Implementasi Klasifikasi Xgboost Untuk Prediksi Dini Risiko Penyakit Jantung Pada Subkelompok Dewasa Muda

Fatia Amalia Maresti [1)], Ayu Andani [2)*], Kent Bryan Yasin [3)], Woro Isti Rahayu[4)]
[1,2,3,4)] Program Studi Sains Data, Universitas Logistik dan Bisnis Internasional (ULBI)

Received: 28 July 2025 Accepted: 27 January 2026 Published: 1 June 2026

ayuandani0302@gmail.com

Kata Kunci: Penyakit Jantung, XGBoost, Prediksi Dini, Dewasa Muda, Klasifikasi.

DSI: Jurnal Data Science Indonesia is licensed under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0).

Abstrak : Penyakit jantung koroner merupakan penyebab utama kematian secara global dan menjadi perhatian di Indonesia, terutama pada kelompok dewasa muda yang sering terdeteksi terlambat. Penelitian ini bertujuan menerapkan algoritma *Extreme Gradient Boosting* (XGBoost) untuk memprediksi risiko serangan jantung secara dini pada subkelompok usia 25–30 tahun. Data yang digunakan merupakan data sekunder dari dataset publik DeepDataLake. Tahap pengolahan data meliputi penyaringan usia, pembersihan data, penyeimbangan kelas (*undersampling*), encoding variabel, konversi fitur biner ke format numerik, serta standardisasi fitur menggunakan StandardScaler. Dataset dibagi menjadi 80% data latih dan 20% data uji menggunakan stratified split. Optimasi hiperparameter dilakukan menggunakan RandomizedSearchCV dengan validasi silang untuk memperoleh konfigurasi terbaik. Evaluasi pada data uji menunjukkan akurasi sebesar 76,15% dan *Area Under the Curve* (AUC) sebesar 0,8443, yang mengindikasikan kemampuan diskriminasi model yang baik dalam membedakan individu berisiko dan tidak berisiko. Penelitian ini bersifat eksploratif berbasis data sekunder sehingga hasilnya tidak dimaksudkan untuk merepresentasikan populasi Indonesia secara nasional.

PENDAHULUAN

Penyakit jantung koroner(PJK) menjadi fokus utama penelitian karena merupakan penyebab kematian tertinggi secara global maupun di Indonesia. Data World Health Organization menegaskan bahwa penyakit jantung adalah penyebab kematian nomor satu di dunia[1]. Di Indonesia, Riskesdas 2013 melaporkan prevalensi penyakit jantung koroner pada penduduk usia > 15 tahun sebesar 1,6 % berdasarkan gejala subjektif[1]. Sementara itu, analisis data Riskesdas 2013–2018 pada 722.329 responden usia ≥ 15 tahun dari 33 provinsi menunjukkan prevalensi penyakit jantung koroner sebesar 1,5 % berdasarkan diagnosa dokter maupun gejala [2]. Tren kenaikan kematian jantung memicu urgensi deteksi dini, karena intervensi pada tahap awal dapat menurunkan angka kejadian dan mortalitas.

Meskipun prevalensi penyakit jantung koroner pada dewasa muda secara keseluruhan lebih rendah ketimbang lansia, penyakit jantung koroner tetap menjadi penyebab utama kematian mendadak pada pasien usia di bawah 45 tahun[3]. Hampir semua pasien muda dengan penyakit jantung koroner memiliki minimal satu faktor risiko kardiovaskular konvensional[3]. Variabel yang dianalisis dalam penelitian ini yaitu hipertensi, diabetes melitus, kolesterol, dan riwayat penyakit jantung telah dikenal kuat berkontribusi pada perkembangan aterosklerosis koroner.

Penelitian memfokuskan pada rentang usia 25–30 tahun sebagai sub kelompok dewasa muda.

Meskipun Riskesdas dan literatur lain belum merinci prevalensi penyakit jantung koroner khusus di rentang usia ini, kelompok 25–30 tahun masuk kategori "di bawah 45 tahun" yang rentan terhadap kejadian penyakit jantung koroner[3]. Selain itu, faktor risiko metabolik seperti hipertensi, diabetes, dan dislipidemia umumnya mulai muncul pada akhir masa dewasa muda, seiring perubahan pola hidup modern. Prediksi dini faktor risiko penyakit jantung koroner pada dewasa muda perlu diperkuat, karena banyak kasus hipertensi atau diabetes pada kelompok usia ini terdeteksi terlambat ketika sudah muncul komplikasi. Skrining rutin meliputi pengecekan tekanan darah, kadar gula, dan profil lipid dapat membantu mengenali individu berisiko tinggi sebelum muncul gejala klinis. Penelitian ini merupakan studi berbasis data sekunder yang menggunakan dataset publik dari DeepDataLake. Dataset tidak menyediakan informasi asal geografis atau fasilitas layanan kesehatan, sehingga hasil penelitian tidak dimaksudkan sebagai representasi populasi Indonesia. Konteks Indonesia pada bagian pendahuluan digunakan untuk menjelaskan urgensi masalah, sedangkan analisis model difokuskan pada subkelompok usia 25–30 tahun.

XGBoost bekerja dengan membangun serangkaian pohon keputusan*(decision trees)* secara berurutan, di mana setiap pohon baru berusaha untuk memperbaiki kesalahan prediksi yang dibuat oleh kombinasi pohon–pohon sebelumnya[4]. Pemilihan algoritma XGBoost untuk klasifikasi risiko dilandasi oleh keunggulannya dalam akurasi dan efisiensi komputasi. Sebuah studi menunjukkan XGBoost mencapai akurasi 93 % dalam memprediksi penyakit jantung koroner, mengungguli Decision Tree (90 %)[5], dan studi lain menyebutkan akurasi model penyakit jantung koroner dari 75,41 % menjadi 89,47 % setelah menerapkan XGBoost dengan penanganan outlier[6]. Keunggulan ini menegaskan XGBoost sebagai pilihan tepat untuk implementasi prediksi dini risiko penyakit jantung pada subkelompok dewasa muda usia 25-30 tahun.

TINJAUAN LITERATUR

Penyakit jantung termasuk dalam kelompok *Cardiovascular Diseases* (CVDs) diperkirakan 19,8 juta orang meninggal karena penyakit kardiovaskular (CVDs) pada tahun 2022[7]. Di Indonesia, prevalensi penyakit jantung berdasarkan diagnosis dokter diperkirakan mencapai 1,5%, atau sekitar 15 dari setiap 1.000 penduduk [8], yang setara dengan ±2,78 juta orang hidup dengan kondisi tersebut. Literatur Indonesia mengidentifikasi berbagai faktor risiko penyakit jantung koroner (PJK) yang dapat dikelompokkan ke dalam faktor yang tidak dapat diubah dan faktor yang dapat diubah. Hipertensi, diabetes melitus, dan dislipidemia merupakan faktor dominan yang secara statistik bermakna terhadap kejadian penyakit jantung koroner di Indonesia[2]. Dalam tinjauan sistematis studi lain, juga menegaskan bahwa hipertensi, diabetes, merokok, obesitas, dan rendahnya aktivitas fisik secara konsisten disebut sebagai determinan utama penyakit jantung koroner dalam berbagai jurnal Indonesia[9]. Pada kelompok dewasa muda, peningkatan prevalensi hipertensi dan diabetes menjadi perhatian khusus karena gejala awal sering tidak terdeteksi tanpa adanya skrining rutin [10]. Secara ringkas, variabel yang dianalisis dalam penelitian ini yaitu umur, hipertensi, diabetes melitus, kolesterol, obesitas, dan riwayat penyakit jantung seluruhnya memiliki dasar yang kuat dalam literatur Indonesia.

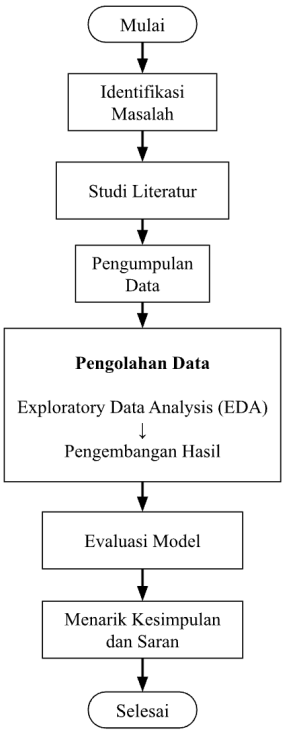
XGBoost bekerja dengan membangun serangkaian pohon keputusan (*decision trees*) secara berurutan, di mana setiap pohon baru berusaha untuk memperbaiki kesalahan prediksi yang dibuat oleh kombinasi pohon–pohon sebelumnya[4]. Pendekatan ini dikenal sebagai ensemble learning melalui teknik boosting, yang secara bertahap menggabungkan weak learners (pohon keputusan dangkal) menjadi strong learner yang lebih kuat dan akurat[5][6]. Salah satu kunci dalam XGBoost adalah penggunaan ekspansi *Taylor orde* kedua pada fungsi kerugian (*loss function*), yang memungkinkan pemanfaatan informasi turunan pertama dan kedua untuk optimasi yang lebih presisi[4]. Selain itu, XGBoost memperkenalkan istilah regularisasi dalam fungsi objektifnya untuk mengontrol kompleksitas model dan secara efektif

mencegah overfitting[5][6]. Sejumlah penelitian menunjukkan bahwa XGBoost memiliki performa yang baik pada tugas klasifikasi, bahkan dapat mengungguli beberapa algoritma lain seperti Naïve Bayes dan Random Forest pada konteks tertentu. Sebagai contoh, pada studi klasifikasi kualitas air minum menggunakan dataset berukuran 2.400 data (pembagian 80% data latih dan 20% data uji) dengan komposisi kelas seimbang (1.200:1.200), evaluasi confusion matrix menunjukkan akurasi 82,29%, *precision* 78,62%, *recall* 85,90%, dan *F1-Score* 82,09%. Temuan ini menunjukkan bahwa XGBoost mampu menghasilkan performa klasifikasi yang kompetitif pada dataset terstruktur, terutama ketika dilakukan pembersihan data, penyeimbangan kelas, dan optimasi parameter[11] .Kemampuan mengatasi data kompleks, XGBoost juga kuat dalam menangani data bermasalah. XGBoost mampu menangani data yang hilang, menghindari *overfitting*, dan menyesuaikan dengan kumpulan data dinamis[12]. Artinya, XGBoost dapat bekerja dengan baik meski ada data yang hilang atau tidak seimbang dan mencegah model menjadi terlalu cocok (overfit) terhadap data latihan. Keunggulan ini penting dalam prediksi penyakit, di mana data klinis seringkali tidak sempurna.

Evaluasi kinerja model pada umumnya menggunakan metrik yang dihitung dari confusion matrix; Akurasi (*Accuracy*) adalah "metrik untuk mengukur sejauh mana model dapat melakukan klasifikasi dengan benar"[5]. Presisi (*Precision*) adalah "persentase label positif yang benar dari seluruh label yang diprediksi sebagai positif".[13] *Recall* (Sensitivitas) adalah "efektivitas pengklasifikasi dalam mengidentifikasi label positif yang benar" [13]. *F1-Score* adalah "metrik yang digunakan untuk menghitung rata – rata harmonis antara presisi dan recall"[5].

METODE PENELITIAN

Metode penelitian digunakan sebagai pedoman penyusunan langkah–langkah penelitian secara sistematis untuk mencapai tujuan penelitian. Penelitian ini menggunakan metode kuantitatif karena berfokus pada pengolahan data numerik dan analisis statistik. Alur penelitian disajikan pada Gambar 1.



Gambar 1. Alur penelitian

Proses penelitian ini terdiri atas beberapa tahapan utama, yaitu:

a. Pengumpulan Data

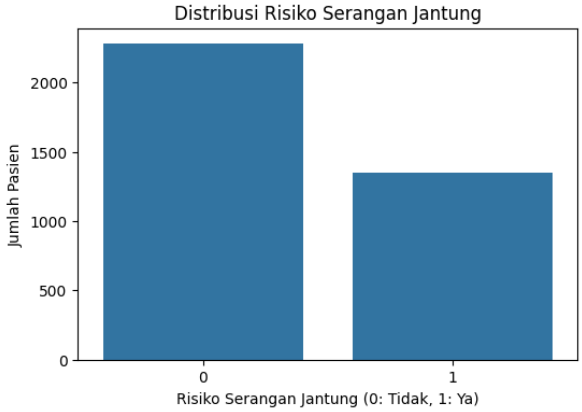
Dataset yang digunakan adalah dataset publik dari DeepDataLake dalam format CSV. Dataset tidak menyediakan metadata asal negara/daerah, fasilitas kesehatan, maupun distribusi demografi spesifik, sehingga penelitian ini diposisikan sebagai studi eksploratif berbasis data sekunder, bukan estimasi prevalensi atau representasi populasi Indonesia. Dataset ini berisi 158.355 rekam pasien dengan 28 variabel yang mencakup demografi, gaya hidup, pemeriksaan klinis, dan riwayat kesehatan. Dataset dipilih karena ukuran besar dan cakupan fitur yang luas, sehingga representatif untuk pembangunan model klasifikasi penyakit jantung. Untuk analisis, peneliti menggunakan 6 fitur utama yang sering disebut dalam literatur: *age, hypertension*, diabetes, *cholesterol_level, obesity*, dan *previous_heart_disease*.

b. Variabel Penelitian

Variabel heart_attack digunakan sebagai variabel target dalam penelitian ini dan bersifat biner, dengan nilai 1 merepresentasikan individu yang tercatat memiliki kejadian atau status risiko serangan jantung, dan nilai 0 merepresentasikan individu tanpa kejadian tersebut. Label ini berasal dari dataset sekunder dan tidak merepresentasikan diagnosis klinis langsung oleh tenaga medis. Distribusi kelas dianalisis sebelum dan sesudah pembagian data untuk memastikan proporsi kelas yang seimbang. Penyeimbangan kelas dilakukan untuk meminimalkan bias model terhadap kelas mayoritas dan meningkatkan interpretabilitas metrik evaluasi seperti recall dan AUC.

c. Pengolahan Data

Data diproses menggunakan Google Colab dengan bahasa pemrograman Python. Tahap awal pengolahan data diawali dengan penyaringan data berdasarkan rentang usia 25–30 tahun untuk memfokuskan analisis pada subkelompok dewasa muda. Selanjutnya dilakukan analisis distribusi variabel target heart_attack menggunakan visualisasi countplot untuk mengamati keseimbangan kelas sebelum proses sampling.



Gambar 2. Distribusi Variabel Target Sebelum Penyeimbangan Data

Gambar 2 menunjukkan distribusi kelas variabel target heart_attack sebelum dilakukan penyeimbangan data. Terlihat adanya ketidakseimbangan kelas yang berpotensi memengaruhi kinerja model klasifikasi. Untuk mengatasi ketidakseimbangan kelas, dilakukan penyeimbangan data dengan metode undersampling, yaitu dengan mengambil masing–masing 1.300 data dari kelas heart_attack = 1 dan heart_attack = 0. Data dari kedua kelas kemudian digabungkan dan diacak (*shuffling*) untuk menghindari bias urutan data.

Pemeriksaan duplikasi dilakukan untuk memastikan tidak terdapat baris data ganda pada dataset hasil sampling. Selanjutnya dilakukan identifikasi tipe data untuk menentukan kebutuhan encoding. Encoding diterapkan pada variabel bertipe *object*, di mana label encoding digunakan untuk variabel dengan dua nilai unik dan one-hot encoding diterapkan pada variabel dengan tiga nilai unik. Setelah proses encoding, beberapa variabel biner dikonversi ke tipe data boolean untuk memastikan konsistensi format data.

Tahap akhir pengolahan data adalah seleksi fitur, dengan memilih enam variabel utama yang

digunakan dalam pemodelan, yaitu *age, hypertension, diabetes, cholesterol_level, obesity*, dan *previous_heart_disease*, serta variabel target *heart_attack*.

Setelah proses encoding, variabel bertipe boolean dikonversi menjadi format numerik (0/1) untuk memastikan kompatibilitas dengan algoritma klasifikasi. Pembagian data dilakukan terlebih dahulu menggunakan stratified train-test split (80% data latih dan 20% data uji) untuk menjaga proporsi kelas. Normalisasi fitur numerik kemudian dilakukan menggunakan StandardScaler dengan skema fit pada data latih dan transform pada data uji melalui ColumnTransformer, guna mencegah kebocoran data (*data leakage*). Tahap normalisasi ini terutama diperlukan untuk model *baseline* yang sensitif terhadap skala fitur (Logistic Regression, SVM, dan KNN); sementara untuk XGBoost yang berbasis pohon keputusan, normalisasi tidak bersifat wajib namun tetap diterapkan untuk konsistensi pipeline pemodelan dan perbandingan antar model.

d. Seleksi Data dan Alur Sampling

Dataset awal terdiri dari 158.355 baris data dengan 28 variabel. Selanjutnya dilakukan penyaringan berdasarkan rentang usia 25 – 30 tahun sehingga diperoleh sebanyak 3.625 baris data. Pemeriksaan *missing value* dilakukan pada dataset awal, *missing value* ditemukan pada kolom yang tidak digunakan dalam pemodelan. Penelitian ini menerapkan seleksi fitur, sehingga analisis hanya menggunakan variabel *age, hypertension, diabetes, cholesterol_level, obesity, previous_heart_disease,* dan target *heart_attack*. Pada subset fitur terpilih tersebut tidak ditemukan *missing value*, sehingga tahap penanganan *missing value* tidak diterapkan pada fitur yang dianalisis. Untuk mengatasi ketidakseimbangan kelas, dilakukan penyeimbangan data dengan metode *undersampling*, yaitu dengan mengambil masing–masing 1.300 data dari kelas positif (heart_attack = 1) dan kelas negatif (heart_attack = 0). Data dari kedua kelas kemudian digabungkan dan diacak untuk menghindari bias urutan sebelum dilakukan pembagian data latih dan data uji. Pembagian data latih dan data uji dilakukan menggunakan stratified split untuk menjaga proporsi kelas target pada kedua subset. Pada tahap evaluasi, data uji berjumlah 520 sampel, sesuai akumulasi nilai pada confusion matrix

e. Penerapan Algoritma XGBoost

XGBoost digunakan untuk membangun model prediksi dini risiko penyakit jantung berdasarkan data klinis pasien. Dataset dipisah menjadi data pelatihan 80% dan data pengujian 20% untuk memungkinkan pelatihan model dan evaluasi pada data yang belum pernah dilihat. Selama pelatihan, dilakukan penalaan (tuning) hiperparameter utama seperti n_estimators, max_depth, dan learning_rate untuk memperoleh performa optimal. Optimasi hiperparameter dilakukan menggunakan RandomizedSearchCV dengan skema validasi silang (*cross-validation*) 6-fold pada data latih, metrik penilaian (*scoring*) roc_auc, jumlah iterasi pencarian 150, serta random_state 42 untuk menjaga reprodusibilitas. Nilai AUC rata–rata terbaik pada cross-validation adalah 0,7814 dan digunakan sebagai dasar pemilihan kombinasi hiperparameter terbaik. Model terbaik hasil tuning kemudian dilatih pada data latih dan dievaluasi pada data uji terpisah, dengan nilai AUC sebesar 0,8443. Konfigurasi ruang parameter yang dieksplorasi ditampilkan pada Tabel 1, sedangkan kombinasi parameter terbaik hasil tuning ditampilkan pada Tabel 2.

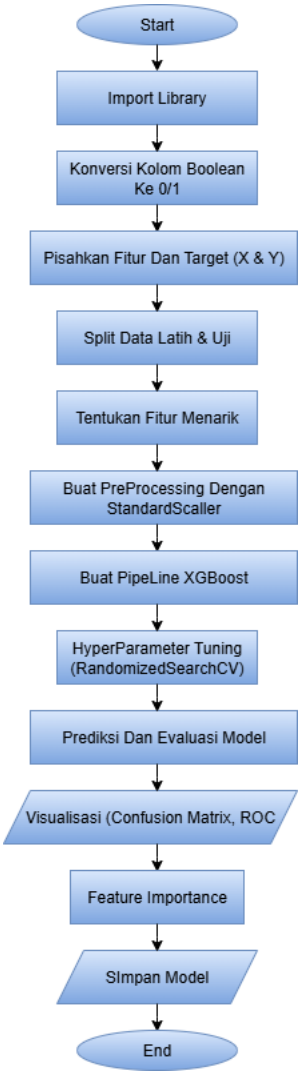
Tabel 1. Konfigurasi Tuning

Parameter	Rentang
classifier__n_estimators	100, 200, 300, 500
classifier__learning_rate	0.01, 0.05, 0.1, 0.2
classifier__max_depth	3, 5, 7, 10
classifier__subsample	0.6, 0.8, 1.0
classifier__colsample_bytree	0.6, 0.8, 1.0
classifier__gamma	0, 0.1, 0.2
classifier__reg_alpha	0, 0.1, 0.5, 1
classifier__reg_lambda	0, 0.1, 0.5, 1

Tabel 2. *Best Hyperparameter*

Parameter	Nilai
classifier__n_estimators	100
classifier__learning_rate	0.01
classifier__max_depth	3
classifier__subsample	0.8
classifier__colsample_bytree	0.6
classifier__gamma	0.2
classifier__reg_alpha	1
classifier__reg_lambda	0.1
Skor CV (mean)	0.7814

Kombinasi hiperparameter terbaik hasil RandomizedSearchCV digunakan untuk melatih model XGBoost akhir pada seluruh data latih. Model kemudian dievaluasi pada data uji menggunakan confusion matrix serta *metrik accuracy, precision, recall, F1-Score*, dan AUC. Untuk keperluan evaluasi komparatif, model baseline (Logistic Regression, Random Forest, SVM, dan KNN) digunakan sebagai pembanding dengan konfigurasi default scikit-learn tanpa proses tuning, sehingga perbandingan berfokus pada efektivitas model utama.



Gambar 3. Alur Penerapan Algoritma XGBoost

HASIL PENELITIAN

Model XGBoost berhasil dibangun untuk memprediksi risiko penyakit jantung pada subkelompok dewasa muda usia 25–30 tahun menggunakan data sekunder berbasis dataset publik menggunakan enam fitur klinis yaitu, umur, hipertensi, diabetes mellitus, kadar kolesterol total, obesitas sentral, dan riwayat penyakit jantung sebelumnya. Penyakit jantung koroner tetap menjadi penyebab utama kematian mendadak pada individu usia di bawah 45 tahun, dan hampir seluruh pasien penyakit jantung koroner usia muda memiliki minimal satu faktor risiko konvensional[3]. Studi di RSUP Haji Adam Malik Medan menunjukkan bahwa 72,7 % pasien penyakit jantung koroner usia muda memiliki riwayat hipertensi, menjadikannya faktor yang sangat dominan dalam kelompok usia ini[10]. Dalam studi yang sama, 61,4 % pasien penyakit jantung koroner memiliki riwayat diabetes terdiagnosis [10]. Kolesterol, bersama dengan diabetes, hipertensi, obesitas, perilaku merokok, ketidakaktifan fisik, dan pola makan tidak sehat merupakan penyebab penting penyakit kardiovaskular [1].

Dalam analisis data Riskesdas 2013–2018, Obesitas sentral memiliki adjusted odds ratio (OR) berkisar antara 1,30 hingga 10,09 terhadap kejadian penyakit jantung koroner bersama dengan faktor risiko lain seperti hipertensi, diabetes, dan kebiasaan merokok[2]. Pasien yang pernah mengalami serangan penyakit jantung koroner dan tidak melakukan aktivitas fisik secara rutin akan berisiko tinggi mengalami kejadian berulang. Hal serupa juga disampaikan oleh P2PTM Kemenkes RI (2019), yang menegaskan pentingnya deteksi dan kontrol jangka panjang pada pasien dengan riwayat penyakit jantung koroner di usia muda[14] Penyakit yang paling mematikan di dunia salah satunya ialah serangan jantung, dimana kasus baru dari penyakit jantung itu sendiri ialah 43.32% dengan jumlah kematian sebesar 12.91%[15]. Oleh karena itu, variabel target heart_attack dipilih sebagai outcome biner (1 = berisiko, 0 = tidak berisiko) untuk memprediksi keberadaan risiko penyakit jantung berdasarkan keenam fitur tersebut. Model diuji pada 520 sampel dengan hasil akurasi sebesar 76,15%, yang berarti sekitar 76% prediksi model sesuai dengan kondisi sebenarnya. Nilai akurasi ini mengindikasikan bahwa model memiliki performa klasifikasi yang cukup baik untuk digunakan sebagai alat bantu skrining awal. Detail metrik evaluasi seperti *precision*, *recall*, dan *F1-Score* untuk masing–masing kelas disajikan pada Tabel 3.

Tabel 3. Classification Report dan Akurasi XGBoost

Metrik	Nilai
Precision (Class 0)	0.76
Recall (Class 0)	0.77
F1-Score (Class 0)	0.76
Precision (Class 1)	0.77
Recall (Class 1)	0.75
F1-Score (Class 1)	0.76

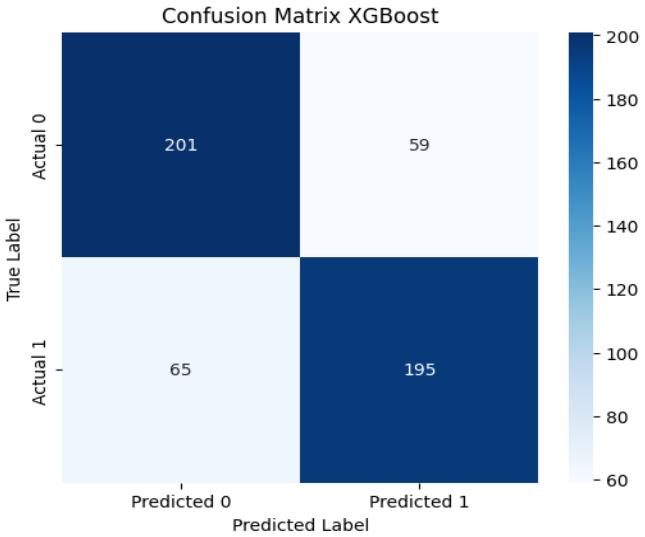
.

Tabel 4. Perbandingan Performa Model Klasifikasi

Model	Accuracy	Recall (Class 1)	F1-Score (Class 1)
Logistic Regression	0.7288	0.74	0.73
Random Forest	0.7135	0.73	0.72
SVM	0.5442	0.28	0.38
KNN	0.6327	0.55	0.60
XGBoost	0.7615	0.75	0.76

Untuk mengevaluasi keunggulan model XGBoost, dilakukan perbandingan dengan beberapa algoritma klasifikasi umum, yaitu Logistic Regression, Random Forest, Support Vector Machine (SVM), dan K-Nearest Neighbor (KNN). Hasil pengujian pada data uji berjumlah 520 sampel menunjukkan bahwa XGBoost memiliki performa terbaik dengan akurasi sebesar 76,15%, *recall* kelas positif sebesar 0,75, dan *F1-Score* sebesar 0,76. Logistic Regression menunjukkan performa yang cukup kompetitif dengan akurasi

72,88%, namun masih berada di bawah XGBoost. Random Forest memberikan hasil yang relatif lebih rendah, sedangkan SVM dan KNN menunjukkan performa yang kurang optimal, terutama pada recall kelas positif. Hasil ini menegaskan bahwa XGBoost lebih efektif dalam mendeteksi individu berisiko penyakit jantung dibandingkan model baseline lainnya pada dataset yang digunakan.

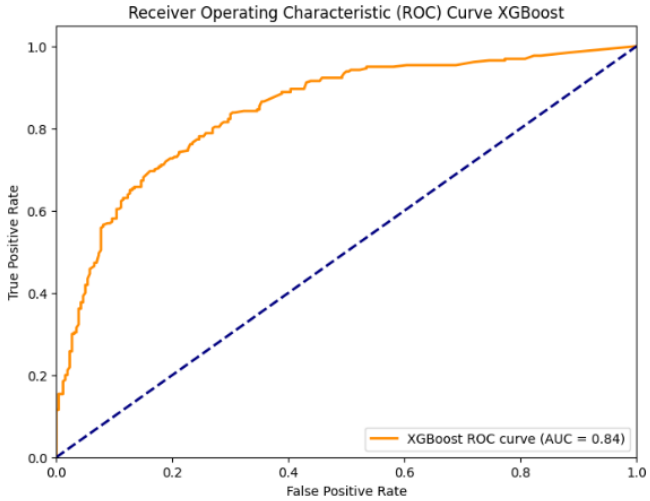


Gambar 4. *Confusion Matrix*

Perbedaan nilai akurasi dengan beberapa studi sebelumnya yang melaporkan akurasi lebih tinggi dapat disebabkan oleh perbedaan karakteristik populasi, jumlah dan jenis fitur yang digunakan, serta perbedaan strategi penyeimbangan kelas dan prosedur validasi. Penelitian ini fokus pada subkelompok usia 25–30 tahun dengan enam fitur klinis utama, sehingga kompleksitas informasi yang tersedia lebih terbatas dibandingkan studi yang menggunakan fitur lebih banyak atau populasi yang lebih heterogen.

Berdasarkan confusion matrix (TP=195, TN=201, FP=59, FN=65), sensitivitas/recall untuk kelas positif adalah 0,75 (195/(195+65)), sedangkan spesifisitas untuk kelas negatif adalah 0,77 (201/(201+59)). False Negative Rate (FNR) sebesar 0,25 menunjukkan bahwa sekitar 25% individu berisiko tidak terdeteksi oleh model. Dalam konteks skrining, false negative lebih berbahaya dibanding false positive karena dapat menunda intervensi dini, sehingga hasil ini perlu dipertimbangkan jika model digunakan sebagai alat bantu skrining awal.

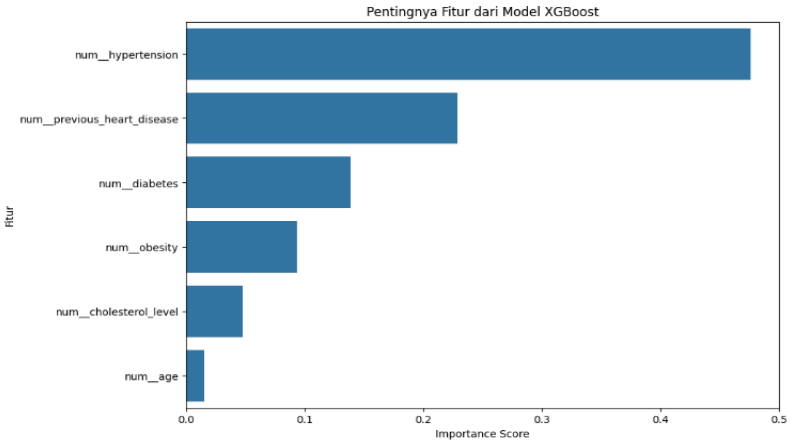
Model menggunakan ambang keputusan (threshold) default 0,5. Penyesuaian threshold berpotensi menurunkan false negative dengan konsekuensi peningkatan false positive; pendekatan ini dapat dipertimbangkan pada implementasi skrining untuk memprioritaskan sensitivitas. Untuk memahami distribusi prediksi model secara lebih rinci, Gambar 4. menunjukkan hasil prediksi model terhadap kelas positif dan negatif. Model mampu mengidentifikasi 195 kasus benar positif (*True Positive*) dan 201 kasus benar negatif (*True Negative*), namun masih terdapat 59 *False Positive* dan 65 *False Negative*. Angka FN yang mencapai 12,5% dari total pengujian menandakan perlunya perhatian klinis, karena pasien berisiko yang tidak terdeteksi dapat kehilangan kesempatan untuk intervensi dini.



Gambar 5. ROC *curve*

Lebih lanjut, performa model dalam membedakan kedua kelas diperkuat oleh *Area Under the Curve* (AUC) sebesar 0,8443. Nilai ini menunjukkan bahwa dalam 84% kasus, model mampu membedakan pasien berisiko dari yang tidak, sehingga valid untuk digunakan dalam konteks klasifikasi medis.

Perbedaan performa akurasi penelitian ini (76,15%) dibandingkan beberapa studi terdahulu yang melaporkan akurasi lebih tinggi (misalnya 93%) dapat disebabkan oleh beberapa faktor. Pertama, perbedaan karakteristik dataset dan populasi akan memengaruhi pola distribusi fitur dan tingkat separabilitas kelas. Kedua, penelitian ini dibatasi pada subkelompok usia 25–30 tahun dan menggunakan enam fitur klinis utama, sehingga informasi yang tersedia untuk pemodelan lebih terbatas dibandingkan studi yang menggunakan populasi lebih luas dan fitur yang lebih beragam. Ketiga, strategi penyeimbangan kelas (*undersampling*) dan prosedur validasi/tuning yang berbeda dapat menghasilkan nilai akurasi yang tidak langsung sebanding antar penelitian. Oleh karena itu, interpretasi performa model pada penelitian ini tidak hanya bertumpu pada akurasi, tetapi juga pada kemampuan diskriminasi yang ditunjukkan oleh nilai AUC sebesar 0,8443.



Gambar 6. Grafik Fitur dari Model XGBoost

Analisis lebih lanjut dilakukan terhadap kontribusi tiap fitur dalam pengambilan keputusan model. Feature importance pada model XGBoost dihitung menggunakan pendekatan *gain-based importance*, berdasarkan importance score pada Gambar 6, fitur hipertensi merupakan prediktor paling dominan, disusul oleh riwayat penyakit jantung dan diabetes. Sementara itu, obesitas, kolesterol, dan usia berkontribusi lebih rendah namun tetap relevan. Temuan ini konsisten dengan literatur bahwa kondisi kronis merupakan indikator kuat dalam risiko penyakit jantung pada usia muda.

Tabel 5. *True Negative*

No.	age	hypertension	diabetes	cholesterol_level	obesity	previous_heart_disease
1	26	1	0	220	0	0
2	25	0	0	209	0	0
3	26	1	0	193	0	0
4	30	0	0	106	0	1
5	29	0	0	231	0	0
6	25	0	0	221	0	0
7	30	0	0	182	0	0
8	25	0	0	164	0	0
9	25	0	0	205	0	0
10	30	0	0	150	0	0

Tabel 6. *True Positive*

No.	age	hypertension	diabetes	cholesterol_level	obesity	previous_heart_disease
1	30	1	0	202	0	0
2	25	0	0	246	0	1
3	30	1	1	128	0	0
4	29	0	1	150	1	1
5	26	0	0	125	0	1
6	30	1	0	211	1	0
7	25	0	0	146	0	1
8	25	0	1	253	0	0
9	25	1	0	213	1	0
10	28	0	0	269	0	1

Contoh konkret dari prediksi model dapat dilihat pada Tabel 5. dan Tabel 6. Data ini menunjukkan bahwa model mampu mengenali pasien sehat dengan karakteristik risiko rendah dan pasien berisiko dengan kombinasi faktor seperti hipertensi, kolesterol tinggi, atau riwayat penyakit jantung. Hal ini menegaskan sensitivitas model dalam mendeteksi pola risiko yang kompleks meskipun dengan jumlah fitur yang terbatas. Namun, masih terdapat keterbatasan. *False Negative* yang cukup signifikan menunjukkan bahwa beberapa faktor risiko tambahan belum tercakup dalam model. Faktor seperti kebiasaan merokok, stres, dan riwayat keluarga tidak dianalisis dalam penelitian ini, namun berpotensi besar dalam meningkatkan prediksi model di masa mendatang. Oleh karena itu, untuk pengembangan lebih lanjut, disarankan penggunaan fitur yang lebih luas serta eksplorasi model klasifikasi lainnya untuk perbandingan performa.

Secara keseluruhan, hasil penelitian ini mendukung penggunaan XGBoost sebagai alat bantu prediktif dalam deteksi dini penyakit jantung pada dewasa muda. Dengan nilai AUC yang tinggi dan distribusi performa yang seimbang, model dapat diandalkan sebagai bagian dari sistem skrining preventif berbasis data, khususnya di populasi usia produktif.

KESIMPULAN

Penelitian berhasil mengimplementasikan algoritma XGBoost untuk prediksi dini risiko penyakit jantung pada subkelompok dewasa muda usia 25–30 tahun berbasis data sekunder. Model yang dibangun menunjukkan performa yang baik dengan akurasi sebesar 76,15% dan nilai AUC sebesar 0,8443, yang mengindikasikan kemampuan diskriminasi model yang kuat dalam membedakan individu berisiko dan tidak berisiko. Nilai presisi, *recall*, dan *F1-Score* yang relatif seimbang pada kedua kelas menunjukkan bahwa model memiliki stabilitas dan konsistensi dalam mengidentifikasi kasus positif maupun negatif. Analisis confusion matrix memperlihatkan bahwa sebagian besar kasus berhasil diklasifikasikan dengan benar, meskipun masih terdapat sejumlah *false negative* yang perlu mendapat perhatian, terutama dalam konteks klinis, namun demikian, proporsi kesalahan tersebut masih berada pada tingkat yang dapat diterima untuk penggunaan sebagai alat bantu skrining awal. Hasil analisis *feature importance* menunjukkan bahwa hipertensi merupakan faktor risiko paling dominan, diikuti oleh riwayat penyakit jantung sebelumnya dan diabetes mellitus, sementara obesitas, kadar kolesterol, dan usia tetap berkontribusi meskipun dengan pengaruh yang lebih rendah.

Secara keseluruhan, XGBoost potensial digunakan sebagai sistem pendukung keputusan untuk skrining dini risiko penyakit jantung pada subkelompok dewasa muda usia 25–30 tahun, namun penelitian ini bersifat eksploratif sehingga temuan perlu divalidasi lebih lanjut menggunakan data klinis terverifikasi dan populasi dengan metadata yang jelas. Untuk penelitian selanjutnya, disarankan penggunaan dataset yang lebih representatif, penambahan variabel risiko lain (misalnya kebiasaan merokok, aktivitas fisik, stres, riwayat keluarga), evaluasi penyesuaian threshold atau pendekatan cost-sensitive untuk menekan false negative, serta validasi eksternal pada dataset berbeda dan penggunaan metode interpretabilitas tambahan (misalnya SHAP) agar model lebih andal dan transparan.

REFERENCES

[1] T. Arini, P. Studi Ilmu Keperawatan, S. Tinggi Ilmu Kesehatan Bhakti Husada Mulia Madiun, and F. Nafiul Umam, "PENGUKURAN LEMAK TUBUH DAN INDEKS MASSA TUBUH SEBAGAI UPAYA PENCEGAHAN RESIKO PENYAKIT KARDIOVASKULAR," vol. IV, no. 1, 2021, [Online]. Available: http://e-journal.lppmdianhusada.ac.id/index.php/jbca

[2] L. Ghani, M. D. Susilawati, and H. Novriani, "Faktor Risiko Dominan Penyakit Jantung Koroner di Indonesia," *Buletin Penelitian Kesehatan*, vol. 44, no. 3, Dec. 2016, doi: 10.22435/bpk.v44i3.5436.153-164.

[3] J. K. Masyarakat, S. Husada, M. I. Jamil, M. Yanni, and M. Noer, "GAMBARAN FAKTOR RISIKO PENYAKIT JANTUNG KORONER PADA USIA MUDA," vol. 1, no. 7, pp. 72–87, 2024, doi: 10.572349/husada.v1i1.363.

[4] Z. A. Zaqi, M. Anshori, and W. T. Kusuma, "Prediksi Dini Penyakit Diabetes pada Wanita dengan Algoritma Klasifikasi Berbasis Pohon," *Jurnal Ilmu Komputer*, vol. 18, no. 1, Apr. 2025, doi: https://doi.org/10.24843/JIK.2025.v18.i01.p05.

[5] D. Mayang Pratiwi and L. Mufidah, *Perbandingan Metode Decision Tree Classifier dan XGBoost Classifier Dalam Memprediksi Penyakit Jantung*, vol. 4, no. 1. 2024. doi: https://doi.org/10.20895/centive.v4i1.336.

[6] L. Azhari, N. Wulandari, F. Adiningrat, and A. D. Alexander, "Model Prediksi Penyakit Jantung dengan Penanganan Outlier Menggunakan Interquartile Range dan Extreme Gradient Boosting," *Journal of Information System Research (JOSH)*, vol. 6, no. 2, Jan. 2025, doi: 10.47065/josh.v6i2.6390.

[7] World Health Organization (WHO), "Cardiovascular diseases." Accessed: Jan. 26, 2026. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[8] W. Wardiyana and T. Herawati, "Literature Review: Health Behavior Changes in Heart Disease Patients Based on Smartphone Game Applications," *Journal of Health and Cardiovascular Nursing*, vol. 3, no. 2, pp. 71–82, Dec. 2023, doi: 10.36082/jhcn.v3i2.1462.

[9] A. D. Hoerul Anwar and A. Farmasi YPF Bandung, "SISTEMATIC REVIEW FAKTOR RESIKO PENYAKIT JANTUNG KORONER DI INDONESIA," 2025, doi: https://doi.org/10.64094/fqanc998.

[10] L. Farida Tampubolon, A. Ginting, F. Ermasta Saragi Turnip, Stik. Santa Elisabeth Medan, J. Bunga Terompet No, and K. Medan Selayang, "GAMBARAN FAKTOR YANG MEMPENGARUHI KEJADIAN PENYAKIT JANTUNG KORONER (PJK) DI PUSAT JANTUNG TERPADU (PJT)," 2023. doi: 10.32583/pskm.v13i3.1077.

[11] M. Dava Maulana *et al.*, "ALGORITMA XGBOOST UNTUK KLASIFIKASI KUALITAS AIR MINUM," 2023. doi: https://doi.org/10.36040/jati.v7i5.7308.

[12] D. T. Murdiansyah, "Prediksi Stroke Menggunakan Extreme Gradient Boosting," *JIKO (Jurnal Informatika dan Komputer)*, vol. 8, no. 2, p. 419, Sep. 2024, doi: 10.26798/jiko.v8i2.1295.

[13] A. Brahmandjati, A. Mizwar, A. Rahim, and F. Asharudin, "Optimasi Prediksi Diabetes Dengan Algoritma XGBoost Dan Teknik Preprocessing Data," 2024. [Online]. Available: https://journal.mediapublikasi.id/index.php/logic/article/view/4963

[14] P. Kharisma Ningtias *et al.*, "HUBUNGAN TINGKAT AKTIVITAS FISIK TERHADAP KEJADIAN RAWAT INAP ULANG PADA PENYAKIT JANTUNG KORONER," 2025. doi: https://doi.org/10.26714/jkj.13.2.2025.349-358.

[15] T. Aniamarta, A. Salsabilla Huda, and F. Lizariani Aqsha, "Causes and Treatments of Heart Attack," *BIOLOGICA SAMUDRA*, vol. 4, no. 1, pp. 22–31, Jul. 2022, doi: 10.33059/jbs.v4i1.3925.