

**Implementasi Algoritma CART (Classification and Regression Tree)
untuk Klasifikasi Risiko Penyakit Jantung Menggunakan Dataset Heart
Disease**

Dosen Fakhri Khusnu Reza Mahfud,M.Kom



Oleh:

Nama : Muhammad Multazim
NIM : 220607110060
Kelas : PSI A
Tanggal : 15 April 2025

JURUSAN PERPUSTAKAAN DAN ILMU INFORMASI
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM
MALANG
2024

BAB I

PENDAHULUAN

1.1 Latar Belakang

Penyakit jantung merupakan salah satu penyebab kematian utama di dunia, termasuk di Indonesia. Menurut data dari World Health Organization (WHO), penyakit jantung berkontribusi besar terhadap angka kematian global setiap tahunnya. Data menunjukkan bahwa sekitar 17,9 juta orang meninggal akibat penyakit jantung setiap tahun, menjadikannya sebagai salah satu masalah kesehatan masyarakat yang paling mendesak. Di Indonesia, prevalensi penyakit jantung terus meningkat, dan ini menjadi tantangan besar bagi sistem kesehatan.

Dalam upaya mitigasi risiko, teknologi informasi memainkan peran penting dalam mendeteksi dan menganalisis faktor-faktor yang mempengaruhi penyakit ini. Dengan perkembangan teknologi big data dan machine learning, analisis dataset medis untuk klasifikasi penyakit menjadi semakin memungkinkan dan efektif. Pendekatan ini memungkinkan para peneliti dan profesional kesehatan untuk menggali informasi yang lebih mendalam dari data pasien, sehingga dapat diidentifikasi pola-pola yang mungkin tidak terlihat melalui metode analisis tradisional.

Algoritma klasifikasi seperti Decision Tree telah banyak digunakan dalam bidang kesehatan untuk menganalisis faktor risiko dan memprediksi kemungkinan terjadinya penyakit. Secara khusus, algoritma CART (Classification and Regression Tree) menawarkan metode klasifikasi yang sederhana, interpretatif, dan efektif untuk dataset berskala kecil hingga menengah. Algoritma ini memungkinkan visualisasi keputusan yang jelas, sehingga memudahkan dokter dalam memahami dan menjelaskan hasil analisis kepada pasien.

Penelitian ini bertujuan untuk menerapkan algoritma CART pada dataset penyakit jantung "Heart Disease" guna membangun model klasifikasi yang dapat membantu dalam prediksi awal penyakit jantung. Dengan menggunakan data yang relevan, penelitian ini diharapkan dapat memberikan kontribusi signifikan terhadap pengembangan sistem diagnosis yang lebih akurat dan efisien, sehingga dapat meningkatkan tingkat keberhasilan pengobatan dan mengurangi angka kematian akibat penyakit jantung.

1.2 Identifikasi Masalah

- Bagaimana membangun model klasifikasi penyakit jantung menggunakan algoritma CART?
- Seberapa akurat model CART dalam mengklasifikasikan risiko penyakit jantung?
- Fitur-fitur apa yang paling berpengaruh dalam klasifikasi penyakit jantung berdasarkan model CART?

1.3 Tujuan Penelitian

- Menerapkan algoritma CART untuk klasifikasi penyakit jantung.
- Mengukur performa klasifikasi model CART pada dataset Heart Disease.
- Mengidentifikasi fitur-fitur penting yang berpengaruh terhadap klasifikasi penyakit jantung.

1.4 Hipotesis

Dalam penelitian ini, hipotesis yang diajukan bertujuan untuk menguji efektivitas algoritma CART dalam klasifikasi penyakit jantung. Berdasarkan latar belakang dan tujuan penelitian yang telah dijelaskan, hipotesis dibagi menjadi dua bagian:

- Hipotesis Nol (H_0)
 H_0 : Algoritma CART tidak mampu mengklasifikasikan penyakit jantung dengan tingkat akurasi yang baik.

Penjelasan: Hipotesis nol ini berfungsi sebagai titik awal untuk analisis. Jika H_0 diterima, maka dapat disimpulkan bahwa algoritma CART tidak memberikan hasil yang memadai dalam mengklasifikasikan risiko penyakit jantung. Hal ini dapat terjadi jika model tidak dapat menangkap pola-pola yang relevan dalam data atau jika fitur yang digunakan tidak cukup informatif. Penelitian ini akan mengukur akurasi model CART dan membandingkannya dengan nilai ambang batas yang telah ditentukan untuk menentukan apakah model tersebut benar-benar tidak efisien.

- **Hipotesis Alternatif (H_1)**

H_1 : Algoritma CART mampu mengklasifikasikan penyakit jantung dengan tingkat akurasi yang baik.

Penjelasan: Hipotesis alternatif ini menunjukkan harapan bahwa algoritma CART dapat memberikan hasil yang positif dalam klasifikasi penyakit jantung. Dengan mengukur performa model melalui metrik seperti akurasi, precision, recall, dan F-score, penelitian ini bertujuan untuk menunjukkan bahwa CART tidak hanya dapat membedakan antara pasien yang berisiko dan tidak berisiko, tetapi juga dapat memberikan model yang dapat diandalkan untuk diagnosis awal. Jika H_1 diterima, maka hasil penelitian akan berkontribusi pada pengembangan alat bantu diagnosa yang lebih baik dalam bidang kesehatan.

- **Signifikansi Hipotesis**

Pentingnya kedua hipotesis ini terletak pada kemampuan untuk memberikan bukti empiris mengenai efektivitas algoritma CART. Dengan menguji kedua hipotesis ini, penelitian ini tidak hanya bertujuan untuk membangun model klasifikasi, tetapi juga untuk memberikan wawasan tentang bagaimana algoritma machine learning dapat diterapkan untuk meningkatkan deteksi dini penyakit jantung, sehingga berdampak positif pada pengelolaan kesehatan masyarakat secara keseluruhan.

1.5 Manfaat Penelitian

- Memberikan kontribusi pada bidang data science medis dalam penerapan machine learning untuk diagnosis penyakit.
- Menyediakan model klasifikasi sederhana namun efektif yang dapat digunakan untuk deteksi dini penyakit jantung.
- Menjadi referensi bagi penelitian lanjutan terkait penerapan algoritma machine learning di bidang kesehatan

1.6 Batasan Masalah

- Data yang digunakan adalah dataset "Heart Disease" yang tersedia secara publik.
- Penelitian ini hanya menggunakan algoritma CART sebagai metode klasifikasi.
- Fokus penelitian terbatas pada evaluasi performa akurasi klasifikasi.

1.7 Sistematika Penulisan

- BAB I Pendahuluan
- BAB II Tinjauan Pustaka
- BAB III Metode Penelitian
- BAB IV Hasil dan Pembahasan
- BAB V Penutup

BAB II

PENDAHULUAN

2.1 Penelitian Terdahulu

Penelitian sebelumnya oleh (Elshewey et al., 2025) menunjukkan bahwa kombinasi algoritma Greylag Goose Optimization (GGO) dan Long Short-Term Memory (LSTM) mampu meningkatkan akurasi klasifikasi penyakit jantung hingga mencapai 99,58%. Penelitian ini membuktikan bahwa pemilihan fitur menggunakan algoritma optimasi metaheuristik seperti GGO secara signifikan dapat meningkatkan performa model prediksi dibandingkan dengan metode optimasi lain.

Penelitian lain oleh (Reddy et al., 2021) mengembangkan sistem prediksi risiko penyakit jantung dengan menggunakan berbagai algoritma machine learning seperti Sequential Minimal Optimization (SMO), Random Forest, dan K-Nearest Neighbor (KNN). Mereka membandingkan performa algoritma berdasarkan dataset Cleveland Heart Disease dan menemukan bahwa tuning hyperparameter (seperti jumlah tetangga pada KNN) dapat meningkatkan akurasi prediksi hingga 86,47%

Selain itu, penelitian oleh (Ozcan & Peker, 2023) membandingkan beberapa metode klasifikasi dalam diagnosis penyakit jantung, dan menemukan bahwa algoritma Decision Tree tetap memiliki keunggulan dalam hal interpretabilitas, walaupun beberapa metode lain menunjukkan akurasi yang kompetitif. Penelitian ini menggarisbawahi pentingnya keseimbangan antara akurasi dan keterbacaan model dalam konteks aplikasi klinis

2.2 Landasan Teori

a. Penyakit Jantung

Penyakit jantung merupakan sekelompok kondisi yang memengaruhi struktur dan fungsi jantung serta pembuluh darah, termasuk di dalamnya penyakit arteri koroner, gagal jantung, dan gangguan ritme jantung seperti aritmia. Menurut (Elshewey et al., 2025), penyakit jantung, terutama penyakit arteri koroner, disebabkan oleh penyempitan atau penyumbatan arteri yang menghambat aliran darah ke otot jantung, sehingga meningkatkan risiko serangan jantung. Penyakit ini seringkali baru terdeteksi setelah terjadinya peristiwa akut, karena gejalanya bisa samar atau menyerupai kondisi lain, khususnya pada populasi lansia. World Health Organization (WHO) mencatat bahwa penyakit jantung tetap menjadi penyebab utama kematian global, dengan sekitar 17,9 juta kematian setiap tahun.

b. Machine Learning

Machine Learning (ML) adalah cabang dari kecerdasan buatan yang memungkinkan sistem komputer belajar dari data untuk membuat prediksi atau keputusan tanpa diprogram secara eksplisit. Dalam konteks prediksi penyakit jantung, machine learning memanfaatkan data klinis seperti usia, tekanan darah, kadar kolesterol, dan hasil tes laboratorium untuk mengenali pola-pola yang mengindikasikan risiko penyakit. (Reddy et al., 2021) menjelaskan bahwa berbagai algoritma machine learning, termasuk klasifikasi dan regresi, dapat meningkatkan akurasi diagnosis dini dengan memanfaatkan teknik evaluasi atribut dan tuning hyperparameter.

c. Decision Tree

Decision Tree adalah model prediktif berbentuk pohon yang digunakan untuk memetakan fitur-fitur input ke dalam label target. Setiap node dalam pohon merepresentasikan sebuah fitur, sedangkan cabang merepresentasikan keputusan berdasarkan nilai fitur tersebut, hingga akhirnya mencapai daun yang menunjukkan hasil klasifikasi. (Ozcan & Peker, 2023) menunjukkan bahwa Decision Tree memiliki keunggulan dalam interpretabilitas dibandingkan metode lain, karena struktur pohonnya memudahkan pemahaman proses pengambilan keputusan oleh tenaga medis

d. CART (Classification And Regression Tree)

Classification and Regression Tree (CART) adalah varian algoritma Decision Tree yang membangun pohon biner berdasarkan kriteria Gini Impurity untuk tugas klasifikasi. Dalam CART, setiap split dibuat dengan tujuan meminimalkan impuritas antar kelompok, sehingga menghasilkan pohon yang optimal untuk prediksi klasifikasi atau regresi. (Ozcan & Peker, 2023) dalam studinya menunjukkan bahwa penggunaan CART dalam prediksi penyakit jantung dapat menghasilkan performa klasifikasi yang tinggi serta memberikan struktur pohon yang sederhana dan efektif.

BAB III

METODE PENELITIAN

3.1 Jenis Penelitian

Penelitian ini merupakan penelitian kuantitatif dengan pendekatan eksperimen. Penelitian kuantitatif dipilih karena fokus utamanya adalah pada analisis numerik untuk membangun model klasifikasi penyakit jantung dan mengevaluasi kinerjanya menggunakan data empiris. Pendekatan eksperimen digunakan untuk menguji efektivitas algoritma Classification and Regression Tree (CART) dalam mengklasifikasikan risiko penyakit jantung berdasarkan sekumpulan data klinis. Algoritma CART dipilih karena memiliki karakteristik sebagai metode klasifikasi yang mampu menghasilkan model yang interpretatif dan mudah dipahami, sehingga sesuai dengan kebutuhan diagnosis medis. Penelitian ini berusaha untuk tidak hanya menilai akurasi model, tetapi juga memahami struktur keputusan yang dihasilkan melalui model pohon keputusan.

3.2 Tempat dan Waktu Penelitian

Penelitian ini dilakukan secara daring (online) tanpa keterikatan pada lokasi fisik tertentu, menggunakan perangkat komputer pribadi yang memenuhi spesifikasi minimum untuk pemrosesan data. Seluruh rangkaian kegiatan penelitian, mulai dari tahap pengumpulan data, preprocessing, implementasi model, hingga evaluasi hasil, dilaksanakan pada bulan Mei 2025. Pemilihan pelaksanaan daring didasarkan pada fleksibilitas akses terhadap sumber daya, kemudahan dalam penggunaan perangkat lunak open-source seperti Python dan Scikit-learn, serta untuk mengoptimalkan waktu pengerjaan.

3.3 Subjek dan Objek Penelitian

a. Subjek Penelitian

Subjek dalam penelitian ini adalah data rekam medis pasien yang direpresentasikan dalam bentuk dataset publik, yaitu Heart Disease Dataset. Dataset ini berisi atribut-atribut penting seperti usia, jenis kelamin, tekanan darah, kadar kolesterol, kadar gula darah, hasil elektrokardiogram, dan parameter medis lainnya yang relevan dalam prediksi risiko penyakit jantung. Penggunaan data rekam medis sebagai subjek penelitian sejalan dengan studi sebelumnya, di mana data klinis pasien digunakan untuk membangun dan mengevaluasi berbagai model machine learning.

b. Objek Penelitian

Objek penelitian ini adalah proses klasifikasi risiko penyakit jantung menggunakan algoritma Classification and Regression Tree (CART). Objek ini mencakup keseluruhan tahapan dari preprocessing data, pembuatan model klasifikasi, hingga evaluasi hasil prediksi.

3.4 Sumber Data

Data yang digunakan dalam penelitian ini diambil dari UCI Machine Learning Repository, yang merupakan repository data publik berstandar internasional dan sering dijadikan sumber data dalam berbagai penelitian machine learning, khususnya di bidang kesehatan. Dataset yang digunakan adalah Cleveland Heart Disease Dataset, yang secara luas telah divalidasi dan digunakan sebagai benchmark untuk pengembangan model prediksi penyakit jantung. Dataset ini terdiri dari 303 sampel dengan 14 atribut utama yang meliputi karakteristik demografis, data medis, dan hasil pemeriksaan klinis. Data ini sangat cocok untuk penelitian ini karena telah terbukti relevan dan representatif untuk studi prediksi penyakit jantung.

3.5 Instrumen Penelitian

Instrumen penelitian yang digunakan dalam penelitian ini meliputi:

a. Perangkat Keras:

- Laptop/PC dengan spesifikasi minimal: prosesor Intel Core i5, RAM 8 GB, penyimpanan 256 GB SSD.

b. Perangkat Lunak:

- Python: bahasa pemrograman utama untuk analisis data.

- Scikit-learn: library Python untuk membangun model CART.
- Pandas: untuk pengolahan, transformasi, dan analisis data tabular.
- Matplotlib: untuk visualisasi hasil klasifikasi dan analisis data.
- c. Dataset:**
 - Heart Disease Dataset dari UCI Machine Learning Repository.

Kombinasi instrumen ini memberikan dukungan penuh terhadap kebutuhan penelitian berbasis data dan pemodelan machine learning sebagaimana yang diterapkan dalam penelitian sebelumnya.

3.6 Teknik Pengumpulan Data

Pengumpulan data dilakukan melalui beberapa tahapan:

- a. Mengunduh Dataset:**
 - Dataset Heart Disease diunduh dari repository publik UCI, memastikan data yang digunakan adalah sumber terbuka dan valid.
- b. Preprocessing Data:**
 - Pembersihan Data: Memastikan tidak ada nilai kosong (missing value). Jika ditemukan, data diperlakukan dengan penghapusan atau imputasi nilai.
 - Normalisasi/Standardisasi: Jika terdapat ketimpangan skala pada data numerik (misal tekanan darah vs kolesterol), maka dilakukan normalisasi untuk meningkatkan performa model klasifikasi.
 - Transformasi Data Kategorikal: Data kategorikal seperti jenis kelamin atau hasil EKG dikodekan secara numerik menggunakan teknik label encoding atau one-hot encoding.

Tahapan ini bertujuan untuk memastikan bahwa data dalam kondisi optimal sebelum digunakan dalam pembangunan model klasifikasi, mengingat kualitas preprocessing berpengaruh besar terhadap performa model.

3.7 Teknik Analisis Data

Analisis data dalam penelitian ini dilakukan melalui tahapan berikut:

- a. Membagi Dataset:**
 - Dataset dibagi menjadi data latih (training set) dan data uji (testing set) dengan rasio umum 80:20.
 - Pembagian ini bertujuan untuk membangun model menggunakan sebagian data dan menguji kinerjanya pada data yang tidak pernah dilihat sebelumnya.
- b. Pembangunan Model CART:**
 - Model klasifikasi dibangun menggunakan algoritma Classification and Regression Tree (CART) dari library Scikit-learn.
 - Kriteria split menggunakan Gini Impurity untuk menentukan pemisahan data terbaik di setiap node.
 - Model dibangun secara rekursif hingga semua node bersifat homogen atau hingga batasan tertentu (seperti kedalaman maksimum pohon) tercapai.
- c. Training dan Testing:**
 - Model dilatih pada data latih untuk mengidentifikasi pola dari atribut klinis yang mempengaruhi risiko penyakit jantung.
 - Model kemudian diuji menggunakan data uji untuk mengukur kemampuannya dalam melakukan generalisasi prediksi pada data baru.
- d. Evaluasi Model:**
 - Performa model dievaluasi menggunakan beberapa metrik:
 - Akurasi: proporsi prediksi benar dari keseluruhan prediksi.
 - Precision: sejauh mana prediksi positif benar-benar positif.
 - Recall (Sensitivity): sejauh mana model mampu menangkap seluruh kasus positif.

- Confusion Matrix: untuk melihat distribusi prediksi benar dan salah dalam bentuk tabel.
- Evaluasi ini memberikan gambaran menyeluruh tentang kekuatan dan kelemahan model dalam klasifikasi.
- e. Interpretasi Model:
 - Pohon keputusan yang dihasilkan dianalisis untuk mengidentifikasi fitur-fitur kunci (seperti usia, tekanan darah, kolesterol) yang paling berkontribusi terhadap prediksi .
 - Interpretasi ini penting untuk mendukung keputusan klinis berbasis data serta meningkatkan kepercayaan tenaga medis terhadap penggunaan machine learning dalam diagnosis.

DAFTAR PUSTAKA

- Elshewey, A. M., Abed, A. H., Khafaga, D. S., Alhussan, A. A., Eid, M. M., & El-Kenawy, E. S. M. (2025). Enhancing heart disease classification based on greylag goose optimization algorithm and long short-term memory. *Scientific Reports*, *15*(1), 1277. <https://doi.org/10.1038/s41598-024-83592-0>
- Ozcan, M., & Peker, S. (2023). A classification and regression tree algorithm for heart disease modeling and prediction. *Healthcare Analytics*. <https://doi.org/10.1016/j.health.2022.100130>
- Reddy, K. V. V., Elamvazuthi, I., Aziz, A. A., Paramasivam, S., Chua, H. N., & Pranavanand, S. (2021). Heart disease risk prediction using machine learning classifiers with attribute evaluators. *Applied Sciences (Switzerland)*, *11*(18). <https://doi.org/10.3390/app11188352>