**JOMLAI: Journal of Machine Learning and Artificial Intelligence**
Vol. 5, No. 2, Juni 2026, pp. 57~64
ISSN: 2828-9102 (Print) | 2828-9099 (Online)
DOI: 10.55123/jomlai.v5i2.5430


# Deteksi Dini Risiko Penyakit Kardiovaskular Menggunakan Algoritma K-Nearest Neighbors

## *Early Detection of Cardiovascular Disease Risk Using the K-Nearest Neighbors Algorithm*

**Leo Fernandy[1], Leonardo[2], Stanley Lim[3], Vincent Liawis[4], Ade Maulana[5]**
[1,2,3,4,5*]Universitas Pelita Harapan, Indonesia

**Article Info**

***Genesis Artikel:***

Diterima, 27 April 2026
Direvisi, 01 Juni 2026
Disetujui, 15 Juni 2026

***Kata Kunci:***

*Kardiovaskular*
*KNN*
K-Nearest Neighbors
*Prediksi*
*Machine Learning*
*Evaluasi Model*

**ABSTRAK**

Kesehatan merupakan aspek fundamental dalam menentukan kualitas hidup manusia. Seiring perubahan pola hidup dan kondisi lingkungan, muncul berbagai tantangan baru dalam sektor kesehatan. Kemajuan teknologi, khususnya kecerdasan buatan, telah membuka peluang besar dalam pengembangan sistem kesehatan yang lebih akurat dan efisien. Salah satu penerapan AI yang berkembang pesat adalah machine learning dalam prediksi penyakit. Penelitian ini bertujuan membangun model prediksi risiko penyakit kardiovaskular menggunakan algoritma *K-Nearest Neighbors* (KNN). Dataset "Cardiovascular Disease" dari Kaggle dengan 68.205 data dan 17 atribut medis digunakan sebagai basis. Tahapan penelitian mencakup praproses data (pembersihan, transformasi kategori, dan normalisasi), pemilihan fitur utama, pelatihan model, serta evaluasi performa. Data dibagi menjadi 80% data latih dan 20% data uji. Hasil menunjukkan nilai $k$ = 41 menghasilkan akurasi tertinggi sebesar 73%. Evaluasi menggunakan precision, recall, dan f1-score menunjukkan performa yang cukup baik, terutama dalam mengidentifikasi pasien dengan risiko tinggi. Model ini berpotensi digunakan sebagai alat bantu deteksi dini penyakit kardiovaskular untuk mendukung pengambilan keputusan medis yang lebih tepat dan preventif.

***Keywords:***

*Cardiovascular*
*KNN*
K-Nearest Neighbors
*Prediction*
*Machine Learning*
*Model Evaluation*

***ABSTRACT***

*Health is a fundamental aspect in determining the quality of human life. Along with changes in lifestyle and environmental conditions, various new challenges have emerged in the healthcare sector. Technological advancements, particularly in artificial intelligence, have opened significant opportunities for the development of more accurate and efficient healthcare systems. One of the most rapidly growing applications of AI is machine learning for disease prediction. This study aims to develop a model for predicting the risk of cardiovascular disease using the* K-Nearest Neighbors *(KNN) algorithm. The "Cardiovascular Disease" dataset from Kaggle, consisting of 68,205 entries and 17 medical attributes, was used as the basis. The research stages included data preprocessing (cleaning, categorical transformation, and normalization), selection of key features, model training, and performance evaluation. The dataset was split into 80% training data and 20% testing data. The experiment showed that* k *= 41 achieved the highest accuracy of 73%. Evaluation using precision, recall, and f1-score indicated fairly good performance, particularly in identifying high-risk patients. This model has the potential to serve as a decision-support tool for early detection of cardiovascular disease, enabling more accurate and preventive medical actions..*



***Penulis Korespondensi:***

Ade Maulana
Program Studi Sistem Informasi,
Universitas Pelita Harapan,
Email: ade.maulana@lecturer.uph.edu

## 1. PENDAHULUAN

Kesehatan merupakan salah satu aspek fundamental yang menentukan kualitas hidup manusia [1]. Perkembangan zaman turut membawa tantangan baru dalam sektor kesehatan, yang dipengaruhi oleh berbagai perubahan dalam pola hidup dan kondisi lingkungan masyarakat. Maka dari itu, untuk mencapai kesejahteraan yang berkelanjutan, diperlukan sinergi upaya peningkatan kesehatan. Sejalan dengan itu, kemajuan teknologi juga telah membuka peluang baru dalam dunia kesehatan. Teknologi seperti kecerdasan buatan, khususnya *machine learning*, mulai dimanfaatkan dalam berbagai aspek [2]. Dari pengolahan data medis, prediksi penyakit, hingga mendukung pengambilan keputusan klinis sekarang telah jauh lebih cepat dan tepat. Inovasi berbasis data ini menghadirkan potensi besar dalam mendorong transformasi pelayanan kesehatan masyarakat menuju sistem yang lebih akurat, efisien, serta berkelanjutan [3].

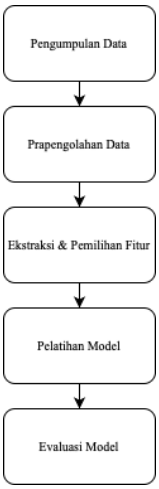
Salah satu tantangan besar dalam dunia kesehatan adalah penyakit kardiovaskular, atau yang dikenal juga dengan istilah *Cardiovascular Disease* (CVD) adalah sekelompok penyakit yang menyerang jantung dan pembuluh darah. Penyakit ini meliputi serangan jantung, stroke, dan penyakit jantung koroner [4]. CVD menjadi salah satu penyebab utama kematian di dunia [5]. Menurut data dari *World Health Organization*, sekitar 17,9 juta orang meninggal setiap tahunnya akibat penyakit ini [6]. Faktor utama yang menyebabkan CVD antara lain adalah gaya hidup tidak sehat, seperti merokok, kurang olahraga, dan pola makan tinggi lemak atau garam [7]. Kebiasaan-kebiasaan tersebut dapat memicu penyumbatan pembuluh darah, tekanan darah tinggi, dan masalah jantung lainnya. Karena gejala awal CVD sering tidak terlihat, deteksi dini sangat penting agar risiko dapat diantisipasi sejak awal [8]. Dengan bantuan teknologi seperti *machine learning*, sistem komputer dapat dilatih untuk memprediksi risiko penyakit ini berdasarkan data kesehatan individu, yang diharapkan mampu mengurangi angka kematian dan meningkatkan kualitas hidup pasien melalui pencegahan yang lebih dini dan tepat sasaran [9].

Dalam upaya mengklasifikasikan risiko penyakit kardiovaskular, berbagai algoritma telah diuji dalam sejumlah penelitian terdahulu. Salah satunya adalah metode *Logistic Regression* yang menghasilkan akurasi sebesar 86,84% pada penelitian yang dilakukan oleh Dharmawan [10]. Sementara itu, pengujian menggunakan algoritma *Random Forest* juga dilakukan oleh Jayaditya & Kadyanan [11] yang menunjukkan akurasi sebesar 73,06%. Di sisi lain, studi oleh Artanti [12] yang menerapkan algoritma *K-Nearest Neighbors* (KNN) pada kasus prediksi penyakit kardiovaskular menunjukkan hasil yang sangat baik, dengan akurasi mencapai 91%. Melihat tingginya akurasi dan kestabilan performa KNN dalam berbagai studi terdahulu, algoritma ini menjadi salah satu metode yang potensial untuk diterapkan lebih lanjut dalam konteks prediksi penyakit kardiovaskular.

Dengan mempertimbangkan urgensi deteksi dini penyakit kardiovaskular serta potensi besar yang dimiliki oleh teknologi *machine learning*, penelitian ini bertujuan untuk mengeksplorasi penerapan metode KNN dalam memprediksi risiko penyakit kardiovaskular. Menggunakan data dari Kaggle yang relevan dan mengoptimalkan parameter model, diharapkan sistem prediksi berbasis KNN ini dapat memberikan solusi preventif yang akurat, efisien, dan aplikatif, mendukung pelayanan kesehatan yang lebih baik dan berkelanjutan.

## 2. METODE PENELITIAN

Penelitian ini menerapkan pendekatan sistematis untuk membangun model prediksi risiko penyakit kardiovaskular menggunakan algoritma K-Nearest Neighbors (KNN). Tahapan dimulai dari prapengolahan data yang mencakup pembersihan data, transformasi atribut, dan normalisasi untuk memastikan kualitas dan konsistensi data. Selanjutnya dilakukan pemilihan fitur berdasarkan relevansi terhadap risiko penyakit, dilanjutkan dengan pelatihan model menggunakan data yang telah dibagi ke dalam data latih dan data uji. Evaluasi model dilakukan menggunakan metrik klasifikasi untuk mengukur performa prediksi. Seluruh proses dirancang guna menghasilkan model yang andal dan dapat digunakan dalam mendukung deteksi dini penyakit kardiovaskular. Gambar 1i merupakan tahapan-tahapan yang dilakukan dalam penelitian ini.




Gambar 1. Alur Pengujian

### 2.1. Pengumpulan Data

Dataset yang berjudul *Cardiovascular Disease* diperoleh dan diunduh melalui website Kaggle. Dataset ini terdiri dari 68.205 entri data dengan 17 atribut yang mencakup berbagai informasi medis terkait penyakit kardiovaskular. Dataset ini dapat diakses melalui platform Kaggle dengan tautan berikut: Kaggle - Cardiovascular Disease Dataset [13]. Berikut adalah keterangan dari atribut dataset yang digunakan:

Tabel 1. Keterangan Atribut Dataset *Cardiovascular Disease*

| No. | Atribut | Keterangan |
|---|---|---|
| 1. | ID | Identifikasi unik untuk setiap pasien |
| 2. | *age* | Usia pasien dalam hari |
| 3. | *age_years* | Usia pasien dalam tahun (diturunkan dari usia dalam hari) |
| 4. | *gender* | Jenis kelamin pasien. Variabel kategorikal (1: Perempuan, 2: Laki-laki) |
| 5. | *height* | Tinggi badan pasien dalam sentimeter (cm) |
| 6. | *weight* | Berat badan pasien dalam kilogram (kg) |
| 7. | ap_hi | Tekanan darah sistolik |
| 8. | ap_lo | Tekanan darah diastolik |
| 9. | *cholesterol* | Tingkat kolesterol. Variabel kategorikal (1: Normal, 2: Di Atas Normal, 3: Jauh Di Atas Normal) |
| 10. | *gluc* | Tingkat glukosa. Variabel kategorikal (1: Normal, 2: Di Atas Normal, 3: Jauh Di Atas Normal) |
| 11. | *smoke* | Status merokok. Variabel biner (0: Tidak Merokok, 1: Merokok) |
| 12. | *alco* | Konsumsi alkohol. Variabel biner (0: Tidak mengonsumsi alkohol, 1: Mengonsumsi alkohol) |
| 13. | *active* | Aktivitas fisik. Variabel biner (0: Tidak aktif secara fisik, 1: Aktif secara fisik) |
| 14. | *cardio* | Ada atau tidaknya penyakit kardiovaskular. Variabel target. Biner (0: Tidak ada, 1: Ada). |
| 15. | bmi | Indeks Massa Tubuh, yang dihitung dari berat badan dan tinggi badan. |
| 16. | *bp_category* | Kategori tekanan darah berdasarkan ap_hi dan ap_lo. Kategori mencakup "Normal", "Tinggi", " Hipertensi Tahap 1", "Hipertensi Tahap 2", dan "Krisis Hipertensi". |
| 17. | *bp_category_encoded* | Bentuk terkodekan dari bp_category untuk keperluan pembelajaran mesin. |

### 2.2. Prapengolahan Data

Sebelum data digunakan untuk melatih model *K-Nearest Neighbors* (KNN), dilakukan prapengolahan data guna meningkatkan kualitas data dan kinerja model. Tahapan ini mencakup penghapusan data duplikat untuk menghindari bias, penyaringan *outlier* agar data ekstrim tidak mempengaruhi hasil prediksi, konversi data kategorik menjadi bentuk numerik agar dapat diproses oleh model, serta normalisasi data dengan metode *Min-Max Scaling* untuk menyamakan skala antar fitur. Langkah-langkah ini penting agar model KNN dapat bekerja secara optimal dan menghasilkan prediksi yang akurat [14].

### 2.3. Ekstraksi dan Pemilihan Fitur

Pada tahap ini, dilakukan pemilihan fitur yang relevan untuk digunakan dalam pelatihan model klasifikasi. Dari total 17 atribut dalam dataset, tiga atribut dihapus, yaitu 'ID' karena tidak memengaruhi proses prediksi, '*age*' dalam satuan hari karena telah direpresentasikan oleh atribut '*age_years*' dalam satuan tahun, serta '*bp_category_encoded*' yang bersifat redundan terhadap atribut '*bp_category*'. Sementara itu, atribut '*cardio*' ditetapkan sebagai variabel target yang menunjukkan keberadaan penyakit kardiovaskular. Pemilihan fitur ini bertujuan untuk menyederhanakan model dan meningkatkan akurasi prediksi.

### 2.4. Pelatihan Model

Setelah melalui tahapan prapengolahan data dan pemilihan fitur, proses pelatihan model dilakukan menggunakan algoritma *K-Nearest Neighbors* (KNN). Dataset dibagi menjadi dua bagian, yaitu data pelatihan sebesar 80% dan data pengujian sebesar 20%. Data pelatihan dinormalisasi menggunakan metode *Standard Scaler* untuk menyamakan skala antar fitur. Model KNN kemudian dilatih dengan data tersebut untuk mengklasifikasikan kemungkinan adanya penyakit kardiovaskular berdasarkan parameter yang tersedia. Proses pelatihan ini bertujuan untuk membangun model yang mampu mengenali pola dan hubungan antara variabel input dengan target secara efektif.

### 2.5. Evaluasi Model

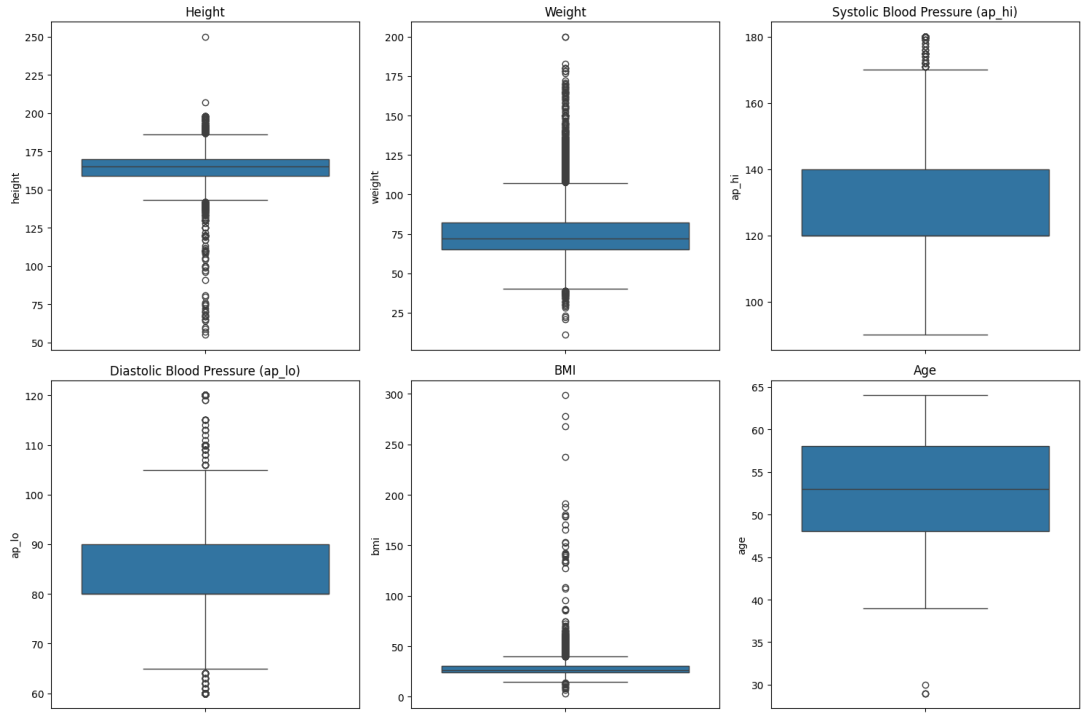
Pada tahap evaluasi model, dilakukan pengujian terhadap performa algoritma KNN yang telah dilatih menggunakan data uji. Evaluasi dilakukan dengan menggunakan metrik akurasi, *precision, recall,* dan *f1-score* yang disajikan dalam *classification report*. Untuk memperoleh performa terbaik, eksperimen dilakukan dengan mencoba berbagai nilai k pada algoritma KNN. Selain *classification report*, evaluasi juga dilengkapi dengan *confusion matrix* untuk memberikan gambaran yang lebih rinci mengenai prediksi model terhadap masing-masing kelas. *Confusion matrix* digunakan untuk menunjukkan jumlah prediksi yang benar dan salah untuk setiap kelas, termasuk *true positives* (TP), *true negatives* (TN), *false positives* (FP), dan *false negatives*

(FN). Penggunaan *confusion matrix* sangat penting terutama dalam kasus klasifikasi biner [15], seperti deteksi penyakit kardiovaskular, karena dapat membantu mengidentifikasi potensi kesalahan yang berdampak besar, seperti kegagalan dalam mengenali individu yang sebenarnya memiliki risiko (false negative).

## 3. HASIL DAN PEMBAHASAN

### 3.1. Prapengolahan Data

Dataset yang digunakan terdiri dari **68.205 entri** dan **17 atribut** yang mencakup berbagai informasi kesehatan. Tahap awal dalam proses prapengolahan dimulai dengan eksplorasi data untuk memahami karakteristik setiap atribut, distribusi nilai, jenis data, serta kemungkinan adanya anomali. Selanjutnya dilakukan pemeriksaan terhadap nilai yang hilang (*missing values*) dan data duplikat. Hasil analisis menunjukkan bahwa tidak terdapat nilai kosong pada atribut manapun. Namun, ditemukan sebanyak 3.206 entri data duplikat, yang kemudian dihapus guna menjaga kualitas dataset sebelum melanjutkan ke tahap berikutnya.




Gambar 2. Visualisasi Outlier

Tahapan selanjutnya adalah mengidentifikasi dan menangani outlier atau nilai ekstrem yang secara statistik maupun secara logika tidak wajar dalam konteks medis. Berdasarkan visualisasi menggunakan boxplot pada gambar di atas, beberapa atribut numerik seperti tinggi badan, berat badan, tekanan darah sistolik dan diastolik, usia, serta indeks massa tubuh (BMI) menunjukkan distribusi data yang menyimpang, dengan nilai yang terlalu rendah atau terlalu tinggi. Nilai ekstrem ini berpotensi mengganggu proses pelatihan model dan menurunkan akurasi prediksi. Untuk mengurangi dampak tersebut, dilakukan penyaringan data dengan batasan nilai yang lebih realistis. Tinggi badan dibatasi antara 120 cm hingga 200 cm guna menghindari data yang tidak masuk akal pada orang dewasa. Berat badan dibatasi dalam rentang 30 kg hingga 160 kg agar entri yang terlalu kecil atau besar dapat dieliminasi. Tekanan darah sistolik difilter dalam rentang 90 hingga 200 mmHg, sedangkan tekanan darah diastolik dibatasi antara 60 hingga 140 mmHg. Nilai BMI yang digunakan berada pada kisaran 10 hingga 60, karena nilai di luar rentang ini sangat jarang terjadi. Usia difokuskan pada individu berumur antara 30 hingga 65 tahun, sesuai dengan fokus analisis terhadap risiko penyakit kardiovaskular yang umumnya terjadi pada kelompok usia tersebut.

Setelah proses pembersihan data, dilakukan konversi pada atribut kategorikal menjadi numerikal agar dapat diproses oleh algoritma pembelajaran mesin. Atribut yang dikonversi adalah *'bp_category'*, yang merupakan klasifikasi tekanan darah. Nilai kategorikal pada atribut ini diubah ke dalam representasi numerik sebagai berikut: *'Normal'* menjadi 0, *'Elevated'* menjadi 1, *'Hypertension Stage 1'* menjadi 2, dan *'Hypertension Stage 2'* menjadi 3. Transformasi ini penting untuk memastikan bahwa model dapat mengenali dan mengolah data secara matematis dalam proses pelatihan.

Setelah proses transformasi data kategorikal, dilakukan normalisasi terhadap atribut numerik seperti '*height', 'weight', 'ap_hi', 'ap_lo', 'age',* dan *'bmi'* menggunakan metode *Min-Max Scaling*. Proses ini bertujuan untuk mengubah nilai-nilai pada masing-masing atribut ke dalam rentang 0 hingga 1, sehingga semua fitur memiliki skala yang seragam. Normalisasi ini penting karena algoritma seperti *K-Nearest Neighbors* (KNN) sangat sensitif terhadap skala data dikarenakan fitur dengan skala yang lebih besar dapat mendominasi perhitungan jarak (*distance*), yang berpotensi menurunkan performa model.

Sebagai bagian dari tahap prapengolahan, dilakukan penghapusan terhadap tiga atribut yaitu *'bp_category_encoded', 'id',* dan *'age'* dalam satuan hari. Atribut *'bp_category_encoded'* dihapus karena hanya digunakan sebagai representasi sementara dari kategori tekanan darah dan tidak diperlukan lebih lanjut dalam proses pelatihan model. Atribut *'id'* dihapus karena bersifat unik untuk setiap entri dan tidak memiliki kontribusi terhadap prediksi, bahkan berpotensi menimbulkan *overfitting*. Sementara itu, *'age'* dalam satuan hari dihapus karena sudah ada atribut yang telah dikonversi ke dalam bentuk yang lebih mudah dipahami, yaitu usia dalam tahun. Atribut *'age_years'* yang merupakan hasil konversi dari *'age'* dalam satuan hari telah diubah labelnya menjadi *'age'*.

Langkah selanjutnya adalah melakukan pemilihan atribut (*feature selection*) untuk memastikan bahwa model yang dibangun menggunakan fitur-fitur yang memiliki relevansi tinggi terhadap risiko penyakit kardiovaskular. Pemilihan atribut ini didasarkan pada bukti ilmiah dari berbagai studi yang menyoroti faktor-faktor risiko utama penyakit kardiovaskular.
Atribut yang digunakan dalam dataset mencakup:

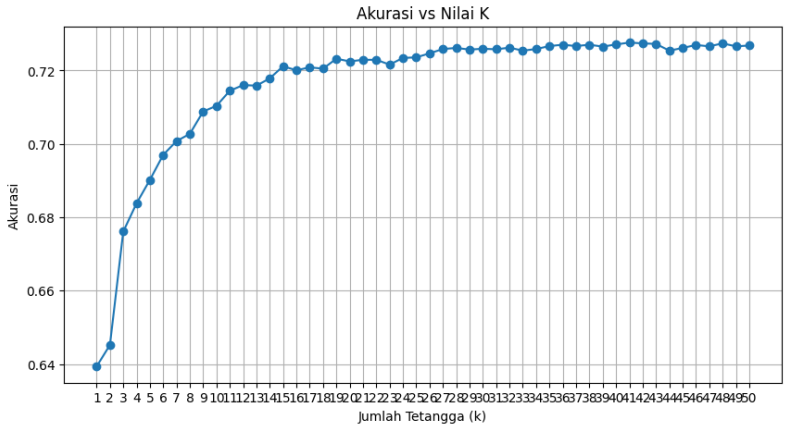
1. Jenis kelamin (gender): Perempuan lanjut usia dilaporkan memiliki risiko penyakit kardiovaskular (CVD) yang lebih tinggi dibandingkan pria seusia [16].
2. Tekanan darah tinggi (ap_hi, ap_lo, bp_category): Tekanan darah tinggi menjadi penyebab utama kematian akibat penyakit kardiovaskular di seluruh dunia. Pada tahun 2021, tekanan darah tinggi menyebabkan sekitar 10,8 juta kematian akibat penyakit kardiovaskular [17].
3. Kolesterol (cholesterol): Tingkat LDL-C (kolesterol jahat) yang tinggi menjadi faktor risiko untuk penyakit kardiovaskular aterosklerotik. Pada tahun 2021, 3,81 juta kematian akibat penyakit kardiovaskular disebabkan oleh tingkat LDL-C yang tinggi [17].
4. Indeks Massa Tubuh (bmi, height, weight): Obesitas terkait erat dengan risiko penyakit jantung dan pembuluh darah, yang menyebabkan 1,95 juta kematian pada 2021 [17].
5. Usia (age): Usia lanjut merupakan faktor risiko *non-modifieable* terhadap penyakit kardiovaskular (CVD), dengan peningkatan risiko signifikan setelah usia 50 tahun [16].
6. Gula darah (gluc): Kadar glukosa tinggi berhubungan erat dengan risiko diabetes, obesitas, dan penyakit jantung, yang menyebabkan 2,3 juta kematian akibat penyakit kardiovaskular pada 2021 [17].
7. Gaya hidup (smoke, alco, active): Gaya hidup tidak sehat seperti merokok, konsumsi alkohol berlebih, dan kurang aktivitas fisik merupakan kontributor utama penyakit kardiovaskular. Pada 2021, merokok menyebabkan sekitar 3 juta kematian kardiovaskular. Konsumsi alkohol yang tinggi (lebih dari 100 g/minggu) menyebabkan 0,4 juta kematian kardiovaskular, dengan risiko meningkat akibat hipertensi dan aritmia. Rendahnya aktivitas fisik juga menyumbang 0,4 juta kematian kardiovaskular [17].

Setelah seluruh tahapan prapengolahan data dilakukan, termasuk penghapusan duplikasi, penanganan outlier, transformasi data kategorikal, normalisasi atribut numerik, dan pemilihan atribut, dataset akhir yang siap digunakan untuk pelatihan model terdiri dari 64.898 entri dan 14 atribut. Dataset ini sudah berada dalam format yang bersih, konsisten, dan siap untuk memasuki tahap selanjutnya yaitu pemodelan dan evaluasi performa algoritma.

### 3.2. Pelatihan Model

Sebelum pelatihan model, dataset dibagi menjadi data latih dan data uji dengan proporsi 80:20, menghasilkan 52.918 data latih dan 12.980 data uji. Setelah itu, dilakukan pelatihan model dengan metode *K-Nearest Neighbors* (KNN). KNN dipilih karena merupakan algoritma sederhana namun efektif untuk masalah klasifikasi, yang bekerja dengan cara mengidentifikasi kedekatan antara data baru dengan data yang sudah ada.

Pelatihan model dilakukan dengan eksperimen untuk menentukan nilai k yang memberikan akurasi tertinggi. Eksperimen dilakukan dengan mencoba berbagai nilai k mulai dari k = 1 hingga k = 51. Hasil eksperimen menunjukkan bahwa nilai k = 41 memberikan akurasi tertinggi, yaitu sebesar 73%.




Gambar 3. Grafik Akurasi vs Nilai *k*

Berikut adalah hasil classification report menggunakan nilai k = 41:

Tabel 2. *Classification Report*

| Classification Report | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| **0** | 0.70 | 0.77 | 0.73 | 6365 |
| **1** | 0.76 | 0.68 | 0.72 | 6615 |
| **Accuracy** | | | 0.73 | 12980 |
| **Macro Avg** | 0.73 | 0.73 | 0.72 | 12980 |
| **Weighted Avg** | 0.73 | 0.73 | 0.72 | 12980 |

Hasil di atas menunjukkan bahwa model dengan nilai *k* = 41 memiliki akurasi sebesar 73%. *Precision, recall, dan f1-score* untuk masing-masing kelas juga menunjukkan performa yang cukup baik, dengan kelas 0 (*No CVD Risk*) memiliki *recall* lebih tinggi, sementara kelas 1 (*CVD Risk*) memiliki *precision* yang lebih tinggi.

### 3.3. Evaluasi Model

Pada tahap ini, dilakukan evaluasi terhadap performa model klasifikasi KNN dalam memprediksi ada atau tidaknya penyakit kardiovaskular. Evaluasi dilakukan dengan menggunakan *confusion matrix*. *Confusion matrix* digunakan untuk memetakan kinerja suatu algoritma klasifikasi dalam bentuk tabel yang menggambarkan hubungan antara prediksi model dan kondisi aktual data.

Melalui *confusion matrix,* nilai akurasi dapat ditentukan dengan pembagian jumlah data yang telah diklasifikasikan secara benar dengan total sampel data yang diuji. Berikut adalah persamaan untuk menghitung akurasi menggunakan confusion matrix:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \text{ x } 100\% \quad (1)$$

Keterangan:

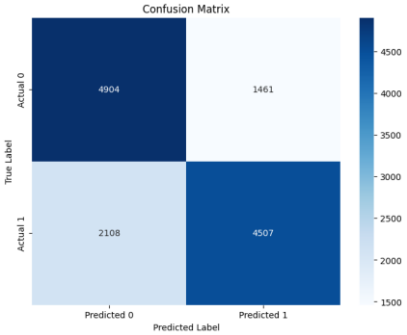
- *True Positive* (TP): jumlah data positif yang berhasil diklasifikasikan dengan benar sebagai positif
- *False Positive* (FP): jumlah data negatif yang keliru diklasifikasikan sebagai positif
- *False Negative* (FN): jumlah data positif yang salah diklasifikasikan sebagai negatif
- *True Negative* (TN): jumlah data negatif yang berhasil diklasifikasikan dengan benar sebagai negatif.

Tabel 3. *Confusion Matrix*

| | *Predicted Negative* (0) | *Predicted Positive* (1) |
|---|---|---|
| *Actual Negative* (0) | *True Negative* (TN) | *False Positive* (FP) |
| *Actual Positive* (1) | *False Negative* (FN) | *True Positive* (TP) |

Gambar 2 menunjukkan *confusion matrix* dari hasil evaluasi model *K-Nearest Neighbors* (KNN) dalam memprediksi risiko penyakit kardiovaskular. Dari visualisasi tersebut, diperoleh hasil sebagai berikut::

- *True Negative* (TN): jumlah data negatif yang berhasil diklasifikasikan dengan benar sebagai negatif, yaitu 4904 data
- *False Positive* (FP): jumlah data negatif yang keliru diklasifikasikan sebagai positif, yaitu 1461 data
- *False Negative* (FN): jumlah data positif yang salah diklasifikasikan sebagai negatif, yaitu 2.108 data
- *True Positive* (TP): jumlah data positif yang berhasil diklasifikasikan dengan benar sebagai positif, yaitu 4.507 data




Gambar 4. *Confusion Matrix*

Berdasarkan gambar *confusion matrix* pada Gambar 4, nilai akurasi dari model klasifikasi pada *Cardiovascular Disease Dataset* dengan menggunakan KNN dengan nilai $k$ = 41 adalah 73% yang bisa dilihat pada tabel 2.

### 3.4. Pembahasan Evaluasi Model

Evaluasi model K-Nearest Neighbors (KNN) dengan nilai k = 41 menunjukkan performa yang cukup baik dalam mengklasifikasikan risiko penyakit kardiovaskular. Berdasarkan *classification report*, model mencapai akurasi sebesar 73%, dengan nilai *precision* sebesar 0.70 untuk kelas 0 (tanpa risiko CVD) dan 0.76 untuk kelas 1 (berisiko CVD). Nilai *recall* untuk kelas 0 adalah 0.77, menunjukkan bahwa sebagian besar individu tanpa risiko berhasil dikenali dengan benar, sedangkan *recall* untuk kelas 1 sebesar 0.68 menunjukkan bahwa masih terdapat sejumlah individu berisiko yang tidak terdeteksi oleh model. Nilai *f1-score* yang seimbang antara kedua kelas (masing-masing sekitar 0.72–0.73) menunjukkan keseimbangan antara *precision* dan *recall*. Rata-rata makro (*macro avg*) dan rata-rata berbobot (*weighted avg*) dari ketiga metrik tersebut semuanya berada di angka 0.73, yang menunjukkan kestabilan kinerja model secara keseluruhan.

Analisis lebih lanjut menggunakan confusion matrix memberikan gambaran yang lebih rinci mengenai performa prediksi model. Sebanyak 4.904 data tanpa risiko kardiovaskular berhasil diklasifikasikan dengan benar sebagai tidak berisiko, yang termasuk dalam kategori *True Negative* (TN). Sementara itu, terdapat 1.461 data tanpa risiko yang justru salah diklasifikasikan sebagai berisiko, yang disebut sebagai *False Positive* (FP). Untuk data yang memang memiliki risiko kardiovaskular, sebanyak 4.507 data berhasil dikenali dengan benar sebagai berisiko atau *True Positive* (TP). Namun, masih terdapat 2.108 data berisiko yang keliru diprediksi sebagai tidak berisiko, yang termasuk dalam kategori *False Negative* (FN).

Secara keseluruhan, model menunjukkan performa yang cukup baik dengan akurasi sebesar 73%, serta keseimbangan antara precision dan recall pada kedua kelas. Hasil ini menunjukkan bahwa model KNN mampu memberikan prediksi yang cukup andal dalam mengklasifikasikan risiko kardiovaskular, dan dapat dijadikan dasar untuk pengembangan lebih lanjut dalam konteks deteksi awal penyakit berbasis data kesehatan.

## 4. KESIMPULAN

Berdasarkan hasil penelitian, dapat disimpulkan bahwa algoritma *K-Nearest Neighbors* (KNN) efektif dalam memprediksi risiko penyakit kardiovaskular dengan akurasi sebesar 73%. Penelitian ini menggunakan dataset dari Kaggle yang mencakup lebih dari 68.000 entri dengan berbagai atribut terkait faktor risiko. Tahapan pengolahan data, seperti penghapusan duplikat dan normalisasi, terbukti penting untuk meningkatkan kualitas input yang digunakan oleh algoritma. Atribut seperti tekanan darah, kolesterol, BMI, serta kebiasaan hidup seperti konsumsi alkohol dan merokok memiliki pengaruh signifikan terhadap hasil prediksi.

Seperti yang telah diharapkan pada bab Pendahuluan, hasil evaluasi model menunjukkan bahwa algoritma KNN mampu mengklasifikasikan pasien dengan baik meskipun masih terdapat potensi perbaikan dalam mengurangi kesalahan klasifikasi pada kategori tertentu. Pengukuran *akurasi*, *precision*, *recall*, dan *f1-score* mengindikasikan bahwa meskipun model ini efektif, terdapat peluang untuk optimasi lebih lanjut pada beberapa kategori data yang sulit diklasifikasikan.

Ke depan, hasil penelitian ini dapat dikembangkan dengan menambah fitur-fitur yang lebih beragam atau mencoba algoritma lain yang lebih kompleks untuk meningkatkan akurasi. Penelitian selanjutnya juga dapat mengarah pada pengembangan aplikasi berbasis mobile atau perangkat wearable untuk deteksi dini risiko kardiovaskular secara real-time. Dengan demikian, berdasarkan hasil dan pembahasan dalam penelitian ini, penerapan algoritma KNN dalam prediksi risiko penyakit kardiovaskular menunjukkan potensi besar, dan studi lebih lanjut di bidang ini dapat membawa manfaat signifikan dalam upaya pencegahan penyakit kardiovaskular.

## UCAPAN TERIMA KASIH

Kami ingin mengucapkan terima kasih kepada semua pihak yang telah mendukung dan berkontribusi dalam penelitian ini. Terima kasih kami sampaikan kepada pembimbing yang telah memberikan arahan, bimbingan, dan masukan yang sangat berharga sepanjang proses penelitian ini. Terima kasih juga kepada teman-teman dan keluarga yang telah memberikan dukungan moral serta semangat selama pengerjaan tugas ini.

Kami juga ingin mengucapkan terima kasih kepada seluruh pihak yang telah menyediakan data dan referensi yang digunakan dalam penelitian ini, serta kepada para ahli yang karya-karyanya telah memberikan inspirasi dan pengetahuan yang mendalam. Semoga hasil dari penelitian ini dapat bermanfaat bagi perkembangan ilmu pengetahuan dan teknologi, khususnya dalam bidang kesehatan.

**REFERENSI**


[1] B. Prabowo, A. Albar, and R. Salim, "Optimalisasi Kesadaran Kesehatan Warga Desa Sarirogo dengan Sosialisasi Hidup Sehat dan Implementasi Medical Check-up,", Fundamentum: Jurnal Pengabdian Multidisiplin, Vol. 2, No. 3, PP. 70–77, Agustus 2024.

[2] R. H. Cahya, A. M. Riadhino, N. R. Adityama and T. L. M. Suryanto, "Sinergi AI Dan Machine Learning Untuk Prediksi Multikeluhan Pada Diagnosis Penyakit Kepala,", JATI: Jurnal Mahasiswa Teknik Informatika, Vol. 9, No. 1, PP. 1469–1476, Februari 2025.

[3] Raspiyahni and Susilawati, "Penerapan Teknologi AI Untuk Meningkatkan Kesehatan Dan Keselamatan Kerja Di Industri Manufaktur,", GJMI: Gudang Jurnal Multidisiplin Ilmu, Vol. 2, No. 6, PP. 651–655, Juni 2024.

[4] M. Martiningsih and A. Haris, "Risiko Penyakit Kardiovaskuler Pada Peserta Program Pengelolaan Penyakit Kronis (Prolanis) Di Puskesmas Kota Bima: Korelasinya Dengan *Ankle Brachial Index* Dan Obesitas,", Jurnal Keperawatan Indonesia, Vol. 22, No. 3, PP. 200–208, September 2019.

[5] D. N. Hafila, Wisudawan, S. Darma, Nurhikmah and Dahlia, "Prevalensi Penyakit Kardiovaskular pada Masa Pandemic Tahun 2020-2021 di RS Arifin Nu'mang Kabupaten Sidrap,", FAKUMI: Jurnal Mahasiswa Kedokteran, Vol. 3, No. 10, PP. 17–28, Oktober 2023.

[6] World Health Organization, "*Cardiovascular diseases (CVDs),*", *World Health Organization: WHO*, 11 Juni 2021, [Online]. Tersedia: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) [Diakses: 13 April 2025].

[7] J. Pane, L. Simorangkir, and P. Saragih, "Faktor-Faktor Risiko Penyakit Kardiovaskular Berbasis Masyarakat,", JPPP: Jurnal Penelitian Perawat Profesional, Vol. 4, No. 4, PP. 1183–1192, September 2022.

[8] A. F. Umara, "Deteksi Dini Penyakit Jantung dan Pembuluh Darah Pegawai,", Media Karya Kesehatan, Vol. 3, No. 2, PP. 122–133, November 2020.

[9] I. Akbar, F. Supriadi, and D. I. Junaedi, "Pemanfaatan Machine Learning Di Bidang Kesehatan,", JATI: Jurnal Mahasiswa Teknik Informatika, Vol. 9, No. 1, PP. 1744–1749, Februari 2025.

[10] S. Dharmawan, V. Fernandes, and H. Halim, "Prediksi Serangan Jantung Dengan Menggunakan Metode Logistic Regression Classifier Dan Adaboost,", Computatio: Journal of Computer Science and Information Systems, Vol. 10, No. 3, PP. 96–103, Januari 2024.

[11] I. K. A. Jayaditya and I. G. A. G. Kadyanan, "Implementasi Random Forest pada Klasifikasi Penyakit Kardiovaskular dengan Hyperparameter Tuning Grid Search,", Jurnal Nasional Teknologi Informasi dan Aplikasinya, Vol. 2, No. 1, PP. 219–226, November 2023.

[12] V. Artanti, M. Faisal, and F. Kurniawan, "Klasifikasi *Cardiovascular Diseases* Menggunakan Algoritma *K-Nearest Neighbors* (KNN),", Techno.COM, Vol. 23, No. 2, PP. 467–479, Mei 2024.

[13] Aidan, "*Cardiovascular Disease,*", *Kaggle*, 2023, [Online]. Tersedia: https://www.kaggle.com/datasets/colewelkins/cardiovascular-disease/data [Diakses: 13 April 2025].

[14] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, Edisi 3. United Kingdom: Elsevier Science, 2011.

[15] S. Sathyanarayanan and B. R. Tantri, "Confusion Matrix-Based Performance Evaluation Metrics,", AJBR: African Journal of Biomedical Research, Vol. 27, No. 4s, PP. 4023–4031, November 2024.

[16] J. L. Rodgers, J. Jones, and S. I. Bolledu, "Cardiovascular Risks Associated with Gender and Aging,", Journal of Cardiovascular Development and Disease, Vol. 6, No. 2, PP. 1–18, April 2019.

[17] M. Vaduganathan, G. Mensah, and J. V. Turco, "The Global Burden of Cardiovascular Diseases and Risk: A Compass for Future Health,", JACC: Journals of the American College of Cardiology, Vol. 80, No. 25, PP. 2361–2371, December 2022.