

KLASIFIKASI RISIKO PENYAKIT JANTUNG BERDASARKAN FAKTOR GAYA HIDUP MENGUNAKAN ALGORITMA KNN

Muhamad Fadli Husna Mubarak¹, Ilamsyah², Raka Danuraya³, Olivia Nurrahmadita³

^{1,3,4}

Ilmu Komputer, Fakultas Teknologi dan Bisnis, Universitas Yatsi Madani

² Sistem Komputer, Fakultas Sains dan Teknologi, Universitas Raharja

E-mail: ¹fadqwe20@gmail.com, ²ilamsyah@raharja.info, ³danuray28@gmail.com,

⁴viiaa 30@gmail.com

ABSTRAK

Penyakit jantung merupakan salah satu penyebab utama kematian di dunia dan berkaitan dengan berbagai faktor gaya hidup serta karakteristik individu. Pemilihan topik adalah karena identifikasi risiko penyakit jantung jarang dilakukan sejak dini. Disamping itu topik ini penting untuk mendukung upaya pencegahan dan meningkatkan kesadaran terhadap kondisi kesehatan. Penelitian ini bertujuan mengklasifikasikan risiko penyakit jantung berdasarkan faktor gaya hidup dan karakteristik individu menggunakan algoritma *K-Nearest Neighbors* (KNN). Penelitian menggunakan metode kuantitatif dengan pendekatan eksperimen melalui pengujian beberapa nilai K dan perbandingan hasil klasifikasi. Data sekunder diperoleh dari dataset *Heart Disease Prediction* pada Kaggle. Dari sekitar 5.000 data, dipilih 500 data secara acak dengan distribusi kelas seimbang sebagai data latih dan 10 data sebagai data uji. Tahapan penelitian meliputi pembersihan data, *discretization* usia, *encoding* fitur kategorikal, penerapan KNN, dan evaluasi hasil klasifikasi. Fitur yang digunakan meliputi usia, jenis kelamin, status merokok, konsumsi alkohol, riwayat stroke, dan gangguan berjalan. Pengujian menggunakan K = 3, 5, dan 7 dengan *Euclidean Distance*. Hasil menunjukkan KNN mampu melakukan klasifikasi dengan hasil relatif konsisten, meskipun terdapat perbedaan pada beberapa data uji. Penentuan nilai K yang sesuai penting untuk memperoleh hasil klasifikasi yang lebih baik. Penelitian ini menunjukkan potensi KNN sebagai pendekatan awal skrining risiko penyakit jantung.

Kata Kunci: penyakit jantung, gaya hidup, KNN, klasifikasi, usia

ABSTRACT

Heart disease is one of the leading causes of death worldwide and is associated with various lifestyle factors and individual characteristics. The topic was selected because early identification of heart disease risk is rarely conducted. In addition, this topic is important for supporting preventive efforts and increasing awareness of health conditions. This study aims to classify heart disease risk based on lifestyle factors and individual characteristics using the K-Nearest Neighbors (KNN) algorithm. The study employed a quantitative method with an experimental approach by testing several K values and comparing the classification results. Secondary data were obtained from the Heart Disease Prediction dataset on Kaggle. From approximately 5,000 records, 500 records were randomly selected with a balanced class distribution as training data, while 10 records were used as testing data. The research stages included data cleaning, age discretization, categorical feature encoding, KNN implementation, and classification evaluation. The features included age, gender, smoking status, alcohol consumption, history of stroke, and difficulty walking. Testing was conducted using K = 3, 5, and 7 with Euclidean Distance. The results showed that KNN was able to perform classification with relatively consistent results, although differences were observed in several testing data points. Appropriate selection of K is important for achieving better classification results. This study demonstrates the potential of KNN as an initial approach for heart disease risk screening.

Keywords: Heart Disease, Lifestyle, KNN, Classification, Age

PENDAHULUAN

Penyakit jantung merupakan salah satu penyebab utama kematian di seluruh dunia. Menurut data Organisasi Kesehatan Dunia (WHO), lebih dari 17 juta orang meninggal setiap tahunnya akibat penyakit kardiovaskular (World Health Organization, 2023). Tingginya angka kematian tersebut menunjukkan bahwa penyakit jantung masih menjadi permasalahan kesehatan yang serius. Permasalahan mendasar atau urgensi yang melatarbelakangi penelitian ini adalah risiko penyakit jantung dapat dipengaruhi oleh berbagai faktor gaya hidup dan karakteristik individu, sementara identifikasi risiko sejak dini belum selalu dilakukan secara optimal. Kondisi tersebut dapat menyebabkan faktor-faktor risiko tidak dikenali lebih awal sehingga upaya pencegahan menjadi kurang optimal. Oleh karena itu, diperlukan suatu pendekatan yang dapat membantu mengidentifikasi dan mengklasifikasikan risiko penyakit jantung berdasarkan faktor-faktor yang mudah diperoleh dari karakteristik dan gaya hidup individu.

Faktor gaya hidup seperti kebiasaan merokok, pola makan tidak sehat, kurangnya aktivitas fisik, konsumsi alkohol, serta stres kronis merupakan beberapa faktor yang berkaitan dengan risiko penyakit jantung. Selain faktor gaya hidup, karakteristik individu seperti usia dan jenis kelamin juga dapat digunakan sebagai informasi dalam proses identifikasi risiko. Dengan berkembangnya teknologi dan ketersediaan data, pendekatan berbasis kecerdasan buatan, khususnya *machine learning*, telah banyak diterapkan untuk memprediksi dan mengklasifikasikan risiko penyakit berdasarkan data individu.

Penelitian mengenai *machine learning* untuk prediksi risiko penyakit kardiovaskular telah dilakukan oleh (Liu et al., 2025). Studi ini merupakan *systematic review and meta-analysis* terhadap penelitian *machine learning* untuk prediksi risiko CVD dan membandingkan performa berbagai model. Salah satu algoritma klasifikasi yang umum digunakan dalam *machine learning* adalah *K-Nearest Neighbors* (KNN). Algoritma KNN memproses data dengan mengukur kedekatan antar titik data menggunakan metode jarak tertentu, seperti *Euclidean Distance* (Sitompul et al., 2025). KNN dikenal karena konsepnya yang sederhana dan kemampuannya dalam melakukan klasifikasi berdasarkan kedekatan karakteristik data. Pendekatan tersebut dapat diterapkan pada data faktor gaya hidup dan karakteristik individu untuk membantu mengelompokkan tingkat risiko penyakit jantung.

Berdasarkan permasalahan tersebut, penelitian ini bertujuan untuk mengimplementasikan algoritma *K-Nearest Neighbors* dalam mengklasifikasikan risiko penyakit jantung berdasarkan faktor gaya hidup dan karakteristik individu. Penelitian ini juga bertujuan untuk mengevaluasi hasil klasifikasi KNN dengan menggunakan beberapa variasi nilai K. Hasil penelitian diharapkan dapat menjadi referensi awal dalam pengembangan sistem deteksi atau skrining risiko penyakit jantung secara otomatis berbasis data.

METODE PENELITIAN

Data yang digunakan dalam penelitian ini merupakan data sekunder yang diperoleh dari dataset *Heart Disease Prediction* yang tersedia pada platform *Kaggle*. Pengumpulan data dilakukan dengan metode observasi dokumentasi, yaitu mengamati dan mengambil data yang telah tersedia pada dataset publik untuk digunakan sebagai bahan penelitian. Dataset tersebut memuat informasi mengenai karakteristik individu dan faktor yang berkaitan dengan kondisi kesehatan, antara lain usia, jenis kelamin, status merokok, konsumsi alkohol, riwayat stroke, gangguan berjalan, serta label risiko penyakit jantung.

Dataset awal yang digunakan terdiri dari sekitar 5.000 data. Selanjutnya, peneliti melakukan seleksi data dengan memilih 500 data secara acak dan memastikan distribusi kelas tetap seimbang, yaitu 250 data dengan label terkena penyakit jantung dan 250 data dengan label tidak terkena penyakit jantung. Data yang telah dipilih kemudian diproses melalui tahapan data *preprocessing* sebelum digunakan dalam proses klasifikasi menggunakan algoritma KNN. Pemilihan data secara seimbang dilakukan untuk mengurangi potensi ketimpangan kelas (class imbalance) dalam proses klasifikasi.

Beberapa cara digunakan untuk pengukuran jarak antara data baru (*data testing*) dengan data lama (*data training*), antara lain manhattan distance (*city block distance*) dan *euclidean distance*. Pengukuran jarak yang paling sering digunakan adalah *euclidean distance*. Jarak *euclidean distance* didefinisikan seperti berikut:

$$d(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

Keterangan:

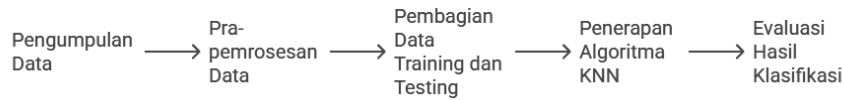
- $d(x, y)$ = Jarak euclidean antara data x dan y
- x_i dan y_i = Nilai dari fitur ke-I pada data x dan data y
- n = Jumlah total fitur yang digunakan dalam perhitungan

Algoritma KNN memiliki beberapa kelebihan, diantaranya ketangguhan terhadap data training yang mengandung banyak noise serta kemampuan menangani data dalam jumlah besar. Namun, algoritma ini juga memiliki beberapa kelemahan, seperti perlunya menentukan jumlah tetangga terdekat (parameter K) dari data target. Penentuan nilai K yang terbaik untuk algoritma ini berdasarkan pada data yang ada. Nilai K yang tinggi dapat mengurangi efek noise pada klasifikasi (Cholil et al., 2021).

Algoritma K-Nearest Neighbor (KNN) merupakan salah satu metode *supervised learning* yang digunakan untuk mengklasifikasikan suatu objek berdasarkan kedekatannya terhadap data latih (training data) serta kelas yang telah ditentukan (Guna & Niska, n.d.). Algoritma ini bekerja dengan cara mengklasifikasikan instance baru berdasarkan mayoritas kelas dari sejumlah tetangga terdekat (nilai K) dalam data latih. KNN banyak digunakan dalam penelitian kesehatan karena kemampuannya dalam melakukan prediksi berbasis kedekatan antar data. Pengumpulan data dalam penelitian ini dilakukan melalui metode observasi terhadap dataset sekunder yang tersedia di situs Kaggle. Dataset tersebut berisi data gaya hidup dan indikator kesehatan individu yang dapat dijadikan dasar dalam mengklasifikasikan tingkat risiko penyakit jantung.

1. Alur Penelitian

Gambar 1 berikut menunjukkan tahapan utama dalam proses penelitian ini. Penelitian dimulai dengan pengumpulan dataset dari situs *Kaggle*, kemudian dilanjutkan dengan proses pra-pemrosesan data untuk menyiapkan data agar dapat digunakan oleh algoritma KNN. Setelah itu, data dibagi menjadi data latih (*training data*) dan data uji (*testing data*). Algoritma KNN kemudian diterapkan untuk melakukan proses klasifikasi terhadap data uji berdasarkan kedekatannya dengan data latih. Langkah terakhir adalah evaluasi hasil klasifikasi menggunakan metrik akurasi untuk menilai kinerja algoritma.

Alur Penelitian Klasifikasi KNN

Gambar 1. Alur penelitian

2. Dataset

Dataset ini berisi informasi mengenai gaya hidup dan kondisi kesehatan responden, yang dapat digunakan untuk memprediksi risiko penyakit jantung. Beberapa fitur utama yang digunakan dalam penelitian ini antara lain:

- Usia
- Jenis Kelamin
- Status Merokok
- Konsumsi Alkohol
- Riwayat Stroke
- Gangguan Berjalan

Seluruh fitur telah dikonversi ke dalam bentuk kategorikal numerik biner (misalnya ya = 1, tidak = 0) atau ordinal (usia dalam rentang tertentu diberi kode 1–13) untuk memudahkan proses klasifikasi menggunakan algoritma KNN. Meskipun dataset asli terdiri dari sekitar 5000 data, dalam penelitian ini digunakan sebanyak 500 data yang dipilih secara acak dan seimbang, yaitu 250 data dengan label “terkena penyakit jantung” dan 250 data dengan label “tidak terkena penyakit jantung”. Pendekatan ini bertujuan untuk menghindari ketimpangan kelas (*class imbalance*) yang dapat memengaruhi kinerja klasifikasi.

Teknik Analisis Data

Teknik analisis data dalam penelitian ini dilakukan secara bertahap untuk mengolah data dan mengetahui kemampuan algoritma *K-Nearest Neighbors* (KNN) dalam mengklasifikasikan risiko penyakit jantung. Tahapan analisis data yang dilakukan adalah sebagai berikut:

1. Analisis Deskriptif

Analisis deskriptif dilakukan untuk mengetahui karakteristik awal dataset yang digunakan dalam penelitian. Analisis meliputi jumlah data, distribusi kelas, serta karakteristik setiap fitur yang digunakan. Dataset awal terdiri dari sekitar 5.000 data, kemudian dipilih 500 data dengan distribusi kelas yang seimbang, yaitu 250 data dengan label terkena penyakit jantung dan 250 data dengan label tidak terkena penyakit jantung. Fitur yang dianalisis meliputi usia, jenis kelamin, status merokok, konsumsi alkohol, riwayat stroke, dan gangguan berjalan.

2. Pra-Pemrosesan Data

Sebelum dilakukan klasifikasi menggunakan algoritma *K-Nearest Neighbor* (KNN), data terlebih dahulu melalui tahap pra-pemrosesan untuk memastikan konsistensi dan kesiapan data dalam proses perhitungan jarak antar fitur serta menyederhanakan dan menyiapkan data agar dapat diolah lebih akurat oleh model klasifikasi (Iathifah et al., 2025). Adapun tahapan pra-pemrosesan yang dilakukan adalah sebagai berikut:

a. Pembersihan Data

Langkah ini dilakukan untuk menghapus data yang tidak lengkap, duplikat, atau mengandung nilai yang tidak valid. Hanya data yang lengkap dan relevan yang digunakan dalam proses klasifikasi.

b. Pengelompokan Nilai (*Discretization*)

Beberapa fitur numerik seperti usia dikelompokkan ke dalam interval tertentu dan diberi nilai diskrit. Misalnya, usia 18–24 dikodekan sebagai 1, usia 25–29 sebagai 2, dan seterusnya hingga usia >80 diberi nilai 13.

c. *Encoding* Fitur Kategorikal

Fitur-fitur seperti status merokok, minum alkohol, stroke, gangguan berjalan, dan jenis kelamin dikonversi ke bentuk numerik biner, dengan skema:

Ya = 1, Tidak = 0 (atau Laki-laki = 1, Perempuan = 0).

d. Tanpa Normalisasi

Karena seluruh fitur telah dikonversi ke dalam bentuk kategorikal diskrit (bukan data kontinu), proses normalisasi tidak diperlukan dalam penelitian ini. Hal ini bertujuan untuk menjaga interpretabilitas nilai fitur dan mempercepat perhitungan jarak.

3. Pembagian Data

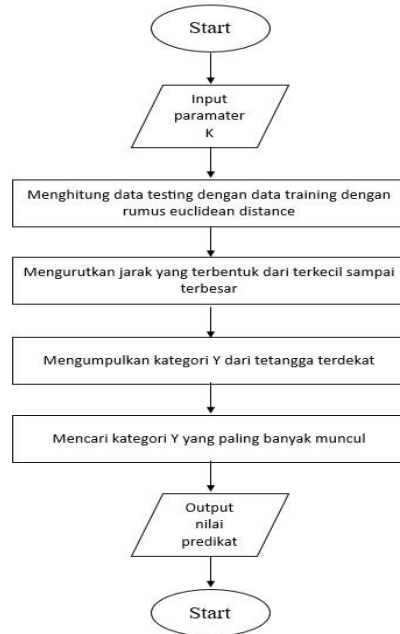
Data yang telah melalui tahap *preprocessing* selanjutnya dibagi menjadi data latih (*training data*) dan data uji (*testing data*). Data latih digunakan sebagai dasar dalam menentukan kedekatan dan kelas data, sedangkan data uji digunakan untuk mengetahui kemampuan KNN dalam melakukan klasifikasi. Dalam penelitian ini digunakan 500 data latih dan 10 data uji. Pembagian data dilakukan secara acak menggunakan fungsi *train_test_split()* dari pustaka *scikit-learn*.

4. Penerapan KNN

K-Nearest Neighbor (KNN) merupakan algoritma klasifikasi berbasis pembelajaran tanpa pelatihan (*lazy learning*), di mana proses klasifikasi dilakukan langsung ketika data uji diberikan, tanpa melalui proses pelatihan model sebelumnya. Algoritma ini bekerja dengan membandingkan data uji dengan seluruh data latih, lalu menentukan kelas berdasarkan mayoritas dari K tetangga terdekatnya. Pengukuran jarak antar data dilakukan menggunakan metode *Euclidean Distance*, yang paling umum digunakan dalam KNN karena kesederhanaannya dan kemampuannya menangani data numerik. Dalam penelitian ini, semua fitur telah dikonversi ke bentuk diskrit (*numerik kategorikal*), sehingga perhitungan jarak tetap dapat dilakukan secara efisien tanpa memerlukan normalisasi tambahan.

5. Eksperimen Nilai K

Setelah seluruh jarak dihitung, model K-NN memilih k data dengan jarak terpendek sebagai tetangga terdekat (Hamidah et al., 2025), lalu menentukan kelas dari data uji berdasarkan mayoritas kelas tetangga tersebut. Nilai K dalam penelitian ini diuji dalam beberapa skenario, yaitu K = 3, 5, dan 7, untuk mengevaluasi pengaruhnya terhadap tingkat akurasi klasifikasi.



Gambar 2. Flowchart algoritma KNN

Implementasi sistem klasifikasi dalam penelitian ini dilakukan menggunakan bahasa pemrograman Python, yang merupakan salah satu bahasa populer dalam bidang analisis data (Sambi Ua et al., 2023). Beberapa pustaka (*library*) pendukung yang digunakan dalam proses implementasi antara lain (Ginting et al., 2026) :

a. *pandas* dan *numpy*

Digunakan untuk manajemen data. Pandas unggul dalam operasi pada data tabular, sedangkan NumPy mengoptimalkan komputasi numerik pada array multidimensi (Japit et al., 2025).

b. *scikit-learn* (*sklearn*)

scikit-Learn adalah *library* Python yang dikembangkan untuk memberikan kemudahan dalam pengkodean machine learning dalam Python, melalui API yang dikembangkan oleh beberapa kontributor dari berbagai negara yang mana biasanya digunakan pada dunia industri dan dunia akademisi (Fahmi, 2024).

c. *matplotlib* dan *seaborn*

Digunakan untuk membuat visualisasi, seperti grafik distribusi data, visualisasi hasil klasifikasi, serta perbandingan akurasi pada berbagai nilai K.

Seluruh proses implementasi dilakukan secara lokal menggunakan perangkat komputer dengan spesifikasi standar, serta editor kode seperti *Jupyter Notebook* atau *Google Colab* untuk kemudahan interaktivitas (Octarina et al., 2025).

HASIL DAN PEMBAHASAN

Bagian ini menyajikan hasil dari penerapan algoritma *K-Nearest Neighbor* dalam klasifikasi risiko penyakit jantung berdasarkan data gaya hidup yang tersedia. Proses analisis dimulai dengan pengolahan data, perhitungan jarak antar data, implementasi sistem menggunakan bahasa pemrograman Python, serta evaluasi performa model menggunakan metrik akurasi.

1. Pengolahan Data

Pengolahan data merupakan tahap penting sebelum dilakukan proses klasifikasi. Pada penelitian ini, pengolahan data dilakukan terhadap 500 data hasil seleksi dari dataset awal sebanyak 5000 data. Data yang digunakan terdiri dari dua kelas yang seimbang, yaitu:

- 250 data yang terkena penyakit jantung (label = 1), dan
- 250 data yang tidak terkena penyakit jantung (label = 0).

Berikut adalah fitur-fitur yang digunakan beserta skema kategorisasinya:

Tabel 1. Fitur dan Kategori

No.	Nama Fitur	Nilai/Kategori	Keterangan
1.	Usia	1-13	Dikelompokkan ke dalam 13 interval usia
2.	Merokok	0 = Tidak, 1 = Ya	Status kebiasaan merokok
3.	Minum Alkohol	0 = Tidak, 1 = Ya	Konsumsi alkohol
4.	Stroke	0 = Tidak, 1 = Ya	Riwayat pernah terkena stroke
5.	Gangguan	0 = Tidak, 1 = Ya	Kesulitan berjalan
6.	Jenis Kelamin	0 = Perempuan, 1 = Laki-laki	Jenis kelamin

Untuk fitur usia, dilakukan pengelompokan sebagai berikut:

- 18–24 tahun = 1
- 25–29 tahun = 2
- 30–34 tahun = 3
- 35–39 tahun = 4
- 40–44 tahun = 5
- 45–49 tahun = 6
- 50–54 tahun = 7
- 55–59 tahun = 8
- 60–64 tahun = 9
- 65–69 tahun = 10
- 70–74 tahun = 11
- 75–79 tahun = 12
- > 80 tahun = 13

Data yang telah dikonversi kemudian dibagi menjadi 500 data latih dan 10 data uji. Pembagian data dilakukan secara acak menggunakan fungsi *train_test_split()* dari pustaka *sklearn.model_selection*.

2. Perhitungan Euclidean Distance

Algoritma *K-Nearest Neighbor* (KNN) bekerja dengan menghitung jarak antara data uji dan seluruh data latih untuk menentukan tetangga terdekat yang akan digunakan dalam klasifikasi. Dalam penelitian ini digunakan 10 data uji yang dibandingkan dengan 500 data latih, dengan menghitung jarak menggunakan metode *Euclidean Distance*. Karena semua fitur telah dikonversi ke bentuk diskrit (misalnya: merokok = 1, tidak = 0; usia dikodekan 1–13 berdasarkan interval usia), maka perhitungan jarak *Euclidean* tetap dapat dilakukan meskipun fitur-fitur tersebut bersifat kategorikal diskrit. Pendekatan ini dianggap cukup baik untuk klasifikasi berbasis data riil yang sudah disederhanakan. Rumus umum untuk menghitung jarak Euclidean antara dua data x dan y adalah sebagai berikut:

$$d(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

- Perhitungan Euclidean Distance

Berikut ini perhitungan jarak antara 3 data uji dengan 1 data latih yang telah dikonversi dan disederhanakan:

Tabel 2. Perhitungan Euclidean Distance

Data	Usia	Merokok	Alkohol	Stroke	Gangguan jalan	Gender	Risiko penyakit jantung
Latih 1	12	1	0	0	1	0	Tinggi
Latih 2	5	1	1	0	0	0	Rendah
Latih 3	6	1	1	0	0	1	Rendah
Uji	3	1	1	0	1	1	?

Hitung jarak antara data uji dan masing-masing data latih:

Jarak ke data Latih 1:

$$\sqrt{(3 - 12)^2 + (1 - 1)^2 + (1 - 0)^2 + (0 - 0)^2 + (1 - 1)^2 + (1 - 0)^2} = \sqrt{83} \approx 9,11$$

Jarak ke data latih 2

$$\sqrt{(3 - 5)^2 + (1 - 1)^2 + (1 - 1)^2 + (0 - 0)^2 + (1 - 0)^2 + (1 - 0)^2} = \sqrt{6} \approx 2,45$$

Jarak ke data Latih 3

$$\sqrt{(3 - 6)^2 + (1 - 1)^2 + (1 - 1)^2 + (0 - 0)^2 + (1 - 0)^2 + (1 - 1)^2}$$

Dari perhitungan tersebut, data uji memiliki tetangga terdekat Latih 2 dan Latih 3, yang keduanya memiliki kelas “Rendah”. Oleh karena itu, data uji diklasifikasikan ke dalam kelas risiko Rendah.

- Penerapan pada Seluruh Dataset

Proses klasifikasi dilakukan terhadap 10 data uji, di mana masing-masing data dibandingkan dengan seluruh 500 data latih. Tujuannya adalah untuk menentukan kelas risiko penyakit jantung dari setiap data uji berdasarkan mayoritas kelas dari tetangga terdekatnya. Nilai K yang digunakan diuji dalam beberapa skenario, yaitu $K = 3$, $K = 5$ dan $K = 7$. Hasil klasifikasi pada setiap nilai K akan dibandingkan berdasarkan akurasi

untuk menentukan nilai K yang paling optimal. Perhitungan dan pemrosesan dilakukan menggunakan fungsi *kneighbors()* dan *KNeighborsClassifier* dari pustaka *sklearn.neighbors*. Proses ini memungkinkan efisiensi yang tinggi meskipun melibatkan ribuan iterasi perhitungan jarak.

3. Implementasi Sistem

Implementasi sistem klasifikasi dilakukan menggunakan bahasa pemrograman Python dengan bantuan beberapa pustaka pendukung, seperti:

- *pandas* dan *numpy* untuk manajemen dan manipulasi data,
- *scikit-learn* untuk implementasi algoritma *K-Nearest Neighbor* (KNN),
- *tabulate* untuk menampilkan data dalam format tabel di terminal,
- *warnings* untuk menghilangkan peringatan non-kritis saat eksekusi.

Dataset yang digunakan telah melalui proses pra-pemrosesan dan dikonversi ke dalam bentuk numerik diskrit. Jumlah data latih yang digunakan adalah 500 data, masing-masing terdiri dari enam fitur utama, yaitu Usia, Jenis kelamin, Merokok, Minum alkohol, Riwayat stroke, Gangguan berjalan. Data latih ini ditampilkan secara ringkas dalam Tabel 3, yang hanya menyajikan sebagian kecil data dari total 500 entri, misalnya entri nomor 1 sampai 13 dan entri ke- 500 sebagai perwakilan struktur dataset. Sisa data tidak ditampilkan dalam dokumen utama maupun lampiran karena keterbatasan ruang dan jumlah yang besar

Tabel 3. Data Latih

No	HeartDisease	Smoking	AlcoholDrinking	Stroke	DiffWalking	Sex	AgeCategory
1	1	1	0	0	1	0	12
2	0	1	1	0	0	0	5
3	0	1	1	0	0	1	6
4	0	1	1	0	0	1	9
5	0	1	1	1	0	1	11
6	0	1	1	0	0	0	13
7	1	1	1	0	0	0	5
8	1	1	1	0	1	1	10
9	1	1	1	0	0	1	12
10	1	1	1	0	0	0	13
11	0	0	1	0	0	0	9
12	0	0	1	0	0	1	10
13	1	0	1	0	0	1	6
500	1	0	0	0	0	1	13

Selanjutnya, digunakan 10 data uji yang juga ditampilkan dalam bentuk tabel

(Tabel 4). Masing-masing data uji akan dihitung jaraknya terhadap seluruh data latih menggunakan rumus *Euclidean Distance*, kemudian ditentukan kelasnya berdasarkan mayoritas dari K tetangga terdekat.

Tabel 4. Data uji

No	Smoking	AlcoholDrinking	Stroke	DiffWalking	Sex	AgeCategory
1	1	1	0	1	1	3
2	1	0	0	0	1	2
3	1	0	0	1	1	12
4	1	0	1	1	0	10
5	1	1	1	1	1	13
6	0	0	0	1	1	13
7	0	0	0	0	0	1
8	1	1	1	1	0	5
9	0	0	1	1	1	4
10	1	1	0	0	1	9

Proses klasifikasi dilakukan menggunakan tiga nilai K yang berbeda, yaitu: K = 3, K = 5, dan K = 7. Hasil klasifikasi dari 10 data uji tersebut ditampilkan dalam tabel 5, 6, dan 7:

Tabel 5. Pengujian dengan K = 3

Data ke	Prediksi
1	Tidak Memiliki Penyakit Jantung
2	Tidak Memiliki Penyakit Jantung
3	Memiliki Penyakit Jantung
4	Memiliki Penyakit Jantung
5	Memiliki Penyakit Jantung
6	Memiliki Penyakit Jantung
7	Tidak Memiliki Penyakit Jantung
8	Memiliki Penyakit Jantung
9	Memiliki Penyakit Jantung
10	Tidak Memiliki Penyakit Jantung

Tabel 6. Pengujian dengan K = 5

Data ke	Prediksi
1	Memiliki Penyakit Jantung
2	Tidak Memiliki Penyakit Jantung
3	Memiliki Penyakit Jantung
4	Memiliki Penyakit Jantung
5	Memiliki Penyakit Jantung
6	Memiliki Penyakit Jantung
7	Tidak Memiliki Penyakit Jantung
8	Tidak Memiliki Penyakit Jantung
9	Memiliki Penyakit Jantung
10	Tidak Memiliki Penyakit Jantung

Tabel 7. Pengujian dengan K = 7

Data ke	Prediksi
1	Memiliki Penyakit Jantung
2	Tidak Memiliki Penyakit Jantung
3	Memiliki Penyakit Jantung
4	Memiliki Penyakit Jantung
5	Memiliki Penyakit Jantung
6	Memiliki Penyakit Jantung
7	Tidak Memiliki Penyakit Jantung
8	Tidak Memiliki Penyakit Jantung
9	Memiliki Penyakit Jantung
10	Memiliki Penyakit Jantung

Seluruh proses perhitungan dan evaluasi klasifikasi dilakukan menggunakan fungsi *Kneighbors Classifier()* dan *kneighbors()* dari pustaka *sklearn.neighbors*, yang memungkinkan efisiensi tinggi meskipun melibatkan ratusan iterasi perhitungan jarak.

4. Evaluasi Metode KNN

Evaluasi terhadap metode klasifikasi *K-Nearest Neighbor* (KNN) dilakukan untuk menilai kemampuan model dalam mengelompokkan data uji ke dalam kategori risiko penyakit jantung. Dalam eksperimen ini, digunakan 10 data uji dan diuji dengan beberapa variasi nilai K , yaitu $K = 3$, $K = 5$, dan $K = 7$, guna mengetahui bagaimana jumlah tetangga memengaruhi hasil klasifikasi. Setiap hasil prediksi dibandingkan secara langsung dengan label aktual yang telah diketahui. Evaluasi dilakukan dengan cara menghitung jumlah prediksi yang sesuai dengan label sebenarnya, yang kemudian menjadi dasar dalam menilai keakuratan model pada masing-masing nilai K . Walaupun tidak menggunakan metode perhitungan metrik seperti *accuracy score*, *precision*, atau *confusion matrix*, pendekatan ini sudah cukup menggambarkan tingkat keberhasilan klasifikasi untuk skala dataset yang digunakan. Model yang menghasilkan jumlah klasifikasi benar terbanyak dianggap memiliki kinerja yang lebih baik dan digunakan sebagai acuan dalam proses klasifikasi akhir.

SIMPULAN

Penelitian ini membahas penerapan algoritma *K-Nearest Neighbor* (KNN) untuk melakukan klasifikasi risiko penyakit jantung berdasarkan faktor gaya hidup dan kondisi kesehatan. Dataset yang digunakan berasal dari situs Kaggle, dengan total 5000 data. Dalam penelitian ini, dipilih 500 data sebagai data latih dan 10 data sebagai data uji, dengan proporsi kelas yang seimbang antara yang terkena dan tidak terkena penyakit jantung. Seluruh fitur dalam dataset telah dikonversi ke bentuk numerik diskrit untuk memudahkan proses klasifikasi. Implementasi sistem dilakukan menggunakan bahasa pemrograman Python dan pustaka *scikit-learn*, dengan pengujian dilakukan pada beberapa nilai K (yaitu $K = 3$, 5 , dan 7) untuk melihat pengaruh jumlah tetangga terhadap hasil klasifikasi.

Berdasarkan hasil pengamatan, metode KNN mampu melakukan klasifikasi dengan hasil yang konsisten dan cukup akurat, meskipun jumlah data uji relatif kecil. Nilai K yang digunakan memberikan variasi hasil yang tidak signifikan, namun tetap menunjukkan bahwa pemilihan K memengaruhi performa model. Dari penelitian ini, dapat disimpulkan bahwa:

1. Algoritma KNN efektif digunakan untuk klasifikasi berbasis data faktor gaya hidup, terutama dengan fitur-fitur yang telah dinormalisasi atau dikonversi ke bentuk numerik diskrit.
2. Pemilihan nilai K yang sesuai menjadi salah satu aspek penting dalam meningkatkan ketepatan klasifikasi.
3. Pra-pemrosesan data, termasuk konversi data kategorikal dan pembersihan data, memberikan kontribusi besar terhadap efektivitas algoritma dalam menghasilkan klasifikasi yang relevan.

DAFTAR PUSTAKA

- Cholil, S. R., Handayani, T., Prathivi, R., & Ardianita, T. (2021). Implementasi Algoritma Klasifikasi K-Nearest Neighbor (KNN) Untuk Klasifikasi Seleksi Penerima Beasiswa. *Indonesian Journal on Computer and Information Technology*, 6(2).
- Fahmi, M. N. (2024). Implementasi Mechine Learning menggunakan Python Library : Scikit-Learn (Supervised dan Unsupervised Learning). *Jurnal Studi Matematika Dan Teknologi*, 1(2).
- Ginting, D. E. L. B., Candiasa, I. M., & Budayana, I. N. (2026). Perbandingan Random Forest Dan Support Vector Machine Dalam Memprediksi Pemilihan Jurusan SMK. *Science and Education Journal*, 5(2).
- Guna, E. A., & Niska, D. Y. (n.d.). IMPLEMENTASI ALGORITMA K-NEAREST NEIGHBOR DALAM SISTEM KLASIFIKASI CALON PENERIMA BEASISWA PROGRAM INDONESIA PINTAR (PIP). *Jurnal Mahasiswa Teknik Informatika*, 9(6).
- Hamidah, A. N., Nugraha, F., & Muzid, S. (2025). Implementasi Data Mining untuk Penentuan Prioritas Perbaikan Jalan Menggunakan Metode Klasifikasi Algoritma K-Nearest Neighbors. *Jurnal Sistem Informasi Dan Teknologi*, 8(1).
- Japit, S., Risyani, Y., Bombongan, C., Selamat, T., & Yuliana. (2025). *Pemanfaatan Pandas, NumPy, Spark, dan TableauPower BI dalam Pengolahan dan Visualisasi Big Data untuk Mendukung Pengambilan Keputusan* . 14(1).
- lathifah, shofi putri, Umam, K., & Handayani, M. R. (2025). EKSPLORASI PENDAPAT PENGGUNA APLIKASI DEEPSEEK DI PLAY STORE DENGAN NATURAL LANGUAGE PROCESSING. *Jurnal Informatika Dan Sistem Informasi*, 6(2).
- Liu, T., Krentz, A., Lu, L., & Curcin, V. (2025). Machine learning based prediction models for cardiovascular disease risk using electronic health records data: systematic review and meta-analysis. *European Heart Journal*, 6(1), 7–22.
- Octarina, S., Puspita, F. M., Yuliza, E., & Indrawati. (2025). PENDAMPINGAN PENGGUNAAN GOOGLE COLAB PADA PEMBELAJARAN PYTHON DAN MACHINE LEARNING BAGI DOSEN MATEMATIKA DI PALEMBANG. *Jurnal Pepadu*, 6(1).
- Sambi Ua, A., Lestriani, D. H., Marpaung, E. S. K., Ong, J., Savinka, M., Nurhaliza, P., & Ningsih, R. Y. (2023). Penggunaan Bahasa Pemrograman Python Dalam Analisis Faktor Penyebab Kanker Paru-Paru. *Jurnal Publikasi Teknik Informatika*, 2(2).
- Sitompul, A. E. N., Yohanna, M., & Silalahi, A. P. (2025). PENERAPAN ALGORITMA K-NEAREST NEIGHBORS DALAM MENGLASIFIKASI PENYAKIT MULTIPLE SCLEROSIS. *Jurnal Manajemen Informatika & Komputerisasi Akuntansi*, 9(2), 307–315.
- World Health Organization. (2023). *Cardiovascular diseases (CVDs)*. WHO.