Jurnal INSAN (Journal of Information Systems Management Innovation)
Volume 5 No. 2 Desember 2025
E-ISSN: 2777-1385

Penerapan Metode Algoritma XGBoost untuk Prediksi Risiko Penyakit Jantung

Fadia Izni Adani[1], Hilda Amalia[2]

[1,2]Program Studi Teknologi Informasi, Fakultas Teknik dan Informasi, Universitas Bina Sarana Informatika
e-mail: [1]adaniizni@gmail.com, [2]hilda.ham@bsi.ac.id

Artikel Info : Diterima : 18-09-2025 | Direvisi : 21-11-2025 | Disetujui : 26-11-2025

Abstrak - Penyakit jantung masih menjadi salah satu penyebab utama kematian di dunia, termasuk di Indonesia, sehingga deteksi dini terhadap faktor risikonya menjadi sangat penting untuk menekan angka kematian dan meningkatkan kualitas hidup pasien. Penelitian ini menerapkan algoritma XGBoost untuk membangun model prediksi risiko penyakit jantung menggunakan dataset *cardio_train.csv* dari Kaggle, yang berisi data kesehatan pasien meliputi usia, jenis kelamin, tekanan darah, kolesterol, serta variabel lain yang relevan. Tahapan penelitian meliputi pra-pemrosesan data, pelatihan model, evaluasi performa menggunakan metrik akurasi dan AUC, serta analisis *feature importance*. Hasil menunjukkan bahwa model XGBoost mampu mencapai *precision* sebesar 0,72, *recall* sebesar 0,77, *F1-score* sebesar 0,74, *akurasi* sebesar 73%, serta nilai AUC sebesar 0,795, yang menandakan kemampuan klasifikasi yang cukup baik. Fitur-fitur seperti tekanan darah sistolik (*ap_hi*), usia (*age*), dan kolesterol merupakan faktor dominan dalam proses prediksi. Dengan hasil ini, XGBoost dapat direkomendasikan sebagai metode dalam pengembangan sistem pendukung keputusan untuk deteksi dini penyakit jantung secara otomatis.

Kata Kunci : *machine learning,* penyakit jantung, XGBoost

Abstract – *Heart disease remains one of the leading causes of death worldwide, including in Indonesia, making early detection of its risk factors crucial to reducing mortality rates and improving patients' quality of life. This study applies the XGBoost algorithm to build a predictive model for heart disease risk using the cardio_train.csv dataset from Kaggle, which contains patient health data such as age, gender, blood pressure, cholesterol, and other relevant variables. The research stages include data preprocessing, model training, performance evaluation using accuracy and AUC metrics, as well as feature importance analysis. The results show that the XGBoost model achieved a precision of 0.72, recall of 0.77, F1-score of 0.74, accuracy of 73%, and an AUC value of 0.795, indicating a fairly good classification capability. Features such as systolic blood pressure (ap_hi), age, and cholesterol were found to be the most dominant factors in the prediction process. Based on these findings, XGBoost can be recommended as a method for developing decision support systems for the early detection of heart disease automatically.*

Keywords : machine learning, heart disease, XGBoost

PENDAHULUAN

Penyakit jantung, yang dalam istilah medis sering disebut sebagai *Cardiovascular Disease* (CVD), merupakan kondisi yang memengaruhi jantung dan pembuluh darah serta menjadi salah satu penyebab kematian tertinggi di dunia. Menurut data World Health Organization (WHO) tahun 2022, penyakit kardiovaskular (CVD) bertanggung jawab atas sekitar 17,9 juta kematian setiap tahun (Soleha et al., 2025). Di Indonesia, penyakit jantung juga menjadi salah satu penyebab kematian tertinggi pada kategori penyakit tidak menular (Hoerul Anwar, 2025). Beberapa faktor utama yang meningkatkan risiko penyakit jantung di Indonesia meliputi tekanan darah tinggi, kadar gula darah yang tinggi, riwayat penyakit jantung dalam keluarga, kebiasaan merokok, obesitas, serta pola hidup yang tidak sehat (Susanti et al., 2024)

Salah satu tantangan utama dalam penanganan penyakit jantung adalah kemampuan yang belum maksimal dalam melakukan deteksi dini terhadap faktor risikonya (Chandra & Prasetyo, 2024). Tidak jarang,

pasien atau penderita penyakit jantung baru menyadari kondisinya karena kurangnya kesadaran akan pentingnya pemeriksaan kesehatan secara rutin dan keterlambatan dalam mendeteksi ketika gejala telah mengindikasikan stadium lanjut.

Pertumbuhan data medis yang pesat dan rumit menghadirkan serangkaian tantangan yang perlu diatasi, sekaligus menawarkan potensi besar untuk penerapan teknologi *machine learning* dan data *mining* dalam pengelolaan dan pemanfaatannya. Data mining merupakan suatu kegiatan interaktif yang melibatkan penggunaan data numerik dan kecerdasan buatan untuk menemukan serta mengenali informasi yang relevan dari sekumpulan data berukuran besar (Yulianti et al., 2022). Kemampuan data *mining* dalam mengidentifikasi informasi relevan dari dataset berukuran besar memungkinkan pemanfaatannya oleh algoritma *machine learning* berbasis *supervised learning* dalam membangun model yang mampu melakukan prediksi berdasarkan data historis pasien.

XGBoost merupakan metode pengembangan dari algoritma *gradient tree boosting* yang menggunakan pendekatan *ensemble* (Rayadin et al., 2024). Algoritma ini sangat efektif dalam menangani permasalahan *machine learning* dengan skala data yang besar, pemilihan XGBoost didasari oleh fitur-fitur tambahannya yang bermanfaat untuk mempercepat proses komputasi dan menghindari *overfitting.* (Yulianti et al., 2022).

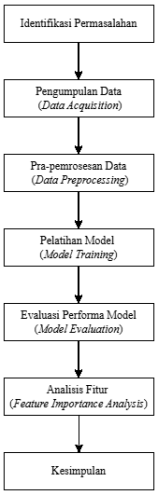
Penelitian terdahulu menunjukkan performa XGBoost yang unggul dibanding algoritma lain. Misalnya, penelitian klasifikasi keputusan kredit menggunakan dataset *Credit Card Approval-With Target* menunjukkan XGBoost lebih akurat dibanding Random Forest (Dachi & Sitompul, 2023). Dalam penelitian prediksi cuaca, XGBoost juga menghasilkan prediksi suhu rata-rata lebih akurat dibandingkan Random Forest dan SVR berdasarkan nilai MAE, MSE, dan R^2 (Syahreza et al., 2024). Namun, penerapan XGBoost dalam prediksi penyakit jantung masih terbatas.

Oleh karena itu, studi ini akan berfokus pada penerapan algoritma XGBoost untuk membangun model prediktif penyakit jantung yang efektif. Peggunaan algortima XGBoost pada dataset penyakit jantung masih belum banyak dilakukan sehingga pada penelitian akan dilakukan pengolahan dataset penyakit jantung dengna metode algoritma XGBoost. Selain itu, penelitian ini akan menganalisis atribut atau fitur data yang paling berpengaruh terhadap prediksi risiko penyakit jantung yang dihasilkan oleh model XGBoost. Kebaruan dari penelitian ini terletak pada penerapan dan optimalisasi algoritma XGBoost secara khusus untuk prediksi risiko penyakit jantung melalui proses *hyperparameter tuning* dan penyeimbangan data, serta analisis interpretabilitas model untuk mengidentifikasi fitur-fitur yang paling berpengaruh terhadap hasil prediksi. Dengan mengoptimalkan kemampuan prediksi dan interpretasi model XGBoost, penelitian ini diharapkan mampu memberikan kontribusi inovatif dalam upaya deteksi dini risiko penyakit jantung secara lebih akurat dan informatif.

METODOLOGI PENELITIAN

1. Tahapan Penelitian

Berikut merupakan tahapan penelitian yang dilakukan dengan pendekatan kuantitatif dan metode pemodelan prediktif:



Sumber: Laporan Penelitian (2025)

Gamba 1 Tahapan Penelitian

Gamba 1 Tahapan Penelitian

Gambar di atas (Rayadin et al., 2024)menyajikan alur penelitian yang penulis lakukan, dengan tahapan sebagai berikut:

a. Identifikasi Permasalahan
   Penelitian ini diawali dengan identifikasi permasalahan terkait meningkatnya angka penyakit jantung yang menjadi salah satu penyebab utama kematian di dunia. Karena itu, diperlukan sebuah sistem yang mampu memprediksi risiko penyakit jantung secara lebih dini dan akurat. Tujuan penelitian ini adalah mengimplementasikan metode algoritma XGBoost adalam membangun model prediktif yang dapat memperkirakan risiko seseorang terkena penyakit jantung berdasarkan data rekam medis atau data karakteristik individu.
b. Pengumpulan Data (*Data Acquisition*)
   Tahapan berikutnya dilakukan dengan mengumpulkan data yang menjadi dasar dalam proses pemodelan. Penelitian ini menggunakan data yang berasal dari Kaggle, yang memuat informasi mengenai kondisi kesehatan pasien, termasuk usia, tekanan darah, kolesterol, jenis kelamin, dan variabel relevan lainnya terkait risiko penyakit jantung. Dataset ini dipilih karena sering digunakan dalam penelitian sejenis dan memiliki struktur yang cocok untuk penerapan metode pembelajaran mesin.
c. Pra-pemrosesan Data (*Data Preprocessing*)
   Setelah data diperoleh, dilakukan tahap pra-pemrosesan data dengan tujuan memastikan bahwa data yang dipakai dalam kondisi bersih dan siap diolah. Proses yang dilakukan melibatkan penanganan data yang hilang (*missing values*), normalisasi atau standarisasi nilai numerik jika diperlukan, serta konversi data kategorikal menjadi bentuk numerik menggunakan teknik encoding. Selain itu, pembagian data dilakukan ke dalam data latih dan data uji dengan rasio tertentu guna memastikan proses pelatihan dan pengujian model berjalan secara optimal.
d. Pelatihan Model (*Model Training*)
   Pada tahap ini, dilakukan pelatihan model menggunakan algoritma XGBoost, yaitu salah satu algoritma *machine learning* berbasis *ensemble learning* yang memiliki kemampuan tinggi dalam klasifikasi data. Pelatihan model dilakukan pada data latih yang telah dipersiapkan melalui tahapan pra-pemrosesan. Dalam upaya meningkatkan kinerja model, dilakukan penyesuaian parameter (*hyperparameter tuning*) serta validasi silang (*cross-validation*) untuk mencegah *overfitting* dan memperkuat kemampuan model dalam menggeneralisasi pada data baru.
e. Evaluasi Performa Model (*Model Evaluation*)
   Setelah model dilatih, dilakukan evaluasi untuk mengukur seberapa baik model dalam melakukan prediksi risiko penyakit jantung. Evaluasi dilakukan menggunakan beberapa metrik seperti akurasi, presisi, *recall*, *F1-score*, dan ROC-AUC. Hasil evaluasi ini menjadi dasar dalam menentukan efektivitas model dalam klasifikasi data serta menunjukkan seberapa layak model diterapkan dalam konteks nyata.
f. Analisis Fitur (*Feature Importance Analysis*)
   Tahap ini bertujuan untuk mengetahui fitur-fitur mana saja yang paling berkontribusi terhadap hasil prediksi. XGBoost memiliki fitur untuk menghitung tingkat kepentingan (*feature importance*) dari setiap variabel. Hasil analisis ini memberikan informasi yang berguna untuk memahami faktor-faktor utama yang mempengaruhi risiko penyakit jantung dan dapat dijadikan bahan pertimbangan dalam pengambilan keputusan medis.
g. Penarikan Kesimpulan dan Saran
   Tahapan terakhir meliputi penarikan kesimpulan yang didasarkan pada hasil evaluasi model dan analisis fitur yang telah dilakukan. Kesimpulan ini merangkum seberapa baik model XGBoost dalam memprediksi risiko penyakit jantung dan fitur apa saja yang paling berpengaruh. Penelitian ini juga memberikan saran agar model yang dikembangkan dapat diintegrasikan ke dalam sistem pendukung keputusan di bidang kesehatan untuk membantu tenaga medis dalam mendeteksi dini potensi penyakit jantung pada pasien.

2. Instrumen Penelitian

Instrumen penelitian adalah alat yang digunakan untuk mengumpulkan data serta mengukur objek atau variabel penelitian secara terstruktur dan sistematis (Muslihin et al., 2022). Untuk mendukung pelaksanaan penelitian, digunakan sejumlah instrumen yang dijelaskan sebagai berikut:

a. Dataset
   Penelitian menggunakan dataset penyakit jantung yang diperoleh dari Kaggle (Johson, 2023), yang mencakup informasi seperti jenis kelamin, usia, tekanan darah, kolesterol, kadar gula darah, dan fitur medis lainnya yang terkait dengan risiko penyakit jantung.
b. Perangkat Lunak (*Software*)
   Pada penelitian ini, platform pemrograman berbasis cloud yang digunakan adalah *Google Colaboratory* (Google Colab). Bahasa pemrograman yang diterapkan adalah Python, dengan memanfaatkan pustaka seperti Pandas untuk pengolahan data, Scikit-learn untuk tahap *preprocessing* dan evaluasi, serta XGBoost dalam proses pembangunan model prediksi. Untuk visualisasi data, digunakan pustaka Matplotlib dan Seaborn.

3. Metode Pengumpulan Data, Populasi, dan Sample Penelitian

Pada penelitian ini, data dikumpulkan menggunakan dua metode, yaitu studi pustaka dan pemanfaatan dataset sekunder. Maka, metode pengumpulan data yang digunakan dalam penelitian ini dapat dijelaskan sebagai berikut:

a. Studi Pustaka
Peneliti melaksanakan studi pustaka dengan menganalisis berbagai sumber ilmiah yang sesuai dengan fokus penelitian. Sumber-sumber tersebut mencakup jurnal nasional dan internasional, artikel ilmiah, serta referensi dari buku yang membahas tentang data mining, machine learning, algoritma XGBoost, dan topik terkait prediksi risiko penyakit jantung. Studi pustaka ini berfungsi sebagai landasan teori dan sebagai pembanding untuk hasil penelitian yang dilakukan.

b. Data Sekunder
Data utama yang menjadi dasar dalam penelitian ini berasal dari dataset sekunder yang dapat diakses melalui platform Kaggle. Dataset sekunder ini terdiri dari kumpulan data kesehatan yang mencakup usia, jenis kelamin, tekanan darah, kolesterol, detak jantung, dan beberapa variabel lain yang terkait dengan risiko penyakit jantung. Dataset ini merupakan data terbuka yang umum digunakan dalam berbagai penelitian ilmiah. Data tersebut dimanfaatkan dalam proses analisis data dan pembangunan model prediksi.

c. Populasi dan Sampel Penelitian
Karena menggunakan data sekunder yang bersumber dari dataset digital yang telah tersedia, sehingga penelitian ini tidak memerlukan penentuan populasi dan sampel seperti dalam penelitian konvensional.

4. Metode Analisis Data

Analisis data dalam penelitian ini dilakukan secara terstruktur untuk mengolah data mentah menjadi informasi yang bermanfaat dalam pengembangan model prediksi risiko penyakit jantung. Tahapan analisis data yang diterapkan adalah sebagai berikut:

a. *Preprocessing Data*
Langkah pertama dalam analisis data mencakup preprocessing, yang melibatkan beberapa kegiatan penting. Pertama, dilakukan pemeriksaan dan penanganan terhadap nilai yang hilang (*missing values*). Selanjutnya, data kategorikal diubah menjadi format numerik jika diperlukan. Proses ini juga mencakup normalisasi atau standarisasi data agar semua fitur memiliki skala yang konsisten. Terakhir, dataset dipisahkan menjadi dua bagian utama, yaitu data latih (*training set*) dan data uji (*testing set*), dengan perbandingan yang umum digunakan, seperti 80:20 atau 70:30.

b. Pelatihan Model dengan Algoritma XGBoost
Setelah proses persiapan data selesai, peneliti membangun model prediksi menggunakan algoritma XGBoost *(Extreme Gradient Boosting*). Algoritma ini termasuk dalam metode *boosting* berbasis pohon keputusan yang dirancang untuk meminimalkan fungsi loss secara optimal melalui pendekatan boosting yang efisien. Pelatihan model dilakukan pada data latih dengan menyesuaikan sejumlah hyperparameter, seperti n_estimators (jumlah pohon), learning_rate (kecepatan pembelajaran), max_depth (kedalaman pohon), dan lainnya.

c. Evaluasi Kinerja Model
Beberapa metrik evaluasi klasifikasi digunakan untuk menilai kualitas model, seperti:

1. Akurasi (*Accuracy*)
Mengukur proporsi prediksi yang benar terhadap total data.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

2. Presisi (*Precision*)
Mengukur ketepatan prediksi positif.

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

3. *Recall* (*Sensitivity*)
Mengukur seberapa baik model menemukan semua data positif.

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

4. F1-Score
Rata-rata harmonis antara *Precision* dan *Recall*.

$$F1\text{-}Score = 2\ \frac{Precision \times Recall}{Precision + Recall}$$

d. Analisis Fitur (*Feature Importance*)

Setelah proses pembangunan dan evaluasi model selesai, dilakukan analisis terhadap kontribusi masing-masing fitur menggunakan fitur bawaan XGBoost yang menghitung tingkat kepentingan (*importance score*). Langkah ini bertujuan untuk mengidentifikasi variabel yang memiliki pengaruh signifikan terhadap prediksi risiko penyakit jantung.

HASIL DAN PEMBAHASAN

1. Dataset

Penelitian ini menggunakan dataset penyakit jantung (*cardiovascular disease*) yang diperoleh dari salah satu sumber terbuka, yaitu platform Kaggle. Dataset ini berisi data pasien yang mencakup berbagai informasi medis dan gaya hidup, yang digunakan untuk memprediksi risiko penyakit jantung.
Dataset tersebut terdiri dari 70.000 data dan 13 atribut, yang mencakup:

a. 1 kolom identitas (id)
b. 11 kolom sebagai fitur input (variabel bebas)
c. 1 kolom target yaitu cardio (variabel terikat)

Adapun penjelasan untuk masing-masing atribut adalah sebagai berikut:

Tabel 1 Deskripsi Atribut Dataset Penyakit Jantung

Nama Kolom	Deskripsi
id	Identitas unik untuk tiap pasien (tidak digunakan dalam pemodelan)
age	Usia pasien dalam satuan hari
gender	Jenis kelamin (1 = pria, 2 = wanita)
height	Tinggi badan pasien dalam cm
weight	Berat badan pasien dalam kg
ap_hi	Tekanan darah sistolik (atas)
ap_lo	Tekanan darah diastolik (bawah)
cholesterol	Tingkat kolesterol (1 = normal, 2 = di atas normal, 3 = sangat tinggi)
gluc	Tingkat glukosa (1 = normal, 2 = di atas normal, 3 = sangat tinggi)
smoke	Apakah pasien merokok (1 = ya, 0 = tidak)
alco	Apakah pasien mengonsumsi alkohol (1 = ya, 0 = tidak)
active	Apakah pasien aktif secara fisik (1 = ya, 0 = tidak)
cardio	Target/label: Risiko penyakit jantung (1 = berisiko, 0 = tidak berisiko)

Sumber: https://Kaggle.com

2. Preprocessing dan Pemodelan

a. Tahap Preprocessing

Tahap ini dilakukan guna menyiapkan data sebelum masuk ke proses pelatihan model. Pada penelitian ini, preprocessing bertujuan untuk membersihkan dan memisahkan data sesuai dengan kebutuhan pemodelan.

Langkah pertama yang dilakukan adalah menghapus kolom id dari dataset. Kolom ini berfungsi sebagai identitas unik untuk masing-masing entri data dan tidak memiliki kontribusi dalam proses klasifikasi, sehingga tidak digunakan dalam analisis lebih lanjut.

Selanjutnya, data dipisahkan menjadi dua bagian utama, yaitu fitur (X) dan label (y). Fitur (X) terdiri dari seluruh atribut kecuali id dan cardio, sedangkan label (y) diambil dari kolom cardio, yaitu variabel target yang menunjukkan status risiko penyakit jantung. Nilai 0 menunjukkan pasien tidak berisiko, sedangkan nilai 1 menunjukkan pasien berisiko mengalami penyakit jantung.

Berikut adalah potongan kode yang digunakan untuk menghapus kolom id, memisahkan fitur dan label, serta memeriksa dimensi data:

```
# Hapus kolom id
df = df.drop(columns=['id'])

# Pisahkan fitur (X) dan target (y)
X = df.drop(columns=['cardio'])  # Semua fitur
y = df['cardio']                 # Target: 0 = sehat, 1 = berisiko

# Cek dimensi
print("X shape:", X.shape)
print("y shape:", y.shape)
```

```
X shape: (70000, 11)
y shape: (70000,)
```

Sumber: Laporan Penelitian (2025)

Gambar 2 Penghapusan Kolom id dan Pemisahan Fitur serta Label

Hasil dari tahapan tersebut menunjukkan bahwa variabel fitur (X) memiliki 11 atribut dengan jumlah 70.000 baris data, dan variabel label (y) memiliki 70.000 data target yang sesuai.

Setelah proses pemisahan, dilakukan pembagian data menjadi data latih dan data uji. Pembagian ini dilakukan dengan rasio 80:20, yaitu 80% untuk data latih dan 20% untuk data uji. Fungsi train_test_split dari pustaka scikit-learn digunakan untuk proses ini, dengan menambahkan parameter *stratify* agar distribusi kelas tetap seimbang pada kedua bagian data.

```
[15] from sklearn.model_selection import train_test_split

     X_train, X_test, y_train, y_test = train_test_split(
         X, y, test_size=0.2, random_state=42, stratify=y
     )
```

Sumber: Laporan Penelitian (2025)

Gambar 3 Kode Pembagian Data Latih dan Data Uji

b. Tahap Pemodelan

Setelah data berhasil dibagi menjadi data latih dan data uji, langkah selanjutnya adalah membangun model prediksi menggunakan algoritma *Extreme Gradient Boostin*g (XGBoost). Algoritma ini dipilih karena dikenal memiliki kinerja yang tinggi dalam berbagai tugas klasifikasi.

Model yang digunakan adalah XGBClassifier dari pustaka XGBoost. Parameter *eval_metric* diatur menjadi *'logloss'* agar sesuai dengan konteks klasifikasi biner dan untuk menghindari peringatan sistem. Proses pelatihan model dilakukan menggunakan data latih yang telah disiapkan sebelumnya. Potongan kode program untuk pelatihan model ditampilkan pada gambar berikut:

```
from xgboost import XGBClassifier

# Buat model
model = XGBClassifier(eval_metric='logloss')

# Latih model
model.fit(X_train, y_train)
```

Sumber: Laporan Penelitian(2025)

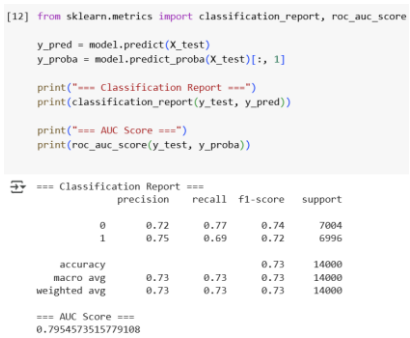
Gambar 4 Kode Pelatihan Model XGBoost

Model yang telah dilatih pada data latih ini kemudian digunakan untuk melakukan prediksi terhadap data uji. Hasil dari prediksi tersebut akan dievaluasi menggunakan metrik evaluasi seperti *precision*, *recall*, *F1-score*, dan AUC (*Area Under Curve*).

3. Hasil Evaluasi Model

Setelah model XGBoost selesai dilatih menggunakan data latih, langkah selanjutnya adalah melakukan evaluasi performa model terhadap data uji (X_test). Evaluasi ini bertujuan untuk mengetahui sejauh mana model mampu melakukan klasifikasi secara akurat terhadap data baru yang belum pernah dilihat sebelumnya.

Evaluasi dilakukan dengan menggunakan dua pendekatan utama, yaitu *classification report* dan AUC (*Area Under Curve*). *Classification report* mencakup metrik-metrik seperti *precision, recall, f1-score,* dan *accuracy*. Sedangkan *AUC score* digunakan untuk mengukur kemampuan model dalam membedakan kelas target secara keseluruhan.



```
[12] from sklearn.metrics import classification_report, roc_auc_score

     y_pred = model.predict(X_test)
     y_proba = model.predict_proba(X_test)[:, 1]

     print("=== Classification Report ===")
     print(classification_report(y_test, y_pred))

     print("=== AUC Score ===")
     print(roc_auc_score(y_test, y_proba))

=== Classification Report ===
              precision    recall  f1-score   support

           0       0.72      0.77      0.74      7004
           1       0.75      0.69      0.72      6996

    accuracy                           0.73     14000
   macro avg       0.73      0.73      0.73     14000
weighted avg       0.73      0.73      0.73     14000

=== AUC Score ===
0.7954573515779108
```

Sumber: Laporan Penelitian (2025)

Gambar 5 Kode Evaluasi Model menggunakan *Classification Report* dan AUC

Cuplikan kode pada gambar di atas menunjukkan proses evaluasi model setelah proses pelatihan selesai dilakukan. Model yang sudah dilatih selanjutnya diuji dengan menggunakan data uji (*X_test*) untuk mengetahui sejauh mana kemampuannya dalam memprediksi risiko penyakit jantung.

Langkah pertama dilakukan dengan memprediksi label dari data uji dan menyimpannya dalam variabel *y_pred*. Selanjutnya, model juga menghasilkan nilai probabilitas prediksi terhadap kelas *1* (yaitu kelas berisiko penyakit jantung), yang disimpan dalam variabel *y_proba*.

Evaluasi performa model dilakukan melalui dua pendekatan utama:

a. *Classification Report*

Berisi metrik evaluasi seperti *precision*, *recall*, dan *f1-score* untuk masing-masing kelas (0 = tidak berisiko, 1 = berisiko). Hasil *classification report* menunjukkan bahwa:

1) Untuk kelas 0:
   a) *Precision* sebesar 0.72, artinya 72% dari prediksi "tidak berisiko" benar.
   b) *Recall* sebesar 0.77, artinya 77% dari data yang benar-benar "tidak berisiko" berhasil dikenali oleh model.
   c) *F1-score* sebesar 0.74, yang merupakan nilai rata-rata harmonis antara *precision* dan *recall*.
2) Untuk kelas 1:
   a) *Precision* sebesar 0.75, *recall* sebesar 0.69, dan *f1-score* sebesar 0.72.
   b) Rata-rata akurasi model secara keseluruhan adalah 73%, artinya 73% dari keseluruhan prediksi model terbukti benar.

b. AUC (*Area Under Curve*)

AUC menilai sejauh mana model mampu membedakan secara keseluruhan antara kelas positif dan negatif. Hasil evaluasi menunjukkan nilai AUC sebesar 0.795, yang berarti model memiliki performa yang cukup baik dalam mengklasifikasikan data, dengan probabilitas mendekati 80% untuk membedakan pasien yang berisiko dan tidak berisiko.

4. Analisis Fitur

Setelah model XGBoost selesai dilatih dan dievaluasi, dilakukan analisis lebih lanjut terhadap kontribusi masing-masing fitur dalam proses prediksi. Proses ini dikenal sebagai *feature importance analysis* dan sangat penting untuk mengetahui fitur-fitur mana saja yang paling berpengaruh dalam menentukan risiko penyakit jantung pada dataset yang digunakan.

Feature importance yang diterapkan dalam penelitian ini ditentukan berdasarkan nilai gain, yaitu seberapa besar suatu fitur meningkatkan keakuratan pemisahan antar kelas pada setiap pembentukan pohon keputusan. Semakin tinggi nilai gain suatu fitur, maka semakin besar pula kontribusinya terhadap prediksi akhir model.

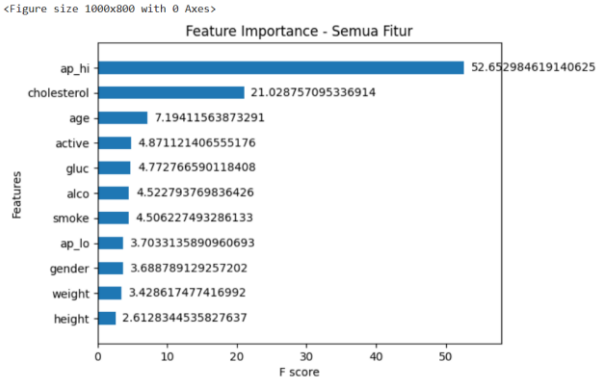
```
import matplotlib.pyplot as plt
from xgboost import plot_importance

# Tampilkan semua fitur, diurutkan berdasarkan gain
plt.figure(figsize=(10, 8))
plot_importance(model, importance_type='gain', title='Feature Importance - Semua Fitur', height=0.5)
plt.grid(False)
plt.show()
```

Sumber: Laporan Penelitian (2025)

Gambar 6 Kode untuk Menampilkan *Feature Importance* berdasarkan Gain.

Setelah kode tersebut dijalankan, model menghasilkan visualisasi tingkat kepentingan fitur yang dapat dilihat pada gambar di bawah.



Sumber: Laporan Penelitian(2025)

Gambar 7 Visualisasi *feature importance* berdasarkan nilai gain

Berdasarkan visualisasi pada gambar di atas, dapat dilihat bahwa beberapa fitur memiliki kontribusi yang lebih tinggi dibandingkan fitur lainnya. Fitur-fitur utama tersebut antara lain:

a. *ap_hi* (tekanan darah sistolik)
b. *age* (usia)
c. *cholesterol* (kadar kolesterol)
d. *weight* (berat badan)
e. *ap_lo* (tekanan darah diastolik)

Fitur *ap_hi* berada pada urutan teratas, menandakan bahwa tekanan darah sistolik merupakan indikator paling penting dalam menentukan risiko penyakit jantung pada dataset ini. Hal ini sesuai dengan literatur medis yang menyatakan bahwa tekanan darah tinggi merupakan salah satu faktor risiko utama penyakit kardiovaskular.

Kebaruan penelitian ini terletak pada implementasi dan optimasi algoritma XGBoost dalam memprediksi risiko penyakit jantung dengan menekankan pada analisis fitur-fitur yang memiliki pengaruh paling signifikan terhadap hasil prediksi. Pendekatan ini tidak hanya berorientasi pada peningkatan akurasi model, tetapi juga pada pemahaman yang lebih mendalam mengenai faktor-faktor yang berkontribusi terhadap munculnya risiko penyakit jantung. Hasil analisis menunjukkan bahwa fitur seperti *age*, *cholesterol*, *weight*, *ap_hi*, dan *ap_lo* memiliki peran penting dalam proses prediksi, yang mengindikasikan bahwa faktor usia, kadar kolesterol, serta tekanan darah, baik sistolik maupun diastolik, memegang peranan penting dalam menentukan tingkat risiko.

Dalam penerapannya, hasil penelitian ini berperan signifikan dalam mendukung pengembangan sistem pendukung keputusan (*decision support system*) di bidang kesehatan, khususnya dalam upaya deteksi dini dan pencegahan penyakit jantung. Dengan mengetahui faktor-faktor utama yang memengaruhi risiko, tenaga medis dapat merancang strategi pencegahan yang lebih tepat, seperti pemantauan tekanan darah dan kadar kolesterol secara berkala. Selain itu, hasil penelitian ini dapat dijadikan landasan untuk penelitian lanjutan yang menggunakan data pasien lokal, sehingga penerapan model prediksi berbasis *machine learning* dapat diadaptasikan secara lebih efektif dalam konteks klinis di Indonesia.

KESIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan, dapat disimpulkan bahwa algoritma XGBoost mampu digunakan secara efektif untuk memprediksi risiko penyakit jantung dengan performa yang cukup baik. Model yang dilatih menggunakan dataset *cardio_train.csv* dari Kaggle menunjukkan hasil evaluasi yang solid, dengan nilai *precision* sebesar 0,72, *recall* sebesar 0,77, *F1-score* sebesar 0,74, akurasi sebesar 73%, dan nilai AUC sebesar 0,795. Hasil tersebut menunjukkan bahwa model XGBoost memiliki kemampuan klasifikasi yang baik dalam membedakan antara pasien yang berisiko dan tidak berisiko terkena penyakit jantung. Selain itu, hasil analisis *feature importance* mengungkap bahwa tekanan darah sistolik (*ap_hi*), usia (*age*), dan kadar kolesterol

merupakan faktor yang paling dominan dalam memengaruhi hasil prediksi. Temuan ini menunjukkan bahwa XGBoost dapat menjadi salah satu metode yang potensial untuk diterapkan dalam pengembangan sistem pendukung keputusan (*decision support system*) guna mendeteksi dini risiko penyakit jantung secara otomatis dan akurat.

Untuk penelitian selanjutnya, disarankan agar dilakukan perbandingan antara algoritma XGBoost dengan algoritma lain seperti Random Forest, Support Vector Machine (SVM), atau Neural Network, sehingga dapat diperoleh hasil yang lebih komprehensif terkait performa berbagai metode dalam prediksi penyakit jantung. Selain itu, karena penelitian ini menggunakan dataset sekunder dari sumber terbuka, penelitian mendatang disarankan untuk memanfaatkan dataset primer yang bersumber dari data pasien lokal, misalnya dari rumah sakit di Indonesia, agar hasil model yang dikembangkan lebih relevan, kontekstual, dan dapat diterapkan secara nyata dalam lingkungan medis.

REFERENSI

Chandra, K., & Prasetyo, J. S. (2024). Prosiding SENAM 2024: Prediksi Penyakit Jantung Koroner Menggunakan Metode K-NN dan Regresi Logistik Berdasarkan Kerangka Kerja CRISP-DM. *Sistem Informasi & Informatika*, *4*, 241–248.

Dachi, J. M. A. S., & Sitompul, P. (2023). Application Of Game Theory In Determining Optimum Marketing Strategy In Marketplace. *JURNAL RISET RUMPUN MATEMATIKA DAN ILMU PENGETAHUAN ALAM*, *2*(2), 87–103. https://doi.org/10.55606/jurrimipa.v2i2.1336

Hoerul Anwar, A. (2025). SISTEMATIC REVIEW FAKTOR RESIKO PENYAKIT JANTUNG KORONER DI INDONESIA. *Https://Journal.Ymci.My.Id/Index.Php/Ijhri/Index*. https://journal.ymci.my.id/index.php/ijhri/index

Johson, A. (2023). *Heart Disease Datasets*. Kaggle. https://www.kaggle.com/datasets/albertjohson/hearth-disease-datasets

Muslihin, H. Y., Loita, A., & Nurjanah, D. S. (2022). Instrumen Penelitian Tindakan Kelas. In *Juni* (Vol. 6, Issue 1).

Rayadin, M. A., Musaruddin, M., Saputra, R. A., & Isnawaty, I. (2024). Implementasi Ensemble Learning Metode XGBoost dan Random Forest untuk Prediksi Waktu Penggantian Baterai Aki. *BIOS: Jurnal Teknologi Informasi Dan Rekayasa Komputer*, *5*(2), 111–119. https://doi.org/10.37148/bios.v5i2.128

Soleha, Kusumajaya, H., & Maryana. (2025). *FAKTOR-FAKTOR YANG BERHUBUNGAN DENGAN KUALITAS HIDUP PADA PASIEN CHF*. http://jurnal.globalhealthsciencegroup.com/index.php/JPPP

Susanti, N., Zahara, A., Fadillah Darus, N., & Zulaila. (2024). FAKTOR RISIKO YANG BERHUBUNGAN DENGAN PENYAKIT JANTUNG KORONER: LITERATUR RIVIEW. *Https://Journal.Universitaspahlawan.Ac.Id/Index.Php/Jkt/Article/View/28417*, *5*(2).

Syahreza, A., Ningrum, N. K., & Syahrazy, M. A. (2024). Perbandingan Kinerja Model Prediksi Cuaca: Random Forest, Support Vector Regression, dan XGBoost. *Edumatic: Jurnal Pendidikan Informatika*, *8*(2), 526–534. https://doi.org/10.29408/edumatic.v8i2.27640

Yulianti, S. E. H., Soesanto, O., & Sukmawaty, Y. (2022). Penerapan Metode Extreme Gradient Boosting (XGBOOST) pada Klasifikasi Nasabah Kartu Kredit. *JOMTA Journal of Mathematics: Theory and Applications*, *4*(1).