F1-score memberikan keseimbangan antara precision dan recall, sementara ROC-AUC digunakan untuk mengukur kemampuan diskriminatif model dalam membedakan kelas pasien

berisiko dan tidak berisiko [15], [16].

METODE

Penelitian ini menggunakan pendekatan kuantitatif eksperimental dengan menerapkan teknik machine learning untuk melakukan klasifikasi risiko penyakit jantung. Dataset yang digunakan merupakan dataset sekunder penyakit jantung yang diperoleh dari Kaggle. Penelitian bertujuan menganalisis dan mengoptimalkan performa algoritma Random Forest XGBoost melalui penerapan feature selection.



Gambar 1 Arsitektur Penelitian

Dataset

Penelitian dapat menggunakan Heart Disease Dataset dari Kaggle yang bersumber atau diadaptasi dari UCI Heart Disease/Cleveland dataset. Dataset ini berisi data klinis pasien yang digunakan untuk menentukan apakah seorang pasien terindikasi memiliki penyakit jantung berdasarkan sejumlah karakteristik kesehatan. Secara umum, dataset Cleveland yang sering digunakan memiliki sekitar 303 observasi dan 13 fitur prediktor + 1 target. Namun, jumlah baris dan format target pada versi Kaggle dapat berbeda karena ada versi yang sudah dibersihkan atau digabungkan. Karena itu, pada laporan penelitian sebaiknya jumlah data mengikuti file Kaggle yang benar-benar digunakan.

Data Preprocessing

Data preprocessing dilakukan untuk meningkatkan kualitas data sebelum digunakan dalam proses pemodelan. Tahap awal dilakukan dengan memeriksa struktur dataset yang meliputi jumlah data, jumlah atribut, tipe data, statistik

deskriptif, serta distribusi variabel target. Selanjutnya dilakukan pemeriksaan terhadap missing value dan data duplikat untuk memastikan kualitas dataset yang digunakan. Setelah proses pembersihan data selesai, dataset dipisahkan menjadi variabel independen (X) dan variabel target (y). Selanjutnya data dibagi menjadi 80% data training dan 20% data testing menggunakan stratified train-test split untuk mempertahankan proporsi kelas pada kedua subset data.

Random Forest

Pada penelitian ini, Random Forest digunakan untuk melakukan klasifikasi risiko penyakit jantung berdasarkan atribut kesehatan yang terdapat pada dataset. Data training yang telah melalui tahap preprocessing dan feature selection digunakan untuk membangun sejumlah Decision Tree. Pada setiap tree dilakukan bootstrap sampling sehingga masing-masing pohon dapat memperoleh komposisi data training yang berbeda. Selain itu, sejumlah fitur dipilih secara acak pada setiap proses pemisahan node sehingga dapat mengurangi korelasi antar-tree.

XGBoost

Extreme Gradient Boosting merupakan algoritma ensemble learning berbasis gradient boosting yang membangun sejumlah Decision Tree secara bertahap atau sekuensial. Berbeda dengan Random Forest yang membangun banyak pohon secara relatif independen menggunakan pendekatan bagging, XGBoost membangun pohon baru untuk memperbaiki kesalahan prediksi yang dihasilkan oleh pohon sebelumnya. Pada setiap iterasi, algoritma menghitung kesalahan (residual) dari prediksi sebelumnya, kemudian membentuk pohon berikutnya untuk meminimalkan kesalahan tersebut. Hasil dari seluruh pohon selanjutnya dikombinasikan secara berbobot untuk menghasilkan prediksi akhir.

Evaluasi Performance

Evaluasi performa merupakan tahap untuk mengukur dan membandingkan kemampuan model Random Forest dan XGBoost dalam memprediksi risiko penyakit jantung berdasarkan data testing yang tidak digunakan pada proses pelatihan model. Evaluasi dilakukan terhadap model baseline dan model hasil optimasi sehingga dapat diketahui pengaruh feature selection dan hyperparameter tuning terhadap performa masing-masing algoritma. Metrik evaluasi yang digunakan meliputi Confusion Matrix, Accuracy, Precision, Recall, F1-Score, dan ROC-AUC [13], [17].

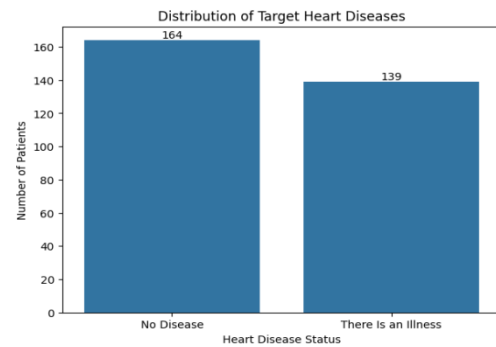
HASIL DAN PEMBAHASAN

Penelitian ini menggunakan Heart Disease Dataset yang diperoleh dari Kaggle. Dataset terdiri atas sejumlah atribut klinis yang digunakan sebagai variabel prediktor, antara lain age, sex, chest pain type (cp), resting blood pressure (trestbps), cholesterol (chol), fasting blood sugar (fbs), resting electrocardiographic results (restecg), maximum heart rate (thalach), exercise-induced angina (exang), oldpeak, slope, ca, dan thal. Variabel target digunakan untuk menunjukkan kelas yang akan diprediksi oleh model.

Dataset

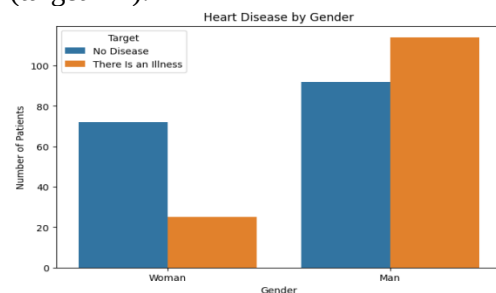
Dataset yang digunakan dalam penelitian ini adalah Cleveland Heart Disease Dataset yang terdiri atas 303 observasi dengan 13 atribut prediktor dan satu atribut target.

Atribut prediktor meliputi karakteristik demografis dan klinis pasien, yaitu age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, dan thal. Variabel target asli memiliki nilai 0 sampai 4 yang selanjutnya dikonversi menjadi klasifikasi biner, yaitu kelas 0 untuk pasien yang tidak terindikasi penyakit jantung dan kelas 1 untuk pasien yang terindikasi penyakit jantung.



Gambar 1 Distribusi Dataset

Grafik pertama menunjukkan distribusi kelas target setelah target asli Cleveland dikonversi menjadi klasifikasi biner. Dari 303 data pasien, terdapat 164 pasien (54,13%) yang masuk kategori tidak terindikasi penyakit jantung (target = 0) dan 139 pasien (45,87%) yang masuk kategori terindikasi penyakit jantung (target = 1).



Gambar 2 Distribusi Dataset by Gender

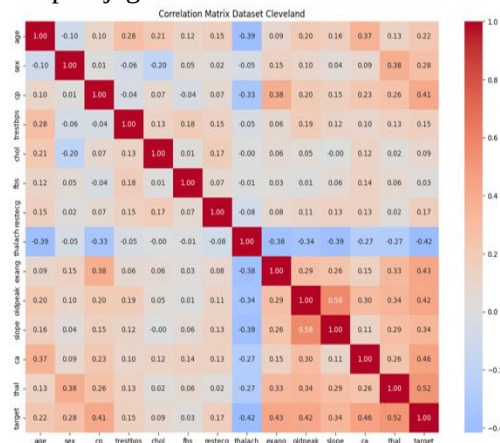
Grafik kedua memperlihatkan hubungan antara usia (age) dengan kelas penyakit jantung. Secara visual, kelompok yang tidak terindikasi penyakit jantung terlihat lebih banyak pada rentang usia yang relatif lebih muda, terutama sekitar usia 40–55 tahun. Sementara itu, distribusi kelompok yang terindikasi penyakit jantung terlihat lebih terkonsentrasi pada rentang usia yang lebih tinggi, khususnya sekitar 55–65 tahun.

Data Preprocessing

Preprocessing dilakukan untuk memastikan bahwa dataset memiliki kualitas yang sesuai untuk proses machine learning. Tahapan yang dilakukan meliputi pemeriksaan missing value, data duplikat, tipe data, nilai ekstrem, serta

encoding terhadap atribut kategorikal apabila diperlukan.

Setelah proses pembersihan selesai, dataset dipisahkan menjadi variabel independen (X) dan variabel target (y). Data kemudian dibagi menggunakan stratified train-test split dengan proporsi 80% sebagai data training dan 20% sebagai data testing. Stratifikasi digunakan agar proporsi masing-masing kelas target pada data training dan testing tetap terjaga.

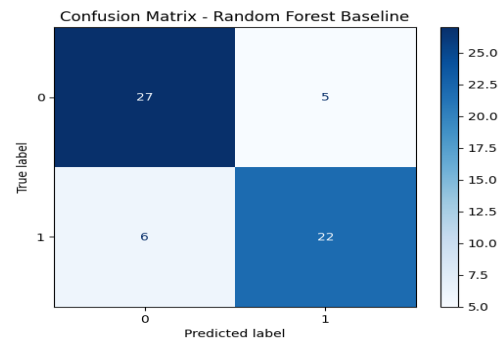


Gambar 3 Correlation Matrix Dataset

Berdasarkan hasil correlation matrix, terdapat beberapa atribut yang menunjukkan hubungan relatif lebih kuat dengan variabel target. Atribut ca memiliki korelasi positif sekitar 0,46, diikuti exang sebesar 0,43, oldpeak sebesar 0,42, dan thal sebesar 0,39. Sementara itu, thalach dan cp menunjukkan korelasi negatif masing-masing sekitar -0,42 dan -0,41.

Random Forest

Berdasarkan hasil pengujian Random Forest Baseline menggunakan data testing, diperoleh 27 True Negative, 22 True Positive, 5 False Positive, dan 6 False Negative. Dari total 60 data testing, sebanyak 49 data berhasil diklasifikasikan dengan benar dan 11 data mengalami kesalahan klasifikasi. Berdasarkan confusion matrix tersebut, Random Forest Baseline menghasilkan accuracy sebesar 81,67%, precision sebesar 81,48%, recall sebesar 78,57%, dan F1-Score sebesar 80,00%.

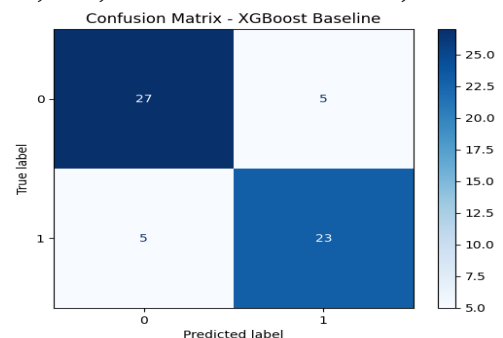


Gambar 4 Confusion Matrix Random Forest

Hasil tersebut menunjukkan bahwa model telah memiliki kemampuan klasifikasi yang cukup baik, tetapi masih terdapat enam kasus positif yang salah diprediksi sebagai kelas negatif. Oleh karena itu, tahap feature selection dan hyperparameter tuning selanjutnya diarahkan untuk meningkatkan kemampuan generalisasi model, khususnya meningkatkan recall dan mengurangi False Negative. Performa model hasil optimasi kemudian dibandingkan dengan baseline untuk menentukan efektivitas proses optimasi Random Forest.

XGBoost

Berdasarkan hasil pengujian XGBoost baseline terhadap 60 data testing, diperoleh 27 True Negative, 23 True Positive, 5 False Positive, dan 5 False Negative. Dengan demikian, sebanyak 50 data berhasil diklasifikasikan dengan benar dan 10 data mengalami kesalahan klasifikasi. Hasil tersebut menghasilkan accuracy sebesar 83,33%, precision sebesar 82,14%, recall sebesar 82,14%, dan F1-Score sebesar 82,14%.

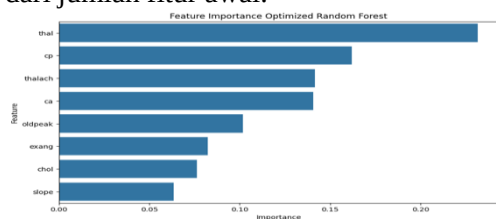


Gambar 5 Confusion Matrix XGBoost

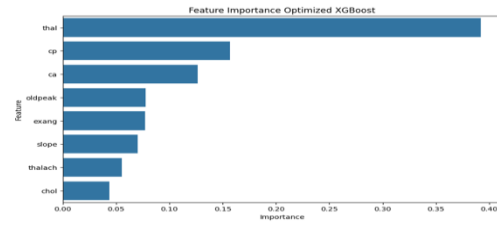
Jika dibandingkan dengan Random Forest baseline yang menghasilkan accuracy 81,67% dan recall 78,57%, XGBoost baseline menunjukkan performa yang lebih baik pada pengujian ini. Selain meningkatkan jumlah True Positive dari 22 menjadi 23, XGBoost juga menurunkan False Negative dari 6 menjadi 5. Meskipun demikian, hasil baseline belum digunakan sebagai kesimpulan akhir karena kedua algoritma selanjutnya perlu melalui proses feature selection dan hyperparameter tuning. Hasil setelah optimasi akan dibandingkan untuk menentukan model yang memberikan performa terbaik dalam klasifikasi penyakit jantung.

Feature Selection

Berdasarkan hasil feature selection, dari 13 fitur awal diperoleh 8 fitur terpilih, yaitu thal, cp, thalach, ca, oldpeak, exang, chol, dan slope. Dengan demikian, proses seleksi fitur berhasil mengurangi lima fitur atau sekitar 38,46% dari jumlah fitur awal.



Gambar 6 Feature Selection Random Forest



Gambar 7 Feature Selection XGBoost

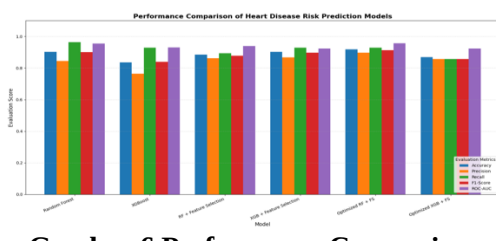
Hasil feature importance menunjukkan bahwa thal menjadi fitur dengan kontribusi terbesar pada kedua algoritma. Pada Random Forest, thal memiliki nilai importance sekitar 0,23 dan diikuti oleh cp, thalach, serta ca. Sementara pada XGBoost, thal menunjukkan dominasi yang lebih besar dengan importance sekitar 0,39, diikuti cp dan ca. Perbedaan urutan kepentingan fitur antara kedua algoritma menunjukkan bahwa Random Forest dan XGBoost memiliki mekanisme yang berbeda dalam memanfaatkan informasi dari masing-masing atribut.

Evaluasi Performance

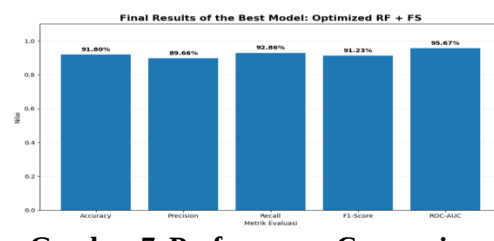
Berdasarkan hasil Evaluasi Performance, terdapat enam skenario model yang dibandingkan, yaitu model awal, model dengan feature selection (FS), dan model hasil optimasi. Secara keseluruhan, Optimized Random Forest + Feature Selection memberikan performa paling seimbang dan menjadi kandidat model terbaik.

Tabel 1 Evaluasi Performance

	Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
0	Random Forest	0.9016	0.8438	0.9643	0.9000	0.9545
1	XGBoost	0.8361	0.7647	0.9286	0.8387	0.9297
2	RF + Feature Selection	0.8852	0.8621	0.8929	0.8772	0.9389
3	XGB + Feature Selection	0.9016	0.8667	0.9286	0.8966	0.9232
4	Optimized RF + FS	0.9180	0.8966	0.9286	0.9123	0.9567
5	Optimized XGB + FS	0.8689	0.8571	0.8571	0.8571	0.9232



Gambar 6 Performance Comparison



Gambar 7 Performance Comparison

Berdasarkan hasil evaluasi terhadap enam skenario pemodelan, Optimized Random Forest berbasis Feature Selection menghasilkan performa keseluruhan terbaik dengan accuracy sebesar 91,80%, precision 89,66%, recall 92,86%, F1-score 91,23%, dan ROC-AUC 95,67%. Nilai tersebut menunjukkan bahwa kombinasi feature selection dan hyperparameter tuning mampu menghasilkan model Random Forest yang memiliki keseimbangan performa yang baik.

Hasil ini menunjukkan bahwa kombinasi feature selection dan hyperparameter tuning mampu meningkatkan performa keseluruhan Random Forest dibandingkan model baseline. Accuracy meningkat dari 90,16% menjadi 91,80%, precision dari 84,38% menjadi 89,66%, F1-score dari 90,00% menjadi 91,23%, dan ROC-AUC dari 95,45% menjadi 95,67%. Meskipun demikian, recall mengalami penurunan dari 96,43% pada Random Forest baseline menjadi 92,86% pada model hasil optimasi.

SIMPULAN

Berdasarkan hasil penelitian, penerapan Random Forest dan XGBoost berbasis feature selection mampu digunakan untuk melakukan klasifikasi penyakit jantung pada Cleveland Heart Disease Dataset. Proses feature selection berhasil mengurangi jumlah fitur dari 13 menjadi 8 fitur, yaitu thal, cp, thalach, ca, oldpeak, exang, chol, dan slope, atau mengurangi sekitar 38,46% fitur awal. Hasil feature importance juga menunjukkan bahwa thal menjadi fitur dengan kontribusi terbesar pada kedua algoritma.

Berdasarkan evaluasi terhadap enam skenario model, Optimized Random Forest + Feature Selection menghasilkan performa keseluruhan terbaik dengan accuracy sebesar 91,80%, precision 89,66%, recall 92,86%, F1-score 91,23%, dan ROC-AUC 95,67%. Dibandingkan

Random Forest baseline, model hasil optimasi meningkatkan accuracy dari 90,16% menjadi 91,80%, precision dari 84,38% menjadi 89,66%, F1-score dari 90,00% menjadi 91,23%, serta ROC-AUC dari 95,45% menjadi 95,67%. Namun, recall mengalami penurunan dari 96,43% menjadi 92,86%, sehingga Random Forest baseline tetap memiliki sensitivitas tertinggi dalam mendeteksi kelas positif.

DAFTAR PUSTAKA

- C. Z. Syahputri, S. Informasi, I. Komputer, and U. D. Jambi, "Optimasi XGBoost Dengan SHAP Untuk Sistem Skrining Penyakit Jantung," no. November, 2025.
- J. Elektronika, I. Terapan, S. K. Parinduri, and A. Pribadi, "Prediksi Risiko Penyakit Jantung Menggunakan Xgboost dengan Optimasi Hyperparameter Berbasis GridsearchCV Penyakit Jantung dan Machine Learning," vol. 4, no. 1, 2026.
- Y. Optarina *et al.*, "Jurnal Software Engineering and Information System (SEIS) OPTIMASI MODEL XGBOOST UNTUK PREDIKSI PENYAKIT JANTUNG," vol. 6, no. 1, 2026.
- D. A. N. K. N. Dalam, "ANALISIS PERBANDINGAN PERFORMA ALGORITMA RANDOM," vol. 5, no. 3, pp. 302–311, 2026.
- A. E. Pranajaya and E. R. Susanto, "Optimasi Random Forest untuk Prediksi Penyakit Jantung Menggunakan SMOTEENN dan Grid Search Sistem Informasi , Fakultas Teknik dan Ilmu Komputer , Universitas Teknokrat Indonesia , Indonesia Random Forest Optimization for Heart Disease Prediction Using SMOTEENN and Grid Search," vol. 5, no. 7, pp. 1965–1979, 2025.
- I. Husni, A. Amin, S. Wibisono, and E. Lestariningsih, "Optimizing Heart

- Disease Prediction with Random Forest and Ensemble Methods,” vol. 11, no. 1, pp. 80–90, 2025.
- A. Pratama, S. Assegaff, J. Jasmir, and N. Nurhadi, “Optimizing Heart Disease Classification Using C4.5, Random Forest, and XGBoost with ANOVA, Chi-Square, and AdaBoost,” vol. 7, no. 2, pp. 1072–1090, 2026.
- M. N. Fadhlurrahman, “Optimization of Heart Failure Risk Prediction Using Random Forest Classifier Algorithm,” pp. 57–64, 2025.
- N. L. Wicaksono, “Optimasi Hyperparameter Ensemble Learning untuk Prediksi Penyakit Jantung berdasarkan Data Pasien Hyperparameter Optimization of Ensemble Learning for Heart Disease Prediction using Patient Data,” vol. 15, pp. 896–906, 2026.
- M. T. P. Mukti, H. Hairani, D. Rahardjo, and M. Rizki, “A Comparison of Logistic Regression, Random Forest, and XGBoost Based on Feature Importance in Heart Failure Prediction,” vol. 5, no. 2, pp. 61–72.
- M. Dzaky and A. P. Kuncoro, “Optimizing XGBoost for Heart Disease Risk Classification Using Optuna and Random Search on the Behavioral Risk Factor Surveillance System (BRFSS) 2023 Dataset,” vol. 10, no. 1, 2026.
- S. Ahmed, N. Elazab, M. M. El-gayar, M. Elmogy, and Y. M. Fouda, “Multi-Scale Vision Transformer with Optimized Feature Fusion for Mammographic Breast Cancer Classification,” pp. 1–30, 2025.
- M. S. Lydia *et al.*, “Mid-Upper Arm Circumference Measurement Using Digital Images: A Top-Down Approach with Panoptic Segmentation Using Mask R-CNN,” vol. 16, no. 8, pp. 392–404, 2025.
- P. H. Hutagalung, S. Informasi, and U. Nasional, “Heart Disease Classification Using Optimised XGBoost and Random Forest with SHAP Explanations,” vol. 10, no. 1, pp. 330–342, 2026.
- A. R. Lubis, P. H. Putra, and F. S. Lubis, “Optimization of principal component analysis and k-nearest neighbors in cultivation area classification red onion,” vol. 23, no. 6, pp. 1579–1589, 2025.
- P. H. Putra, D. Selvida, and M. S. Novelan, “Application of Apriori Algorithm in Data Mining to Find Consumer Purchasing Patterns in Supermarkets,” vol. 4, no. 1, pp. 221–226, 2025.
- D. Selvida, P. Hasan, and M. Mughnyanti, “Penerapan Algoritma Apriori dan K-Nearest Neighbor Untuk Analisis Pola Prediksi Pembelian,” vol. 4, no. 1, pp. 54–63, 2026.