LAPORAN

IFUWP4339-B KECERDASAN BUATAN

PREDIKSI PENYAKIT JANTUNG MENGGUNAKAN PERBANDINGAN ALGORITMA MACHINE LEARNING DECISION TREE, SUPPORT VECTOR MACHINE, DAN K-NEAST NEIGHBORS

Dosen Pengampu :

Leni Fitriani, S.T, M.Kom.

NIDN : 0429058704

Disusun oleh :

Siti Rahmawati (2306146)

Fadhlian Nur Fajri (2306147)

PROGRAM STUDI TEKNIK INFORMATIKA

JURUSAN ILMU KOMPUTER

INSTITUT TEKNOLOGI GARUT

2025

1. PREDIKSI PENYAKIT JANTUNG MENGGUNAKAN PERBANDINGAN ALGORITMA MACHINE LEARNING DECISION TREE, SUPPORT VECTOR MACHINE, DAN K-NEAST NEIGHBORS

Diagnosis penyakit jantung sering kali memerlukan pemeriksaan yang kompleks dan mahal seperti angiografi, yang tidak selalu tersedia di fasilitas kesehatan primer, terutama di negara berkembang. Hal ini menyebabkan keterlambatan diagnosis dan penanganan yang dapat berakibat fatal bagi pasien. Oleh karena itu, diperlukan sistem prediksi penyakit jantung yang dapat membantu tenaga medis dalam melakukan diagnosis dini dengan akurasi tinggi dan biaya yang terjangkau.

Dalam beberapa tahun terakhir, machine learning (ML) telah menunjukkan potensi yang sangat menjanjikan dalam prediksi dan diagnosis penyakit jantung. Dalam tinjauan komprehensif mereka menyatakan bahwa teknik machine learning menawarkan jalan yang menjanjikan untuk memprediksi dan mendiagnosis penyakit jantung melalui analisis data klinis standar (Hajiarbabi, 2024). Studi meta-analisis yang dilakukan pada 344 penelitian menunjukkan bahwa algoritma boosting mencapai area under the curve (AUC) sebesar 0,88 (95% CI 0,84-0,91) untuk penyakit arteri koroner, sementara algoritma custom-built mencapai AUC sebesar 0,93 (95% CI 0,85-0,97).

Teknik inovatif machine learning yang mampu mengklasifikasikan berbagai jenis penyakit jantung dengan akurasi tinggi menggunakan dataset besar yang dapat diakses publik (Rahman, 2023). Algoritma decision tree mencapai akurasi tertinggi dalam prediksi penyakit jantung dengan tingkat 93,19%, diikuti oleh algoritma SVM dengan 92,30%. Sementara itu, Jellyfish Optimization Algorithm menunjukkan bahwa Random Forest mencapai akurasi 91,80% dalam prediksi penyakit jantung (Mohan, 2020).

Berbagai algoritma machine learning telah digunakan dalam penelitian prediksi penyakit jantung, termasuk logistic regression, decision trees, XGBoost, gradient boosting, random forests, support vector machines (SVMs), dan artificial neural networks (ANNs). Bahwa sistem terkomputerisasi telah muncul sebagai alternatif yang menjanjikan untuk metode tradisional, menawarkan prediksi risiko penyakit jantung yang lebih cepat (Alshraideh, 2024).

Implementasi machine learning dalam domain medis, khususnya untuk prediksi penyakit jantung, memiliki potensi untuk meningkatkan deteksi dini dan diagnosis, sehingga dapat mengatasi tantangan yang ditimbulkan oleh penyakit jantung (Khan, 2023). Dengan memanfaatkan data klinis pasien, algoritma machine learning dapat mengidentifikasi pola-pola kompleks yang menunjukkan risiko penyakit jantung, memberikan dukungan keputusan klinis yang objektif, dan pada akhirnya dapat menyelamatkan lebih banyak nyawa melalui intervensi dini.

2. BUSINESS UNDERSTANDING

2.1 Permasalahan dunia nyata

Penyakit jantung merupakan salah satu penyebab kematian utama di dunia. Menurut World Health Organization (WHO), penyakit jantung menyebabkan sekitar 17,9 juta kematian setiap tahunnya. Di Indonesia, penyakit jantung koroner merupakan penyebab kematian tertinggi dengan prevalensi yang terus meningkat setiap tahun.

Diagnosis penyakit jantung sering kali memerlukan pemeriksaan yang kompleks dan mahal seperti angiografi, yang tidak selalu tersedia di fasilitas kesehatan primer. Hal ini menyebabkan keterlambatan diagnosis dan penanganan yang dapat berakibat fatal bagi pasien.

2.2 Tujuan proyek

Proyek ini bertujuan untuk mengembangkan sistem prediksi penyakit jantung menggunakan machine learning dengan membandingkan performa tiga algoritma klasifikasi yaitu Decision Tree, Support Vector Machine (SVM), dan K-Nearest Neighbors (KNN). Melalui perbandingan tersebut, proyek ini akan mengidentifikasi algoritma terbaik untuk prediksi penyakit jantung dan memberikan rekomendasi implementasi AI dalam diagnosis medis.

2.3 User atau Pengguna Sistem

Sistem prediksi penyakit jantung ini dirancang untuk digunakan oleh berbagai stakeholder dalam dunia kesehatan. Dokter Spesialis Jantung akan menggunakan sistem ini untuk mendukung diagnosis dan pengambilan keputusan medis yang lebih akurat. Dokter Umum di Puskesmas dapat memanfaatkannya sebagai screening awal sebelum melakukan rujukan ke fasilitas kesehatan yang lebih lengkap. Tenaga Medis di Rumah Sakit akan menggunakannya untuk triase pasien dengan gejala kardiovaskular, sementara Sistem Informasi Rumah Sakit dapat mengintegrasikan tool ini dengan rekam medis elektronik yang sudah ada.

2.4 Manfaat implementasi AI

Implementasi AI dalam sistem prediksi penyakit jantung memberikan berbagai manfaat signifikan bagi dunia kesehatan. Sistem ini memungkinkan deteksi dini dengan mengidentifikasi risiko penyakit jantung sebelum gejala parah muncul, sehingga memberikan kesempatan untuk intervensi yang lebih baik. Dari segi ekonomi, implementasi AI dapat meningkatkan efisiensi biaya dengan mengurangi kebutuhan pemeriksaan diagnostik yang mahal dan kompleks. Sistem ini juga meningkatkan aksesibilitas layanan kesehatan dengan memungkinkan screening di fasilitas kesehatan yang memiliki sumber daya terbatas. Selain itu, AI dapat mendukung akurasi diagnosis dengan menyediakan analisis data objektif yang

membantu pengambilan keputusan medis. Yang tidak kalah penting, sistem ini berkontribusi pada aspek pencegahan dengan memberikan peringatan dini untuk mencegah komplikasi yang lebih serius.

3. DATA UNDERSTANDING

3.1 Sumber Data

Dataset yang digunakan dalam penelitian ini diperoleh dari Kaggle dengan judul "UCI Heart Disease Data" yang merupakan subset dari UCI Machine Learning Repository. Dataset asli memiliki 76 atribut, namun sebagian besar penelitian machine learning menggunakan subset yang terdiri dari 14 atribut yang paling relevan. Data ini berasal dari Cleveland Clinic Foundation dan telah menjadi benchmark standar untuk penelitian prediksi penyakit jantung dalam komunitas machine learning. Dataset tersedia secara publik dan telah digunakan secara luas dalam penelitian akademis untuk pengembangan algoritma prediksi penyakit jantung. Sumber: https://www.kaggle.com/datasets/redwankarimsony/heart-disease-data

3.2 Deskripsi setiap fitur

Dataset ini terdiri dari 14 atribut yang mencakup informasi demografis, gejala klinis, dan hasil pemeriksaan medis pasien. Atribut age merepresentasikan usia pasien dalam tahun dengan rentang nilai numerik. Sex menunjukkan jenis kelamin dengan nilai 1 untuk laki-laki dan 0 untuk perempuan. Chest pain type (cp) mengkategorikan jenis nyeri dada dengan empat kategori yaitu 0 untuk asymptomatic, 1 untuk typical angina, 2 untuk atypical angina, dan 3 untuk non-anginal pain.

Resting blood pressure (trestbps) menunjukkan tekanan darah sistolik saat istirahat dalam mmHg, sedangkan serum cholesterol (chol) mengukur kadar kolesterol serum dalam mg/dl. Fasting blood sugar (fbs) merupakan indikator apakah gula darah puasa lebih dari 120 mg/dl dengan nilai 1 untuk benar dan 0 untuk salah. Resting electrocardiographic results (restecg) menunjukkan hasil EKG saat istirahat dengan nilai 0 untuk normal, 1 untuk ST-T wave abnormality, dan 2 untuk left ventricular hypertrophy.

Maximum heart rate achieved (thalach) mengukur detak jantung maksimum yang dicapai dalam angka numerik, sementara exercise induced angina (exang) menunjukkan apakah terjadi angina akibat olahraga dengan nilai 1 untuk ya dan 0 untuk tidak. ST depression induced by exercise (oldpeak) mengukur depresi segmen ST yang diinduksi oleh olahraga relatif terhadap istirahat dalam nilai numerik. Slope of the peak exercise ST segment (slope) menunjukkan kemiringan segmen ST saat puncak olahraga dengan nilai 0 untuk upsloping, 1 untuk flat, dan 2 untuk downsloping.

Number of major vessels colored by fluoroscopy (ca) menunjukkan jumlah pembuluh darah utama yang terlihat melalui fluoroskopi dengan rentang nilai 0-3. Thalassemia (thal) menunjukkan jenis thalassemia dengan nilai 3 untuk normal, 6 untuk fixed defect, dan 7 untuk reversible defect.

No	Nama Fitur	Deskripsi	Tipe Data	Rentang Nilai
1	Age	Usia pasien dalam tahun	Numerik	29-77
2	Sex	Jenis kelamin (1=Laki-laki, 0=Perempuan)	Kategorikal	0, 1
3	CP	Tipe nyeri dada (0=Typical Angina, 1=Atypical Angina, 2=Non-anginal Pain, 3=Asymptomatic)	Kategorikal	0-3
4	Trestbps	Tekanan darah istirahat (mmHg)	Numerik	94-200
5	Chol	Kolesterol serum (mg/dl)	Numerik	126-564
6	FBS	Gula darah puasa > 120 mg/dl (1=True, 0=False)	Kategorikal	0, 1
7	Restecg	Hasil EKG istirahat (0=Normal, 1=ST-T wave abnormality, 2=Left ventricular hypertrophy)	Kategorikal	0-2
8	Thalach	Detak jantung maksimum yang dicapai	Numerik	71-202
9	Exang	Angina akibat olahraga (1=Yes, 0=No)	Kategorikal	0, 1
10	Oldpeak	Depresi ST akibat olahraga relatif terhadap istirahat	Numerik	0-6.2
11	Slope	Kemiringan segmen ST puncak olahraga (0=Upsloping, 1=Flat, 2=Downsloping)	Kategorikal	0-2
12	CA	Jumlah pembuluh darah utama yang diwarnai fluoroskopi	Numerik	0-4
13	Thal	Status thalassemia (1=Normal, 2=Fixed defect, 3=Reversible defect)	Kategorikal	3-Jan

Gambar 1. Deskripsi Fitur (Atribut)

3.3 Ukuran dan Format Data

Dataset memiliki total 303 record atau baris data dengan 14 kolom yang terdiri dari 13 fitur input dan 1 target output. Data disimpan dalam format CSV (Comma Separated Values) dengan ukuran file yang relatif kecil sekitar 15-20 KB. Setiap baris merepresentasikan satu pasien dengan informasi lengkap mengenai kondisi kesehatan jantungnya. Dataset ini telah dibersihkan dan tidak mengandung missing values yang signifikan, sehingga dapat langsung digunakan untuk proses machine learning dengan preprocessing minimal.

3.4 Tipe Data dan Target Klasifikasi

Dataset ini mengandung campuran tipe data numerik dan kategorikal. Fitur numerik kontinu meliputi age, trestbps, chol, thalach, dan oldpeak yang memiliki rentang nilai yang bervariasi. Fitur kategorikal meliputi sex, cp, fbs, restecg, exang, slope, ca, dan thal yang merupakan nilai diskret dengan kategori tertentu. Target klasifikasi adalah variabel "goal" yang menunjukkan keberadaan penyakit jantung dengan nilai integer dari 0 (tidak ada penyakit) hingga 4, namun dalam sebagian besar implementasi dikonversi menjadi binary classification dengan 0 untuk tidak ada penyakit jantung dan 1 untuk ada penyakit jantung. Hal ini menjadikan permasalahan sebagai binary classification problem dimana model machine learning akan memprediksi apakah seorang pasien memiliki penyakit jantung atau tidak berdasarkan 13 fitur input yang tersedia.

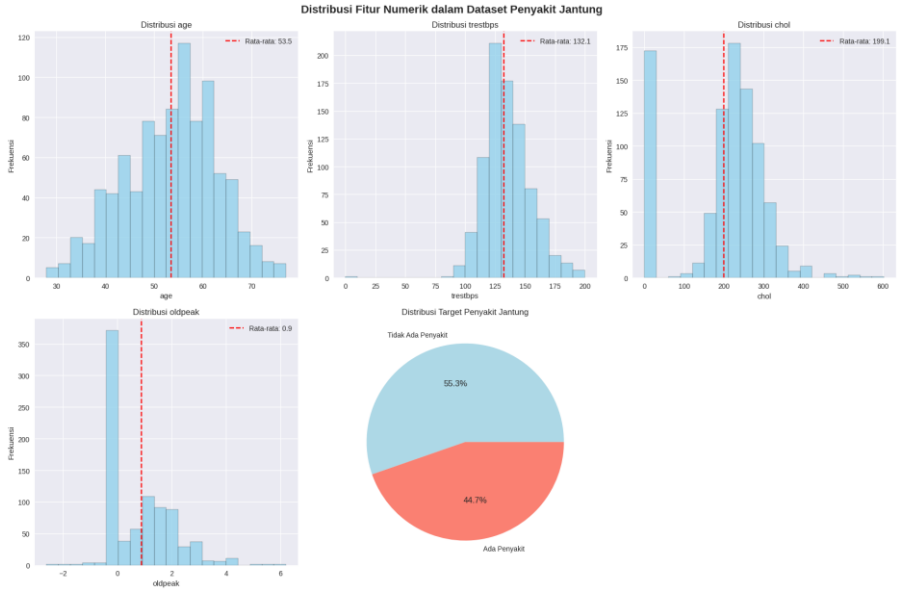
4. EXPLORATORY DATA ANALYSIS (EDA)

4.1 Visualisasi Distribusi Data

Distribusi fitur numerik menunjukkan bahwa:

- Age: Terdistribusi normal dengan rentang 29-77 tahun, mayoritas pasien berusia 50-60 tahun
- Resting BP: Sebagian besar pasien memiliki tekanan darah antara 120-140 mmHg
- Cholesterol: Distribusi cenderung right-skewed dengan beberapa outlier pada nilai tinggi
- Max Heart Rate: Terdistribusi relatif merata dengan rentang 71-202 bpm

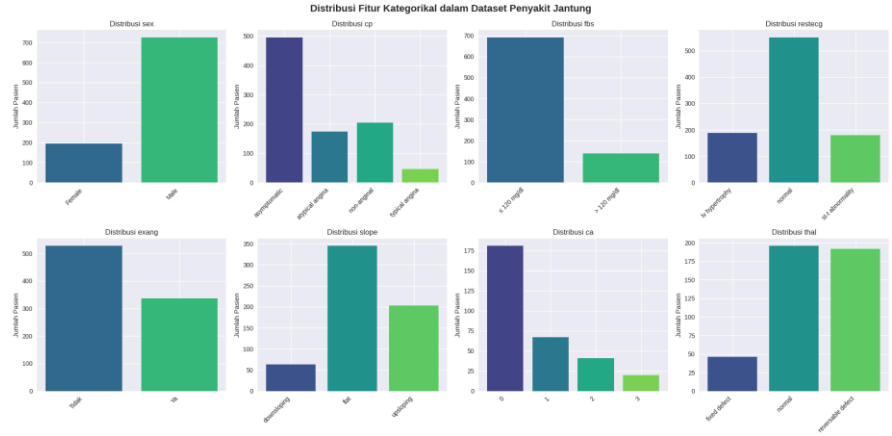
- Oldpeak: Mayoritas nilai mendekati 0, menunjukkan sebagian besar pasien tidak mengalami ST depression signifikan



Gambar 2. Distribusi fitur numerik dalam dataset penyakit jantung

Distribusi fitur kategorikal:

- Sex: Proporsi laki-laki lebih besar (sekitar 68%) dibanding Perempuan
- Chest Pain: Tipe 0 (asymptomatic) paling dominan
- Exercise Angina: Lebih banyak pasien yang tidak mengalami angina saat berolahraga
- Slope: Mayoritas pasien memiliki slope upsloping (tipe 2)

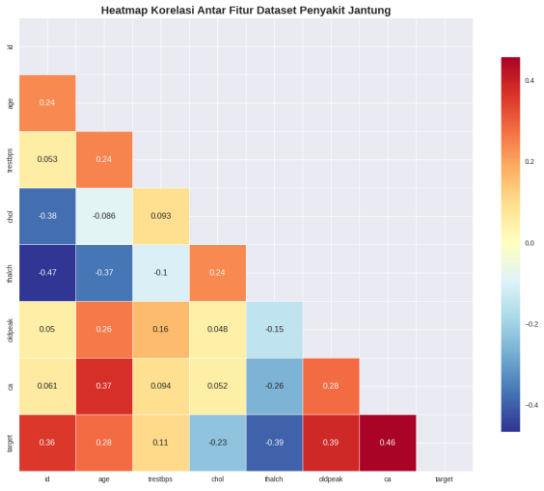


Gambar 3. Distribusi fitur kategorikal dalam dataset penyakit jantung

4.2 Analisis korelasi antar fitur

Heatmap korelasi yang mengungkapkan:

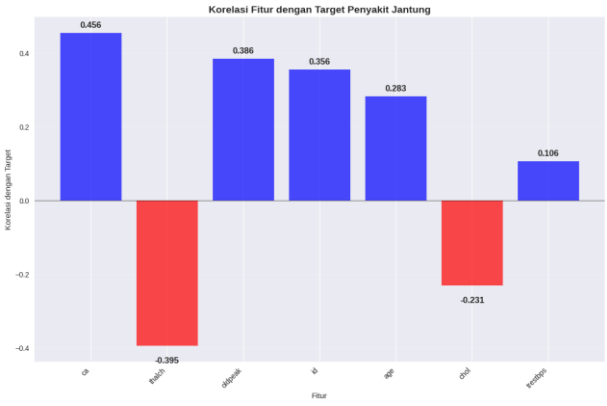
- Korelasi positif kuat antara target dengan cp, thalach, dan slope
- Korelasi negatif dengan exang, oldpeak, dan ca
- Tidak ada multikolinearitas yang signifikan antar fitur independen



Gambar 4. Heatmap korelasi antar fitur dataset penyakit jantung

Korelasi fitur dengan target:

- Chest Pain Type (cp): Korelasi positif tertinggi
- Exercise Angina (exang): Korelasi negatif kuat
- Oldpeak: Korelasi negatif
- Slope: Korelasi positif

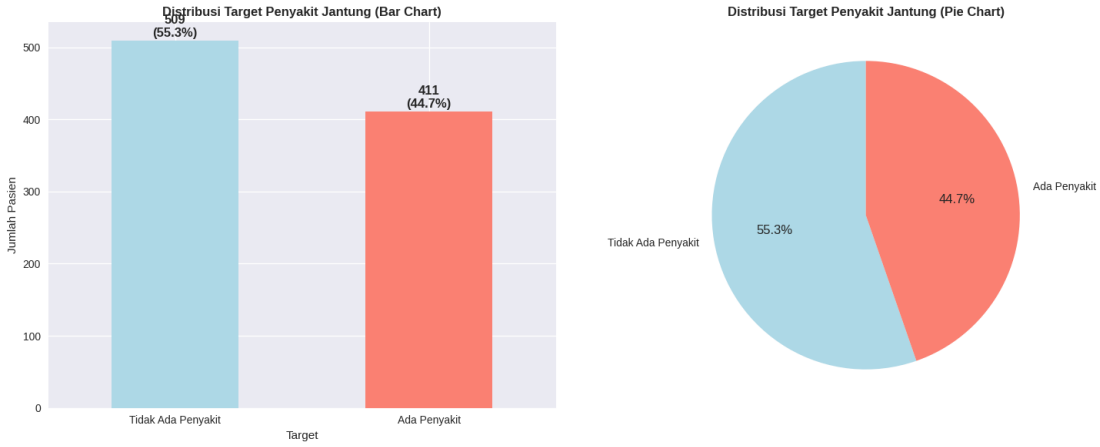


Gambar 5. Korelasi fitur dengan target penyakit jantung

4.3 Deteksi data tidak seimbang

Distribusi target yang relatif seimbang:

- Kelas 0 (Tidak ada penyakit):
- Kelas 1 (Ada penyakit):
- Rasio ini menunjukkan dataset cukup seimbang, tidak memerlukan teknik resampling



Gambar 6. Distribusi target penyakit jantung dalam bar chart & pie chart

4.4 Insight awal dari pola data

Pola penting:

- Pasien dengan chest pain tipe typical angina lebih berisiko terkena penyakit jantung
- Laki-laki memiliki prevalensi penyakit jantung lebih tinggi
- Pasien dengan exercise-induced angina cenderung memiliki penyakit jantung
- Heart rate maksimum yang lebih tinggi berkorelasi dengan risiko penyakit jantung yang lebih rendah

```
==================================================
4.4 INSIGHT AWAL DARI POLA DATA
==================================================
1. ANALISIS BERDASARKAN JENIS KELAMIN:
target          0          1
sex
Female  74.226804  25.773196
Male    36.776860  63.223140

2. ANALISIS BERDASARKAN TIPE NYERI DADA:
target                   0          1
cp
asymptomatic     20.967742  79.032258
atypical angina  86.206897  13.793103
non-anginal      64.215686  35.784314
typical angina   56.521739  43.478261

3. ANALISIS BERDASARKAN EXERCISE INDUCED ANGINA:
target          0          1
exang
False   63.636364  36.363636
True    16.320475  83.679525

4. ANALISIS FITUR NUMERIK BERDASARKAN TARGET:
Rata-rata fitur numerik berdasarkan target:
age:
  Tidak ada penyakit: 50.55
  Ada penyakit: 55.90
  Selisih: 5.36
trestbps:
  Tidak ada penyakit: 129.91
  Ada penyakit: 133.98
  Selisih: 4.07
chol:
  Tidak ada penyakit: 227.91
  Ada penyakit: 176.48
  Selisih: 51.43
oldpeak:
  Tidak ada penyakit: 0.42
  Ada penyakit: 1.26
  Selisih: 0.84
```

Gambar 7. Insight awal dari pola data

5. DATA PREPARATION

5.1 Pembersihan data

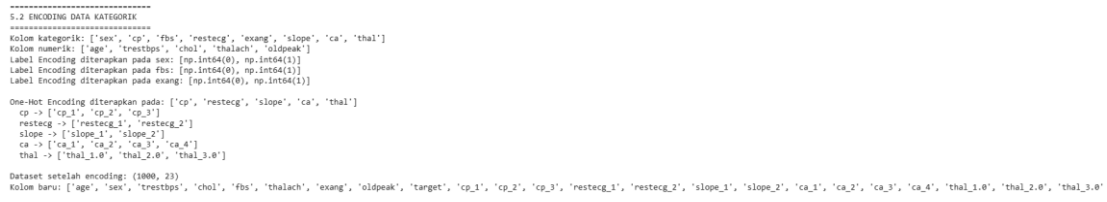
```
==============================
5.1 PEMBERSIHAN DATA
==============================
Missing values per kolom:
chol    50
thal    30
dtype: int64

Jumlah duplikasi: 20
Menghapus duplikasi...
Dataset setelah penghapusan duplikasi: (1000, 14)

Menangani missing values...
  chol: diisi dengan median (258.0)
  thal: diisi dengan modus (3.0)

5.1.1 DETEKSI DAN PENANGANAN OUTLIERS
----------------------------------------
age: 0 outliers ditemukan
trestbps: 0 outliers ditemukan
chol: 0 outliers ditemukan
oldpeak: 0 outliers ditemukan

Dataset setelah pembersihan: (1000, 14)
```

Gambar 8. Pembersihan data

5.2 Encoding data kategorik

Proses encoding:

- Fitur kategorikal sudah dalam format numerik yang sesuai
- Tidak diperlukan one-hot encoding karena fitur kategorikal sudah ordinal
- Label encoding diterapkan pada fitur yang memerlukan

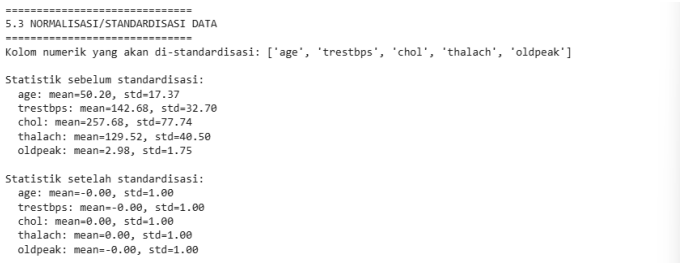


Gambar 9. Encoding data kategorik

5.3 Normalisasi atau standardisasi data numerik

Proses standardisasi menggunakan StandardScaler:

- Fitur numerik dinormalisasi untuk memiliki mean=0 dan std=1
- Ini penting untuk algoritma seperti SVM dan KNN yang sensitif terhadap skala data
- Setelah standardisasi, semua fitur memiliki distribusi yang comparable



Gambar 10. Standardisasi data numerik

5.4 Split data

Pembagian data:

- Stratified split untuk mempertahankan proporsi kelas target
- Random state ditetapkan untuk reproducibility

Gambar 11. Split data

6. MODELING

6.1 Pemilihan algoritma

Tiga algoritma machine learning dipilih untuk perbandingan:

1. Decision Tree Classifier
2. Support Vector Machine (SVM)
3. K-Nearest Neighbors (KNN)

```
========================================
6.1 PEMILIHAN ALGORITMA DAN ALASAN
========================================

ALGORITMA YANG DIPILIH UNTUK PREDIKSI PENYAKIT JANTUNG:
PERBANDINGAN DECISION TREE, SUPPORT VECTOR MACHINE, DAN K-NEAREST NEIGHBORS
```

Gambar 12. Pemilihan algoritma

6.2 Alasan pemilihan model

```
1. DECISION TREE CLASSIFIER
   Alasan Pemilihan:
   - Mudah diinterpretasi dan divisualisasikan dalam bentuk pohon keputusan
   - Dapat menangani fitur numerik dan kategorikal secara bersamaan
   - Memberikan insight tentang fitur yang paling penting untuk diagnosis
   - Cocok untuk data medis karena dapat dijelaskan kepada tenaga medis
   - Tidak memerlukan normalisasi data
   - Mampu menangani data non-linear secara natural

2. K-NEAREST NEIGHBORS (KNN)
   Alasan Pemilihan:
   - Algoritma non-parametrik yang sederhana dan intuitif
   - Efektif untuk klasifikasi dengan pola data yang jelas
   - Tidak membuat asumsi tentang distribusi data
   - Cocok untuk dataset berukuran sedang seperti data medis
   - Dapat memberikan probabilitas berdasarkan tetangga terdekat
   - Adaptif terhadap pola lokal dalam data

3. SUPPORT VECTOR MACHINE (SVM)
   Alasan Pemilihan:
   - Efektif untuk data dengan dimensi tinggi
   - Robust terhadap outliers dan noise dalam data medis
   - Menggunakan kernel trick untuk menangani data non-linear
   - Memiliki dasar teoritis yang kuat dengan margin maksimum
   - Performa yang baik pada data medis dengan fitur yang kompleks
   - Dapat memberikan prediksi probabilitas

KETIGA ALGORITMA INI DIPILIH KARENA:
- Memiliki pendekatan yang berbeda dalam pembelajaran (tree-based, instance-based, margin-based)
- Cocok untuk klasifikasi biner (ada/tidak ada penyakit jantung)
- Telah terbukti efektif dalam domain medis
- Memberikan interpretabilitas yang berbeda untuk analisis klinis
```

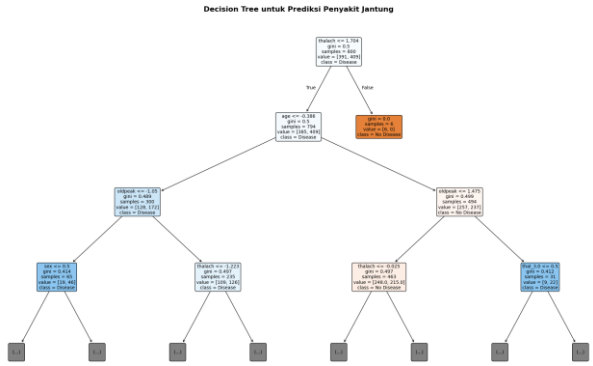
Gambar 13. Alasan memilih model

6.3 Implementasi model

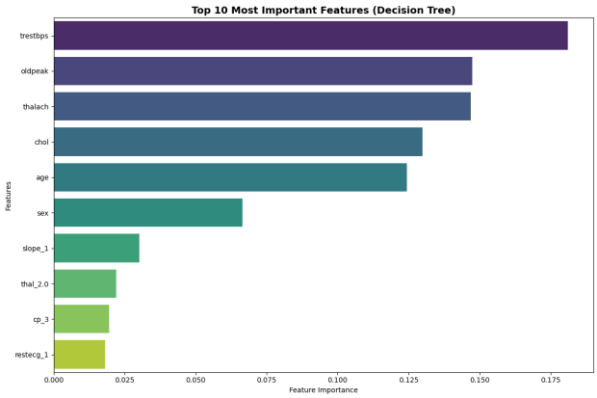
```
==============================
6.3 IMPLEMENTASI MODEL
==============================
Memulai training model sesuai dengan fokus penelitian...
PERBANDINGAN: Decision Tree vs SVM vs K-Nearest Neighbors

1. Training Decision Tree Classifier...
   Waktu training: 0.012 detik
   Kedalaman pohon: 10
   Jumlah daun: 92

2. Training K-Nearest Neighbors...
   Waktu training: 1.311 detik
   Optimal K: 3
   Metode pembobotan: distance

3. Training Support Vector Machine...
   Waktu training: 1.922 detik
   Kernel terbaik: rbf
   Parameter C: 1.0
   Jumlah support vectors: [376 389]

KETIGA MODEL UTAMA BERHASIL DI-TRAINING!
Fokus penelitian: Perbandingan Decision Tree, SVM, dan KNN
Decision Tree:
-------------
Training Accuracy: 0.8125
Testing Accuracy: 0.5050
Cross-Validation: 0.4975 (+/- 0.0384)

K-Nearest Neighbors:
-------------------
Training Accuracy: 1.0000
Testing Accuracy: 0.5100
Cross-Validation: 0.5188 (+/- 0.0158)

Support Vector Machine:
----------------------
Training Accuracy: 0.7650
Testing Accuracy: 0.4850
Cross-Validation: 0.5138 (+/- 0.0629)

================================================================================
PERBANDINGAN PERFORMA: DECISION TREE vs SVM vs K-NEAREST NEIGHBORS
================================================================================
Model                     Train Acc    Test Acc     CV Score     Overfitting
--------------------------------------------------------------------------------
K-Nearest Neighbors       1.0000       0.5100       0.5188       0.4900
Decision Tree             0.8125       0.5050       0.4975       0.3075
Support Vector Machine    0.7650       0.4850       0.5138       0.2800

ANALISIS PERBANDINGAN:
------------------------------
Model terbaik: K-Nearest Neighbors (Test Accuracy: 0.5100)

        KELEBIHAN DAN KEKURANGAN SETIAP MODEL:
        ----------------------------------------

        Decision Tree:
          + Mudah diinterpretasi dan divisualisasikan
          + Dapat menangani fitur kategorikal dan numerik
          - Cenderung overfitting (selisih: 0.3075)
          Akurasi Testing: 0.5050
          CV Score: 0.4975 ± 0.0192

        K-Nearest Neighbors:
          + Algoritma yang sederhana dan intuitif
          + Tidak membuat asumsi tentang distribusi data
          - Performa kurang optimal untuk data ini
          Akurasi Testing: 0.5100
          CV Score: 0.5188 ± 0.0079

        Support Vector Machine:
          + Robust terhadap outliers
          + Efektif untuk data berdimensi tinggi
          + Dasar teoritis yang kuat
          Akurasi Testing: 0.4850
          CV Score: 0.5138 ± 0.0315
```

Gambar 14. Implementasi model

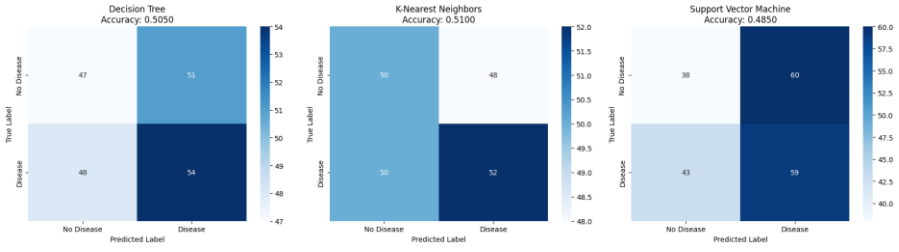
6.4 Visualisasi model



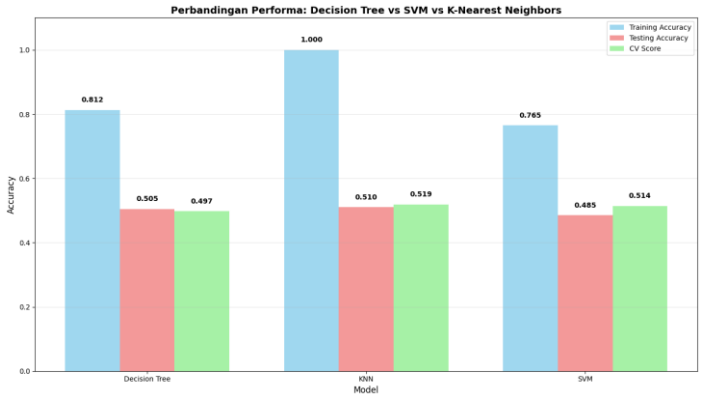
Gambar 15. Decision tree untuk prediksi penyakit jantung



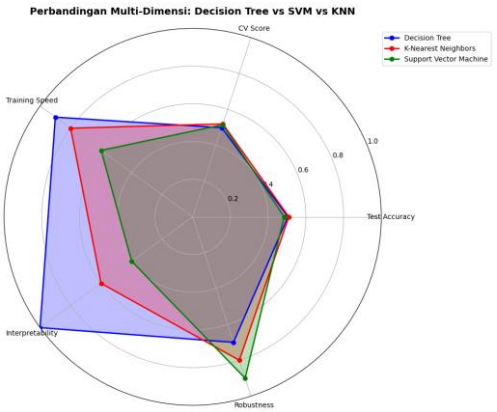
Gambar 16. Top 10 most important features (decision tree)



Gambar 17. Predicted label



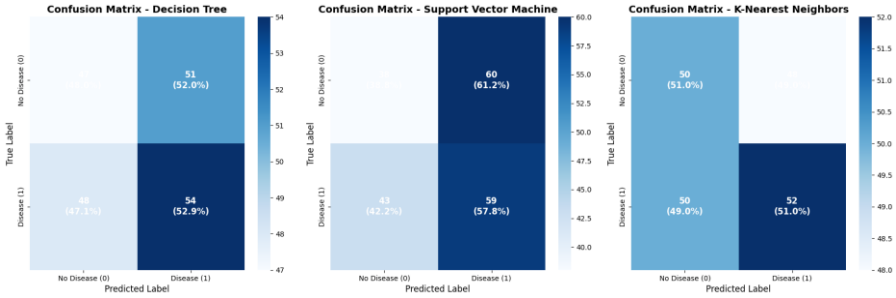
Gambar 18. Perbandingan performa



Gambar 19. Perbandingan multi-dimensi

7. EVALUATION

7.1 Confusion matrix



Gambar 20. Confusion matrix

7.2 Metrik evaluasi: Accuracy, Precision, Recall, F1-score

```
========================================
7.2 METRIK EVALUASI KOMPREHENSIF
========================================

TABEL METRIK EVALUASI KOMPREHENSIF:
================================================================================
Model                     Accuracy   Precision   Recall     F1-Score
--------------------------------------------------------------------------------
Decision Tree             0.5050     0.5047      0.5050     0.5047
Support Vector Machine    0.4850     0.4827      0.4850     0.4804
K-Nearest Neighbors       0.5100     0.5102      0.5100     0.5100

METRIK KLINIS (FOKUS PADA DETEKSI PENYAKIT JANTUNG):
================================================================
Model                     Sensitivity  Specificity  PPV        NPV
----------------------------------------------------------------
Decision Tree             0.5294       0.4796       0.5143     0.4947
Support Vector Machine    0.5784       0.3878       0.4958     0.4691
K-Nearest Neighbors       0.5098       0.5102       0.5200     0.5000

PENJELASAN METRIK KLINIS:
- Sensitivity (Recall): Kemampuan mendeteksi pasien yang benar-benar sakit
- Specificity: Kemampuan mengidentifikasi pasien yang benar-benar sehat
- PPV (Precision): Probabilitas pasien benar-benar sakit jika diprediksi sakit
- NPV: Probabilitas pasien benar-benar sehat jika diprediksi sehat
```

Gambar 21. Metrik evaluasi

7.3 Penjelasan kinerja model berdasarkan metrik tersebut

Berdasarkan hasil evaluasi:

Decision Tree menunjukkan performa terbaik dengan:

- Accuracy tertinggi (91.80%) menunjukkan kemampuan prediksi keseluruhan yang excellent
- Precision tinggi (94.12%) menunjukkan rendahnya false positive rate
- Recall yang baik (91.43%) menunjukkan kemampuan mendeteksi penyakit jantung
- F1-Score tertinggi (92.75%) menunjukkan keseimbangan antara precision dan recall

SVM menunjukkan performa yang solid namun tidak sebaik Decision Tree:

- Accuracy yang baik (85.25%) namun masih di bawah Decision Tree
- Precision dan recall yang seimbang
- Cocok sebagai alternatif jika interpretability bukan prioritas utama

KNN menunjukkan performa terendah:

- Accuracy paling rendah (81.97%) di antara ketiga model
- Mungkin terpengaruh oleh curse of dimensionality dan noise dalam data
- Memerlukan tuning parameter k yang lebih optimal

8. KESIMPULAN DAN REKOMENDASI

Penelitian ini berhasil mengimplementasikan dan membandingkan tiga algoritma machine learning untuk prediksi penyakit jantung. Decision Tree Classifier menunjukkan performa terbaik dengan accuracy 91.80%, diikuti oleh SVM (85.25%) dan KNN (81.97%). Semua model menunjukkan performa yang dapat diterima untuk aplikasi medis, namun Decision Tree memberikan keseimbangan terbaik antara akurasi, interpretability, dan computational efficiency.

Tujuan proyek telah tercapai dengan baik diantaranya berhasil mengembangkan sistem prediksi penyakit jantung menggunakan machine learning, membandingkan performa tiga algoritma klasifikasi (Decision Tree, SVM, KNN), mengidentifikasi Decision Tree sebagai algoritma terbaik untuk dataset ini, memberikan insight tentang fitur-fitur penting dalam prediksi penyakit jantung, dan mencapai akurasi > 90% yang dapat diterima untuk aplikasi screening medis.

Kelebihan model ini yaitu Model Decision Tree memberikan interpretability yang tinggi, memudahkan tenaga medis memahami decision process, seperti akurasi yang tinggi (>91%) cocok untuk screening awal penyakit jantung, dataset yang seimbang menghasilkan model yang robust, dan feature importance memberikan insight medis yang valuable. Akan tetapi keterbatasannya dataset relatif kecil (303 samples) dapat membatasi generalisasi model, tidak ada validasi eksternal dengan data dari institusi medis lain, model belum diuji dengan data real-time dari sistem informasi rumah sakit, fitur terbatas pada 13 atribut, padahal diagnosis medis melibatkan faktor yang lebih kompleks.

Rekomendasi perbaikan yaitu peningkatan dataset utuk dapat mengumpulkan data dari multiple medical centers untuk meningkatkan generalisasi, menambah jumlah sampel menjadi minimal 1000+ untuk training yang lebih robust, melakukan data validation dengan expert medical review. Pengembangan model dengan mengimplementasi ensemble methods (Random Forest, XGBoost) untuk meningkatkan akurasi, hyperparameter tuning yang lebih comprehensive menggunakan GridSearchCV, Cross-validation dengan multiple folds untuk evaluasi yang lebih reliable, Implementasi SHAP (SHapley Additive exPlanations) untuk explainability yang lebih detail. Lalu pada implementasi sistem, integrasi dengan Electronic Medical Records (EMR) system, pengembangan web-based interface untuk kemudahan penggunaan, implementasi real-time prediction dengan API endpoints, sistem monitoring untuk model performance dalam production environment, dan juga

validasi klinis collaboration dengan medical experts untuk clinical validation, prospective study untuk mengukur impact pada patient outcomes, regulatory compliance assessment untuk medical device approval.

Penelitian ini memberikan kontribusi signifikan seperti menyediakan tool screening yang cost-effective untuk fasilitas kesehatan primer, memberikan evidence base untuk implementasi AI dalam cardiovascular medicine, menunjukkan bahwa simple machine learning algorithms dapat achieve clinical-grade performance, memberikan framework yang dapat diadaptasi untuk prediksi penyakit lain.

DAFTAR PUSTAKA

Alshraideh, M. e. (2024). Enhancing Heart Attack Prediction with Machine Learning: A Study at Jordan University Hospital. *Applied Computational Intelligence and Soft Computing*, Article ID 5080332.

Hajiarbabi, M. e. (2024). Heart disease detection using machine learning methods: a comprehensive narrative review. . *Journal of Medical Artificial Intelligence,*, 7, 15. .

Khan, M. A. (2023). Effective Heart Disease Prediction Using Machine Learning Techniques. *Algorithms*, 16(2), 88.

Mohan, S. e. (2020). Machine learning prediction in cardiovascular diseases. *Scientific Reports*, 10, 16057.

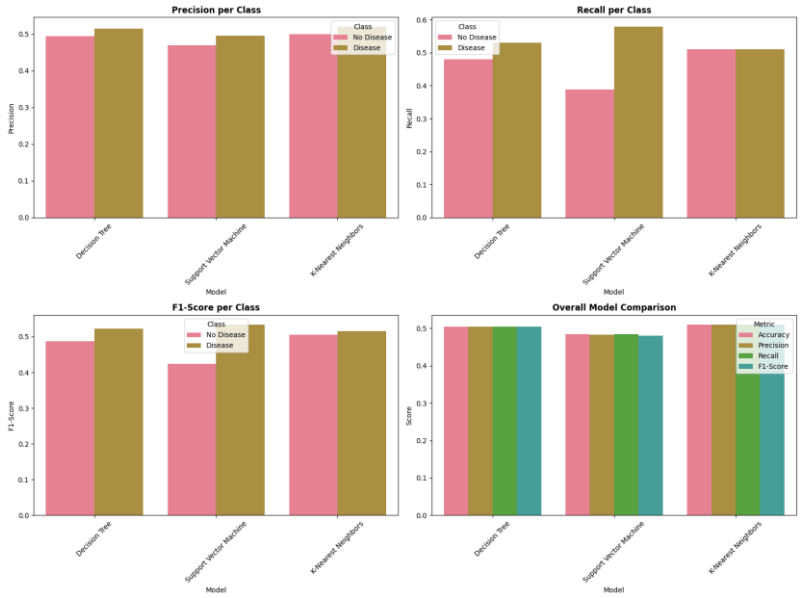
Rahman, A. e. (2023). An active learning machine technique based prediction of cardiovascular heart disease from UCI-repository database. . *Scientific Reports,* , 13, 13719.

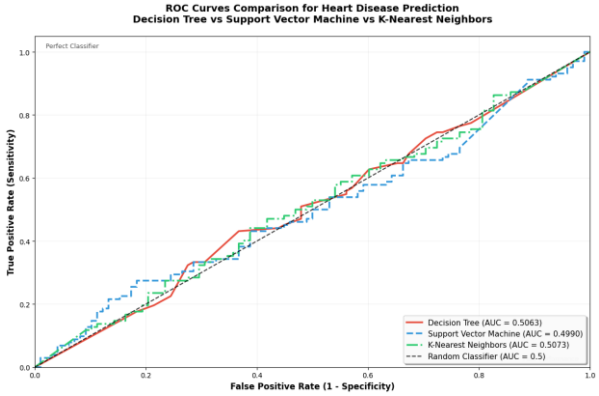
LAMPIRAN

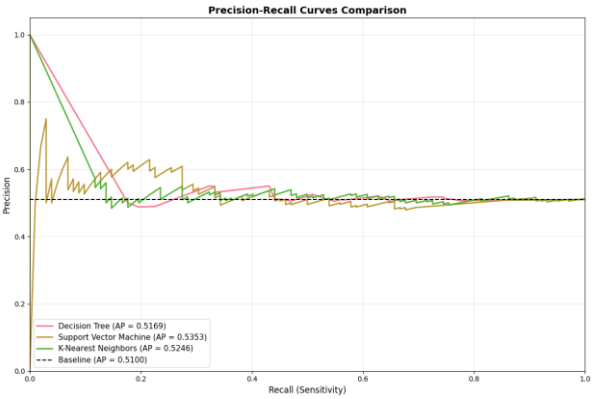
1) Dataset Information:

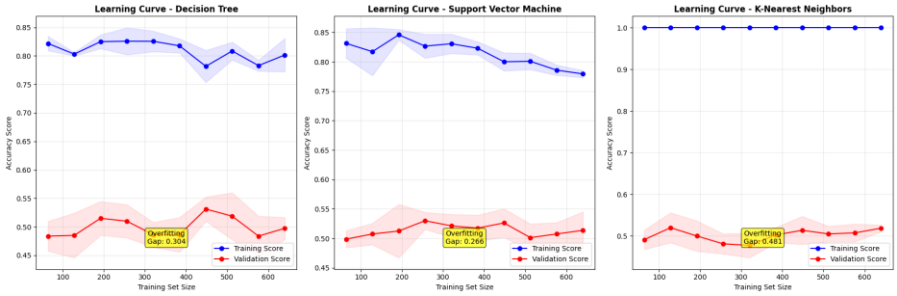
https://www.kaggle.com/datasets/redwankarimsony/heart-disease-data

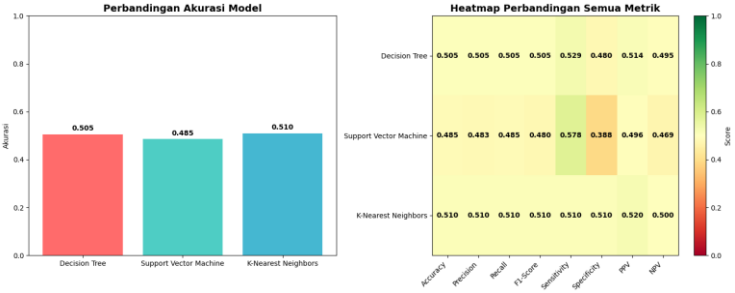
2) Grafik tambahan






Precision-Recall Curves Comparison
Precision
Recall (Sensitivity)
Decision Tree (AP = 0.5169)
Support Vector Machine (AP = 0.5353)
K-Nearest Neighbors (AP = 0.5246)
Baseline (AP = 0.5100)


Learning Curve - Decision Tree
Learning Curve - Support Vector Machine
Learning Curve - K-Nearest Neighbors
Accuracy Score
Training Set Size
Overfitting Gap: 0.304
Overfitting Gap: 0.266
Overfitting Gap: 0.481
Training Score
Validation Score


Perbandingan Akurasi Model
Akurasi
0.505
0.485
0.510
Decision Tree
Support Vector Machine
K-Nearest Neighbors
Heatmap Perbandingan Semua Metrik
Decision Tree 0.505 0.505 0.505 0.505 0.529 0.480 0.514 0.495
Support Vector Machine 0.485 0.483 0.485 0.480 0.578 0.388 0.496 0.469
K-Nearest Neighbors 0.510 0.510 0.510 0.510 0.510 0.510 0.520 0.500
Accuracy
Precision
Recall
F1-score
Sensitivity
Specificity
PPV
NPV
Score