

Penerapan Algoritma K-Nearest Neighbor (KNN) Untuk Klasifikasi Resiko Penyakit Jantung

Aprillia Wulan Nanda Dari*, Ika Nur Fajri

Fakultas Ilmu Komputer, Program Studi Sistem Informasi, Universitas Amikom Yogyakarta, Yogyakarta
Jl. Ring Road Utara, Ngringin, Condongcatur, Kec. Depok, Kabupaten Sleman, Daerah Istimewa Yogyakarta, Indonesia

Email: ^{1,*}wulanaprillia@students.amikom.ac.id, ²fajri@amikom.ac.id

Email Penulis Korespondensi: wulanaprillia@students.amikom.ac.id

Submitted: 08/10/2024; Accepted: 22/10/2024; Published: 23/10/2024

Abstrak—Penyakit jantung merupakan salah satu penyakit mematikan di dunia, dimana terdapat gangguan fungsi jantung dan pembuluh darah yang menyebabkan nyeri dada, detak jantung tidak teratur, dan kesulitan bernafas. Menurut data World Health Organization (WHO) terdapat 17,9 juta kematian setiap tahunnya akibat penyakit jantung. Kesulitan dalam melakukan klasifikasi penyakit jantung secara akurat dan cepat menjadi suatu permasalahan yang signifikan. Dari permasalahan tersebut, peneliti melakukan penelitian data mining menggunakan algoritma KNN untuk mengklasifikasikan resiko penyakit jantung dengan mengambil data dari website resmi kaggle. Dalam penelitian ini terdapat 4 tahap yaitu pengumpulan data, pembentukan model, evaluasi mode, dan interface prediksi. Dengan menggunakan algoritma KNN hasil analisis didapatkan akurasi 83%, presisi 0,88, recall 0,77 dan f1-score 0,82. Dengan hasil data evaluasi model tersebut menunjukkan bahwa klasifikasi risiko penyakit jantung menggunakan algoritma KNN memiliki performa yang cukup baik. Hasil dari pemodelan tersebut kemudian disajikan dalam bentuk website dengan melakukan deployment model.

Kata Kunci: Data Mining; K-Nearest Neighbor; Klasifikasi Risiko; Model Evaluasi; Penyakit Jantung

Abstract—Heart disease is one of the deadliest diseases in the world, where there is a disruption in the function of the heart and blood vessels that causes chest pain, irregular heartbeat, and difficulty breathing. According to data from the World Health Organization (WHO), there are 17.9 million deaths each year due to heart disease. The difficulty in classifying heart disease accurately and quickly is a significant problem. From this problem, researchers conducted data mining research using the KNN algorithm to classify the risk of heart disease by taking data from the official Kaggle website. In this study, there are 4 stages, namely data collection, model formation, mode evaluation, and prediction interface. By using the KNN algorithm, the analysis results obtained an accuracy of 83%, precision 0.88, recall 0.77 and f1-score 0.82. With the results of the model evaluation data, it shows that the classification of heart disease risk using the KNN algorithm has quite good performance. The results of the modeling are then presented in the form of a website by deploying the model.

Keywords: Data Mining; K-Nearest Neighbor; Risk Classification; Evaluation Model; Heart Disease

1. PENDAHULUAN

Penyakit jantung merupakan kondisi dimana terjadi sebuah kelainan atau gangguan pada fungsi jantung dan pembuluh darah. Gangguan yang menyebabkan nyeri dada, rasa tertekan, detak jantung tidak teratur, kesulitan bernafas, serta bengkak pada kaki dan perut. Penyebab dari terjadinya penyakit jantung bermacam-macam diantaranya stres, kurang beraktivitas, pola makan yang buruk, obesitas, hipertensi, merokok, riwayat keluarga hingga umur seseorang. Gejala penyakit jantung yang dirasakan dapat berbeda-beda, salah satu gejalanya adalah bengkak pada kaki dan perut serta sesak nafas[1].

Menurut data dari World Health Organization (WHO), penyakit jantung adalah salah satu penyebab kematian nomor satu secara global sebanyak 17,9 juta kematian setiap tahunnya. Di Indonesia, penyakit jantung juga menjadi penyebab kematian utama dibuktikan dengan angka kematian yang terus meningkat setiap tahunnya. Kemenkes RI menyatakan pada tahun 2023 angka kematian akibat penyakit jantung mencapai angka 650.000 penduduk pertahun[2].

Masalah utama dalam penanganan penyakit jantung adalah sulitnya melakukan klasifikasi risiko dengan cepat dan akurat. Pada saat ini diagnosis masih dilakukan secara manual bergantung pada gejala, riwayat medis, serta faktor lain seperti usia dan jenis kelamin, belum terdapat metode yang tepat dan efektif untuk melakukan identifikasi secara otomatis. Sangat penting dilakukan klasifikasi risiko untuk memastikan penanganan yang maksimal dan mencegah komplikasi berkelanjutan. klasifikasi dilakukan dengan memanfaatkan teknik Data Mining. Data mining merupakan proses pengolahan data yang dilakukan dengan mengekstrak informasi pada sebuah data. Tujuan dari dilakukannya data mining adalah sebagai sarana untuk menjelaskan, konfirmasi, dan eksplorasi data[3]. Terdapat beberapa metode algoritma klasifikasi dalam data mining, diantaranya Decision Tree, Bayesian Classification, Neural Network, K-Nearest Neighbor (KNN) dan Support Vector Machine (SVM).

Terdapat beberapa penelitian terkait klasifikasi menggunakan teknik data mining. Pada penelitian yang dilakukan oleh Uktupi Nijunniyahayah dkk. (2024) membahas tentang implementasi algoritma KNN untuk memprediksi penjualan alat Kesehatan[4]. Dengan mengidentifikasi tantangan yang dihadapi perusahaan seperti kurangnya stok dan penumpukan barang penelitian ini dapat menganalisis data penjualan yang digunakan untuk memahami kebutuhan alat kesehatan pelanggan. Metode KNN dalam penelitian ini digunakan untuk melakukan prediksi penjualan dengan membagi 3 kriteria diantaranya paling laris, laris dan tidak laris. Data informasi penjualan yang digunakan dalam merupakan data penjualan tahun 2020 hingga 2022. Dengan menerapkan

algoritma KNN dalam melakukan prediksi, perusahaan dapat melakukan pengelolaan barang secara efisien dan dapat terhindar dari kekurangan stok barang yang tidak diinginkan. Dengan hasil akurasi penelitian sebesar 95% menunjukkan bahwa metode yang diimplementasikan dapat menjadi acuan dalam perencanaan penjualan yang baik dimasa yang akan datang. Penelitian ini memiliki kelemahan yang terkait dengan data penelitian. Penelitian yang dilakukan oleh Uktupi Nijunniyah dkk. (2024) kurang dalam memberikan Informasi proses pengumpulan data, metode pembersihan data serta parameter algoritma K-Nearest Neighbor (KNN) dan tidak terdapat informasi mengenai validitas dan reliabilitas data yang digunakan[4].

Penelitian yang dilakukan oleh Rismala dkk. (2023) membahas tentang bagaimana penerapan metode K-nearest Neighbor (KNN) untuk memprediksi tentang penjualan sepeda motor yang terlaris [5]. Tujuan dilakukannya penelitian ini adalah mengatasi peningkatan penjualan sepeda motor pada perusahaan setiap bulannya, menjadikan penting untuk mengetahui jenis sepeda motor apa yang paling diminati oleh konsumen. Perusahaan melakukan prediksi penjualan untuk membedakan antara sepeda motor yang laris dan tidak laris dengan teknik data mining. Data mining digunakan dalam mengumpulkan dan menganalisa data penjualan dengan memahami preferensi konsumen dengan semakin meningkatnya penggunaan sepeda motor sebagai alat transportasi. Hasil dari penelitian yang dilakukan oleh Rismala dkk. (2023) memiliki akurasi sebesar 96,15% dengan menggunakan nilai $k = 5$. Hasil akurasi tersebut menunjukkan bahwa metode atau model yang diimplementasikan dapat membantu dalam mengambil keputusan strategi pemasaran dan menjadi solusi dalam memprediksi penjualan sepeda motor. Terdapat kelemahan dalam penelitian yang dilakukan oleh Rismala dkk. (2023) diantaranya tidak terdapat batasan data yang digunakan, informasi terkait metode penelitian tidak rinci dan variable yang digunakan tidak diinformasikan secara jelas [5].

Penelitian yang dilakukan oleh Maulana Fanyuri (2020) membahas tentang analisis algoritma klasifikasi KNN dalam menghitung akurasi kepuasan pelanggan pada PT. Tirgatra Komunikatama [6]. Metode dalam penelitian ini adalah metode data mining dengan algoritma KNN. penelitian dengan menggunakan data primer yang diperoleh dari hasil kuesioner dan data sekunder yang diperoleh dari observasi lapangan. Dengan membandingkan penelitian-penelitian sebelumnya terkait penggunaan metode tersebut yang memiliki hasil akurasi prediksi maksimal. Dan dengan menggunakan data yang seimbang, penelitian [6], dilakukan untuk mengetahui tingkat akurasi kepuasan pelanggan pada PT. Trigatra Komunikatama. Penelitian yang dilakukan Maulana Fanyuri (2020) memberi acuan perusahaan dalam meningkatkan performa layanan. Hasil akurasi 83.33% menunjukkan bahwa metode ini terbukti efektif dan akurat dalam mengklasifikasi kepuasan pelanggan. Informasi yang didapatkan dari penelitian tersebut dapat digunakan perusahaan untuk meningkatkan tingkat kepuasan pelanggan atau konsumen dengan memahami faktor-faktor yang memiliki kontribusi akan kepuasan pelanggan.

Dengan didasarkan penelitian yang telah dilakukan, klasifikasi data mining penting dilakukan untuk mengetahui pola data untuk diambil menjadi sebuah Keputusan. Untuk meningkatkan pengembangan metode klasifikasi dan prediksi risiko penyakit jantung pada penelitian ini digunakan algoritma K-Nearest Neighbor (KNN). Algoritma KNN merupakan metode pembelajaran mesin yang digunakan untuk klasifikasi data dengan berdasarkan pada kemiripan data baru dengan data latih [7]. Pada penelitian sebelumnya fokus utamanya adalah pada prediksi penjualan atau kepuasan pelanggan, penelitian ini dilakuakn dengan tujuan untuk menerapkan algoritma KNN untuk mengklasifikasi data pasien berdasarkan gejala dan faktor resiko lain seperti umur dan jenis kelamin, sehingga memungkinkan prediksi seseorang terkena penyakit jantung atau tidak. Selain untuk melakukan klasifikasi risiko penelitian ini juga memiliki tujuan untuk mengimplementasikan model prediksi penyakit jantung ke dalam sistem perangkat lunak melalui proses deployment model. Dengan demikian, hasil prediksi dapat diintegrasikan ke dalam sistem klinis yang membantu dalam melakukan diagnosis dan memberikan perawatan yang tepat[8].

2. METODOLOGI PENELITIAN

Pada penelitian ini terdapat empat proses atau tahapan yang dilakukan secara berurutan diantaranya yaitu pengumpulan dataset, pembentukan model, evaluasi mode, dan interface prediksi. Tahapan proses tersebut dapat dilihat pada Gambar 1.. Setiap jalannya alur proses sangat penting untuk dilakukan. Detail penjelasan masing-masing proses akan dijabarkan pada bab ini.



Gambar 1. Alur Metodologi Penelitian

2.1 Pengumpulan Dataset

Tahap awal dalam penelitian ini adalah pengumpulan data, yang akan menjadi dasar untuk dilakukannya analisis dan hasil. Pada penelitian ini data yang digunakan adalah data Heart Disesase Dataset dari platform kaggle, website pengumpulan data sebagai bahan pembelajaran (<https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset>). Didalam dataset tersebut terdapat 1025 data record dan 14 atribut[9].

2.2 Pembentukan Model

Setelah melakukan tahap awal dan data berhasil diperoleh dari website kaggle, tahap berikutnya adalah pembentukan model / penerapan model. Dalam pembentukan model dilakukan juga preprocessing. Preprocessing merupakan Langkah analisis data dan pembelajaran yang melibatkan pembersihan dan persiapan data mentah untuk menghilangkan inkonsistensi, data tidak lengkap dan redundansi data awal[10]. Langkah ini memastikan bahwa data yang akan dianalisis memiliki kualitas yang optimal, yang secara keseluruhan dapat mempengaruhi performa model. Tujuan dari preprocessing ini adalah untuk memastikan bahwa data mentah dapat diolah menjadi data yang matang dan siap digunakan[10][11].

Model yang digunakan dalam penelitian ini adalah algoritma K-Nearest Neighbor (KNN). K-Nearest Neighbor (KNN) merupakan salah satu metode algoritma dalam machine learning yang digunakan untuk mengklasifikasikan data berdasarkan kedekatan dengan tetangga terdekat dalam data pelatihan[4]. Dalam algoritma KNN klasifikasi ketetanggaan digunakan sebagai nilai prediksi[6]. Konsep dasar dari algoritma ini adalah dengan mencari jarak terdekat antar data satu dengan K tetangga terdekatnya dalam data pelatihan[7][15]. Dimana jumlah kelas yang paling banyak dengan dengan jarak terdekat akan menjadi kelas Dimana data tersebut berada[12]. Pembentukan model dilakukan agar dapat membedakan kelas dalam dataset mentah. Dibutuhkan dua jenis data yaitu data training dan data testing yang berasal dari hasil preprocessing[13][14]. Dalam pembentukan model data training dan data testing dibutuhkan dalam proses klasifikasi dengan algoritma KNN.

2.3 Evaluasi Model

Tahap setelah pembentukan model yaitu evaluasi model. evaluasi model dilakukan dengan tujuan untuk melihat keseluruhan model dan meninjau model serta memastikan model berjalan dengan baik hingga mencapai tujuan yang telah ditentukan. Evaluasi pada penelitian ini dilakukan dengan memperhatikan Confusion Matrix untuk mengetahui sejauh mana klasifikasi dapat memprediksi data kelas[7][10].

Tabel 1. Tabel Classifier

	C1	C2
C	TP	FN
C	FP	TN

Tabel 1 merupakan istilah yang digunakan untuk menganalisis kemampuan classifier. Tabel tersebut berisi True Positive (TP), True Negative (TN), False Negative (FN) dan False Positive (FP). Penilaian kinerja didasarkan pada perhitungan rata-rata kinerja seperti berikut [7] :

- Akurasi : sebuah matrik yang menunjukkan persentase prediksi positif dan benar secara keseluruhan. Nilai akurasi yang tinggi menunjukkan bahwa model secara keseluruhan memiliki kemampuan memprediksi data dengan baik[7]. Rumus akurasi adalah pada rumus 1.

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

- Presisi : sebuah matrik yang memberitahukan proporsi prediksi yang sebenarnya positif dan hanya fokus terhadap hal-hal positif. Nilai presisi yang tinggi dengan hasil prediksi yang positif menunjukkan bahwa model harus berhati-hati[7]. Rumus presisi adalah pada rumus 2.

$$\text{Presisi} = \frac{TP}{TP+FP} \quad (2)$$

- Recall : sebuah matrik yang melihat sisi positif melalui sudut pandang yang berbeda hingga diketahui proporsi kasus positif aktual yang diidentifikasi dengan benar. Nilai recall yang tinggi menunjukkan model efektif dalam menjangkau kasus positif dan menghindari hilangnya titik data positif yang seharusnya[7]. Rumus recall adalah rumus 3.

$$\text{Recall} = \frac{TP}{TP+FN} \quad (3)$$

- F1-Score : sebuah matrik yang memberikan rata-rata keseimbangan antara presisi dan recall. Skor F! yang tinggi menunjukkan model memberikan keseimbangan yang baik antara mengidentifikasi dan memprediksi data positif dengan benar[7]. Rumus F1-Score adalah rumus 4.

$$\text{F1-Score} = 2 \frac{(\text{Presisi} \times \text{Recall})}{\text{Presisi} + \text{Recall}} \quad (4)$$

2.4 Interface Prediksi

Tahap terakhir dalam penelitian ini adalah interface prediksi. Deployment model tahapan terakhir dan tahapan paling menantang pada Machine Learning. Deployment merupakan proses pembuatan model pada lingkungan produksi, yang ditujukan agar dapat memberikan prediksi ke sistem perangkat lunak. Deployment penting untuk dilakukan untuk mengekstraksi prediksi handal dan memaksimalkan nilai model machine learning yang telah dibuat[8]. Deployment model ini bertujuan untuk mengetahui bagaimana sajian model yang telah dibuat agar dapat digunakan oleh orang-orang sekitar serta memanfaatkan teknologi yang ada[10].

3. HASIL DAN PEMBAHASAN

Adapun hasil dan pembahasan berdasarkan metode penelitian yang telah dicantumkan di atas penelitian klasifikasi risiko penyakit jantung dengan algoritma KNN memiliki beberapa hasil.

3.1 Pengumpulan Dataset

Tahap ini merupakan tahap pertama dalam penelitian ini. Heart Disease Dataset merupakan kumpulan data penyakit jantung yang diupload oleh johnsmith88 melalui website kaggle <https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset>. Pada Gambar 2, merupakan sajian dataset yang berjumlah 1025 data dengan 14 atribut diantaranya usia, jenis kelamin, tipe nyeri dada, tekanan darah, kolesterol, gula darah puasa, hasil elektrokardiografi, denyut jantung, angina olahraga, oldpeak, kemiringan segment, jumlah pembuluh darah utama, thal (0 = normal, 1 = cacat tetap, 2 = cacat yang dapat diperbaiki), dan kelas target (0 = tidak ada penyakit jantung, 1 = ada penyakit jantung).

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	52	1	0	125	212	0	1	168	0	1.0	2	2	3	0
1	53	1	0	140	203	1	0	155	1	3.1	0	0	3	0
2	70	1	0	145	174	0	1	125	1	2.6	0	0	3	0
3	61	1	0	148	203	0	1	161	0	0.0	2	1	3	0
4	62	0	0	138	294	1	1	106	0	1.9	1	3	2	0
...
1020	59	1	1	140	221	0	1	164	1	0.0	2	0	2	1
1021	60	1	0	125	258	0	0	141	1	2.8	1	1	3	0
1022	47	1	0	110	275	0	0	118	1	1.0	1	1	2	0
1023	50	0	0	110	254	0	0	159	0	0.0	2	0	2	1
1024	54	1	0	120	188	0	1	113	0	1.4	1	1	3	0

1025 rows × 14 columns

Gambar 2. Dataset Penyakit jantung

3.2 Pembentukan Model

Setelah tahap pengumpulan data, tahap selanjutnya adalah pembentukan model. pembentukan model pada penelitian ini melalui beberapa tahap diantaranya preprocessing (pemisahan kolom target, pembagian data latih dan data uji, penyeimbangan data), penskalaan fitur, membuat model, melatih model, dan memprediksi label.

```
# pemisahan kolom target dalam dataset
X = data.drop('target', axis=1)
y = data['target']
```

Gambar 3. Syntax Pemisahan Kolom

Pada Gambar 3, terdapat kode atau syntax preprocessing data, preprocessing data tersebut dilakukan untuk memisahkan semua fitur (variabel dependen) dalam data. X mengambil semua kolom kecuali target untuk digunakan sebagai fitur inputan. Y mengambil kolom target untuk digunakan sebagai label yang akan diprediksikan.

```
# Preprocessing data
# Membagi dataset menjadi data latih dan data uji
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

Gambar 4. Syntax Pembagian Dataset

Selanjutnya adalah Gambar 4, pembagian dataset yang dilakukan dengan kode "X_train", dengan maksimal 80% data digunakan untuk pelatihan dan 20% sisanya digunakan untuk pengujian. Parameter random_state=42 digunakan untuk memastikan pembagian data yang konsisten setiap kali kode dijalankan. Setelah data berhasil di bagi selanjutnya dilakukan penyeimbangan data menggunakan SMOTE dengan menambahkan data sintetik pada kelas minoritas.

```
Distribusi target sebelum balancing:  
Counter({1: 526, 0: 499})
```

Gambar 5. Data Tidak Seimbang

```
Distribusi target setelah balancing:  
Counter({0: 423, 1: 423})
```

Gambar 6. Data Seimbang

Pada Gambar 5 menunjukkan bahwa data tidak seimbang dimana data pada kelas 1 sebesar 526 dan data pada kelas 0 sebesar 499. Setelah dilakuakn penyeimbangan data didapatkan Gambar 6 dimana pada gambar tersebut didapatkan data seimbang diaman kelas 1 dan kelas 0 sebesar 423 .

```
# Penskalaan fitur  
scaler = StandardScaler()  
X_train = scaler.fit_transform(X_train)  
X_test = scaler.transform(X_test)
```

Gambar 7. Syntax Pembagian Data dan Penskalaan Fitur

Setelah preprocessing data dilakukan, langkah selanjutnya pada Gambar 7 adalah penskalaan fitur dilakukan menggunakan "StandardScaler" untuk menormalkan data. Hal ini penting untuk model algoritma K-Nearest Neighbors (KNN), yang skala fiturnya memiliki dampak kuat pada hasil. Transformasikan data pelatihan menggunakan "scaler.fit_transform(X_train)", di mana mean dan deviasi standar dihitung berdasarkan data pelatihan dan setiap fitur diskalakan. Selanjutnya, data pengujian juga diskalakan menggunakan "scaler.transform(X_test)" dengan mean dan deviasi standar yang sama dengan data pelatihan untuk memastikan penskalaan yang konsisten pada kedua subset.

```
# Membuat model KNN  
k = 5 # Jumlah tetangga terdekat yang akan digunakan  
knn_classifier = KNeighborsClassifier(n_neighbors=k)  
  
# Melatih model  
knn_classifier.fit(X_train, y_train)  
  
# Memprediksi label  
y_pred = knn_classifier.predict(X_test)
```

Gambar 8. Syntax Pembuatan Model

Dalam proses pembuatan model pada Gambar 8, nilai `k = 5` menentukan jumlah tetangga terdekat yang akan diperhitungkan algoritma saat melakukan prediksi. Dalam hal ini, 5 tetangga terdekat akan digunakan untuk menentukan hasil klasifikasi. Model KNN dibuat menggunakan metode 'knn_classifier yang berarti model akan mempertimbangkan 5 tetangga terdekat saat melakukan prediksi. Model kemudian dilatih, di mana data pelatihan berskala digunakan untuk memungkinkan model "mempelajari" pola dari data sehingga dapat digunakan untuk membuat prediksi berdasarkan hal baru. Setelah model dilatih, langkah selanjutnya adalah memprediksi label berdasarkan data pengujian. Biasanya, proses prediksi diimplementasikan menggunakan 'y_pred = knn_classifier.predict(X_test)' yang setelah menerapkan model ke data pengujian, akan menyimpan hasil prediksi di variabel. Hasil prediksi tersebut dapat dilihat pada Gambar 9.

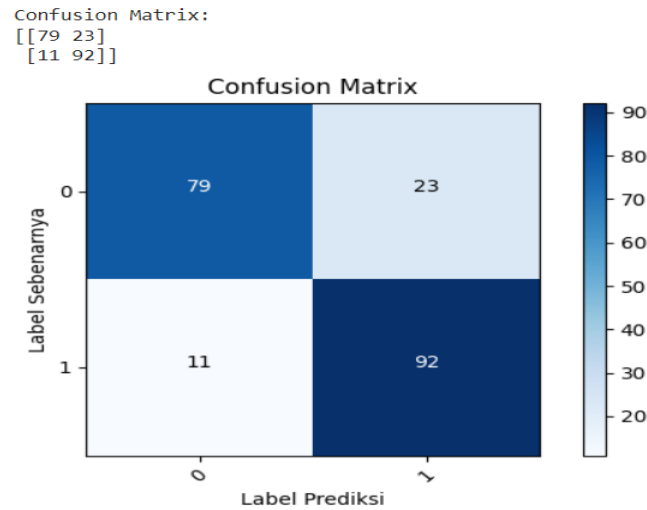
```
Hasil prediksi label untuk data uji:  
[1 1 0 1 0 0 0 0 1 0 1 0 1 1 0 1 0 1 1 0 1 0 1 1 1 1 0 1 0 1 1 1 1 1 1  
 0 1 1 0 1 0 1 0 1 0 0 0 0 0 0 1 1 0 1 0 1 1 0 0 0 1 1 0 1 0 1 0 0 0 1 0 0 0 1  
 1 1 0 1 1 1 0 0 0 0 0 0 0 0 1 1 0 1 0 0 0 1 1 1 1 0 0 0 0 1 0 1 1 0 1 0 1 0 1 1  
 1 1 0 1 1 0 1 1 0 1 0 1 1 0 0 0 0 1 0 0 1 1 0 1 1 0 1 1 0 1 1 1 0 1 1 1 1 1  
 0 0 1 0 1 1 0 0 0 1 0 1 0 1 0 0 0 0 0 0 1 1 0 1 1 0 1 1 1 0 1 1 1 0 1 1 1 1  
 1 1 1 1 1 1 1 1 1 0 0 0 0 0 1 1 1 1 1 1 0 1]
```

```
Label sebenarnya dari data uji (y_test):  
[1 1 0 1 0 1 0 0 1 0 1 0 1 1 0 0 0 1 1 0 1 0 0 0 1 1 1 1 0 0 0 1 0 1 1 1 1  
 1 1 1 0 0 1 0 0 0 0 0 0 1 1 0 0 0 1 1 0 0 0 1 1 1 0 1 0 0 1 0 0 1 0 0 0 1  
 1 1 0 0 0 1 0 0 0 0 0 1 0 1 0 0 0 0 0 0 1 1 1 1 0 0 0 0 1 0 0 1 0 1 0 1 0 1 0  
 1 1 0 1 1 0 1 1 0 1 1 0 1 1 0 1 0 1 1 0 1 1 0 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1  
 0 0 0 0 1 1 0 0 0 1 0 0 1 1 0 0 1 1 0 0 1 1 0 0 1 1 0 1 1 1 0 0 1 1 0 1 0 1  
 1 1 0 1 1 1 0 0 0 0 1 0 0 1 1 1 1 1 0 0]
```

Gambar 9. Hasil Prediksi

3.3 Evaluasi Model

Evaluasi model pada penelitian ini menggunakan confusion matrix dengan melihat dari nilai akurasi, presisi, recall, dan f1-Score. Hasil output evaluasi model confusion matrix dapat dilihat pada Gambar 10.



Gambar 10. Confusion Matrix

Pada Gambar 10, hasil confusion matrix menunjukkan 79 data negatif (0) berhasil diprediksi sebagai negatif (0), sementara 92 data positif (1) juga berhasil diprediksi sebagai data positif (1). Sementara itu terdapat pula 23 data negatif (0) salah diprediksi sebagai data positif (1) dan 11 data positif (1) salah diprediksi sebagai negatif (0). Dengan data hasil confusion matrix, performa model diketahui dengan menghitung akurasi, presisi, recall dan f1-score. Berikut perhitungan performa model dengan melihat tabel classifier pada table 1.

a. Akurasi

$$\begin{aligned} \text{Akurasi} &= \frac{TP+TN}{TP+TN+FP+FN} \times 100\% \\ \text{Akurasi} &= \frac{79+92}{79+92+11+23} \times 100\% \\ \text{Akurasi} &= 0.834146341 \times 100\% = 83\% \end{aligned}$$

b. Presisi

$$\begin{aligned} \text{Presisi} &= \frac{TP}{TP+FP} \\ \text{Presisi} &= \frac{79}{79+11} = 0.877777778 \\ \text{Presisi} &= 0.88 \end{aligned}$$

c. Recall

$$\begin{aligned} \text{Recall} &= \frac{TP}{TP+FN} \\ \text{Recall} &= \frac{79}{79+23} = 0.774509804 \\ \text{Recall} &= 0.77 \end{aligned}$$

d. F1-Score

$$\begin{aligned} \text{F1-Score} &= 2 \frac{(\text{Presisi} \times \text{Recall})}{\text{Presisi} + \text{Recall}} \\ \text{F1-Score} &= 2 \frac{(0.88 \times 0.77)}{0.88 + 0.77} \\ \text{F1-Score} &= 0.82133333 = 0.82 \end{aligned}$$

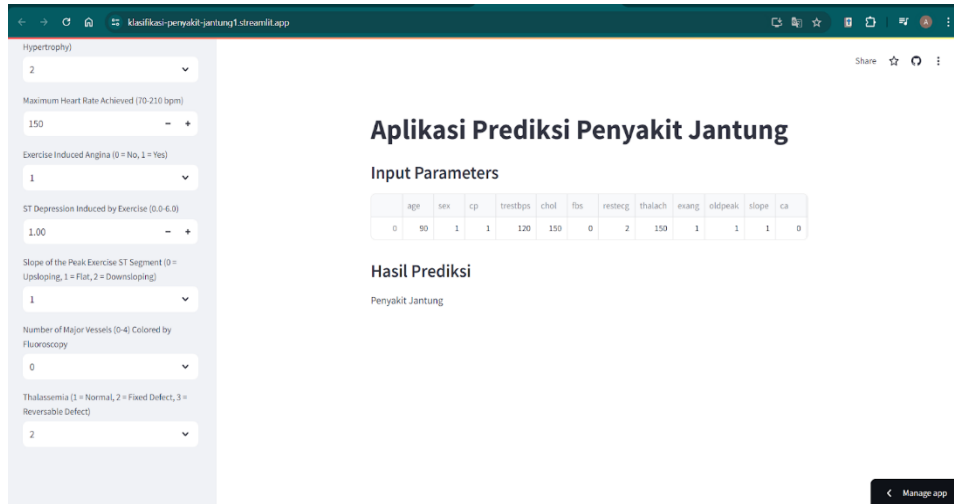
Berdasarkan pada evaluasi hasil klasifikasi dengan nilai K=5 pada algoritma K-Nearest Neighbor (KNN) didapatkan nilai akurasi sebesar 83 %, presisi 0.88, recall 0.77, dan f1-score 0.82. perhitungan tersebut dapat dilihat pada Gambar 11.

Classification Report:					
	precision	recall	f1-score	support	
0	0.88	0.77	0.82	102	
1	0.80	0.89	0.84	103	
accuracy			0.83	205	
macro avg	0.84	0.83	0.83	205	
weighted avg	0.84	0.83	0.83	205	

Gambar 11. Hasil Performa Model

3.4 Interface Prediksi

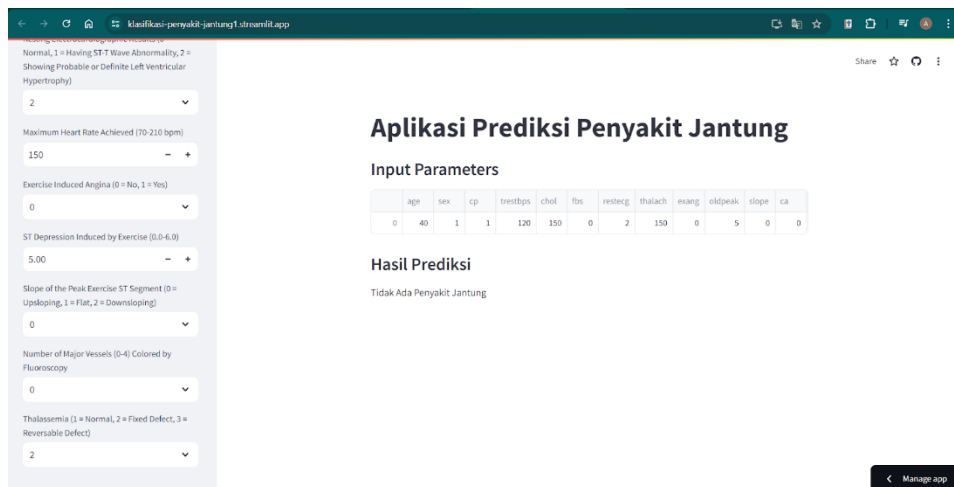
Setelah model berhasil diimplementasikan dan dilakukan evaluasi kinerja model, tahap selanjutnya adalah menyajikan model ke dalam aplikasi berbasis web. Penyajian ini dilakukan dengan melakukan deployment model menggunakan software streamlit.



The screenshot shows a web application titled "Aplikasi Prediksi Penyakit Jantung". On the left, there is a sidebar with input fields for various medical parameters: Hypertrophy (dropdown: 2), Maximum Heart Rate Achieved (70-210 bpm) (slider: 150), Exercise Induced Angina (0 = No, 1 = Yes) (dropdown: 1), ST Depression Induced by Exercise (0.0-6.0) (slider: 1.00), Slope of the Peak Exercise ST Segment (0 = Upsloping, 1 = Flat, 2 = Downsloping) (dropdown: 1), Number of Major Vessels (0-4) Colored by Fluoroscopy (dropdown: 0), and Thalassemia (1 = Normal, 2 = Fixed Defect, 3 = Reversible Defect) (dropdown: 2). On the right, there is a table for "Input Parameters" with columns: age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca. The table contains one row of data: 60, 1, 1, 120, 150, 0, 2, 150, 1, 1, 1, 0. Below the table, the "Hasil Prediksi" (Prediction Result) is displayed as "Penyakit Jantung". At the bottom right, there is a "Manage app" button.

Gambar 12. Interface Prediksi 1

Gambar 12, merupakan hasil interface prediksi dengan hasil prediksi positif beresiko penyakit jantung. Prediksi dilakukan dengan memasukkan data-data yang telah dicantumkan seperti umur, jenis kelamin, tipe nyeri dada, kolesterol, tekanan darah, gula darah, dan lain-lain.



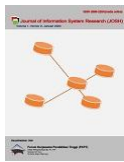
The screenshot shows the same web application as Gambar 12, but with different input parameters. The "Input Parameters" table now contains one row of data: 40, 1, 1, 120, 150, 0, 2, 150, 0, 5, 0, 0. The "Hasil Prediksi" (Prediction Result) is displayed as "Tidak Ada Penyakit Jantung". The "Manage app" button is still present at the bottom right.

Gambar 13. Interface Prediksi 0

Gambar 13, merupakan hasil interface prediksi dengan hasil prediksi negatif tidak beresiko penyakit jantung. Prediksi dilakukan dengan memasukkan data-data yang telah dicantumkan seperti umur, jenis kelamin, tipe nyeri dada, kolesterol, tekanan darah, gula darah, dan lain-lain.

4. KESIMPULAN

Setelah melakukan klasifikasi dengan menggunakan algoritma K-Nearest Neighbors (KNN) dimana didalam dataset penyakit jantung tersebut terdapat 1025 data pasien dengan 14 atribut, dapat disimpulkan bahwa model prediksi penyakit jantung yang digunakan memiliki performa yang cukup baik. Dengan nilai akurasi 83,% dari seluruh kasus yang diuji. Dalam mengidentifikasi pasien yang tidak memiliki penyakit jantung, model memiliki kemampuan yang baik dengan tingkat presisi 88%. Dari 102 kasus yang diprediksi sebagai tidak memiliki penyakit jantung, 79 diantaranya benar-benar tidak memiliki penyakit jantung. Namun, model juga mengalami kesalahan prediksi di mana 23 kasus salah diprediksi sebagai memiliki penyakit jantung. Sementara itu, dalam mengidentifikasi pasien yang memiliki penyakit jantung, model juga menunjukkan performa yang baik dengan tingkat presisi 80%. Dari 103 kasus yang diprediksi sebagai memiliki penyakit jantung, 92 di antaranya benar-benar memiliki penyakit jantung. Namun, terdapat 11 kasus yang salah diprediksi sebagai tidak memiliki penyakit



jantung. Dalam hal menemukan pasien yang sebenarnya memiliki penyakit jantung, model dapat mengenali dengan benar 89% dari mereka. Sedangkan dalam menemukan pasien yang sebenarnya tidak memiliki penyakit jantung, model dapat mengenali dengan benar 77% dari mereka. Secara keseluruhan, model K-Nearest Neighbor (KNN) memiliki performa yang baik dalam mengklasifikasikan kedua kelas, dengan skor F1 0,82 untuk kelas tanpa penyakit jantung dan 0,84 untuk kelas dengan penyakit jantung. Meskipun demikian, masih terdapat ruang untuk peningkatan dalam mengurangi kesalahan prediksi dan meningkatkan akurasi keseluruhan. Dengan kata lain, model yang digunakan untuk memprediksi penyakit jantung ini cukup andal dan dapat digunakan sebagai alat bantu dalam mendeteksi penyakit jantung dengan tingkat keakuratan yang cukup tinggi, meskipun masih terdapat beberapa kesalahan prediksi yang perlu diminimalisir.

REFERENCES

- [1] S. F. apt. Yasmin Azhar, “Penyakit Jantung,” <https://www.klikdokter.com/penyakit/masalah-jantung-dan-pembuluh-darah/penyakit-jantung> (accessed Sep. 30, 2024).
- [2] Humas Fakultas Kedokteran Universitas Brawijaya, “World Heart Day 2023: Use Heart Know Heart,” Prasetya Online, 2023. <https://prasetya.ub.ac.id/world-heart-day-2023-use-heart-know-heart/> (accessed Sep. 30, 2024).
- [3] R. Setiawan, “Apa itu Data Mining dan Bagaimana Metodenya?” <https://www.dicoding.com/blog/apa-itu-data-mining/> (accessed Jun. 13, 2024).
- [4] U. Nijunnihayah and S. S. Hilabi, “Implementation of the K-Nearest Neighbor Algorithm to Predict Sales of Medical Devices in Medical Devices Implementasi Algoritma K-Nearest Neighbor untuk Prediksi Penjualan Alat Kesehatan pada Media Alkes,” vol. 4, no. April, pp. 695–701, 2024.
- [5] R. Rismala, I. Ali, and A. Rizki Rinaldi, “Penerapan Metode K-Nearest Neighbor Untuk Prediksi Penjualan Sepeda Motor Terlaris,” JATI (Jurnal Mhs. Tek. Inform., vol. 7, no. 1, pp. 585–590, 2023, doi: 10.36040/jati.v7i1.6419.
- [6] M. Fansyuri, “Analisa algoritma klasifikasi k-nearest neighbor dalam menentukan nilai akurasi terhadap kepuasan pelanggan (study kasus pt. Trigatra komunikatama),” Humanika J. Ilmu Sos. Pendidikan, dan Hum., vol. 3, no. 1, pp. 29–33, 2020.
- [7] M. Heydarian, T. E. Doyle, and R. Samavi, “MLCM: Multi-Label Confusion Matrix,” IEEE Access, vol. 10, pp. 19083–19095, 2022, doi: 10.1109/ACCESS.2022.3151048.
- [8] N. Hafidhoh, A. P. Atmaja, G. N. Syaifuddin, I. B. Sumafta, S. M. Pratama, and H. N. Khasanah, “Machine Learning untuk Prediksi Kegagalan Mesin dalam Predictive Maintenance System,” J. Masy. Inform., vol. 15, no. 1, pp. 56–66, 2024, doi: 10.14710/jmasif.15.1.63641.
- [9] D. Lapp, “Heart Disease Dataset,” 2019. <https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset> (accessed May. 2, 2024).
- [10] Syahril Dwi Prasetyo, Shofa Shofiah Hilabi, and Fitri Nurapriani, “Analisis Sentimen Relokasi Ibukota Nusantara Menggunakan Algoritma Naïve Bayes dan KNN,” J. KomtekInfo, vol. 10, pp. 1–7, 2023, doi: 10.35134/komtekinfo.v10i1.330.
- [11] J. Muliawan and E. Dazki, “Sentiment Analysis of Indonesia’s Capital City Relocation Using Three Algorithms: Naïve Bayes, Knn, and Random Forest,” J. Tek. Inform., vol. 4, no. 5, pp. 1227–1236, 2023, doi: 10.52436/1.jutif.2023.4.5.1436.
- [12] Fritz Al, “Pra-pemrosesan dan Visualisasi Data untuk Model Machine Learning,” ICHI.PRO, Feb. 12, 2020. <https://ichi.pro/id/pra-pemrosesan-dan-visualisasi-data-untuk-model-machine-learning-156112039253027#> (accessed Jun. 12, 2024).
- [13] M. Ramdhani et al., “Prediksi Capaian Bulanan Pajak Daerah Kabupaten Bandung Barat Menggunakan Metode Logistic Regression,” J. Inf. Syst. Res., vol. 5, no. 4, pp. 881–890, 2024, doi: 10.47065/josh.v5i4.5330.
- [14] P. R. Sihombing and A. M. Arsani, “Comparison of Machine Learning Methods in Classifying Poverty in Indonesia in 2018,” J. Tek. Inform., vol. 2, no. 1, pp. 51–56, 2021, doi: 10.20884/1.jutif.2021.2.1.52.
- [15] D. Cahyanti, A. Rahmayani, and S. A. Husniar, “Analisis performa metode Knn pada Dataset pasien pengidap Kanker Payudara,” Indones. J. Data Sci., vol. 1, no. 2, pp. 39–43, 2020, doi: 10.33096/ijodas.v1i2.13.
- [16] R. Y. Parapat, E. Sandjaya, S. A. Nurfadhilah, M. M. Fetok, N. Hikmah, and Salaffudin, “Scientica Scientica,” Eval. Keselam. Kerja Di PT. Timah Ind. Dengan Menggunakan Metod. HIRARC, vol. 2, pp. 251–255, 2024.
- [17] M. Gamma, A. Hakim, and F. Irwiensyah, “Analisis Sentimen Terhadap Ulasan Pengguna Pada Aplikasi BCA Mobile Menggunakan Metode Naïve Bayes,” J. Inf. Syst. Res., vol. 5, no. 4, pp. 911–921, 2024, doi: 10.47065/josh.v5i4.5343.
- [18] M. Tam, “Analisis Penerimaan Pengguna E-Wallet DANA Menggunakan,” vol. 5, no. 4, pp. 891–900, 2024, doi: 10.47065/josh.v5i4.5334.
- [19] R. Rifaldi, J. Indra, A. R. Pratama, and A. R. Juwita, “Analisis Sentimen Pemboikotan Produk dengan Pendekatan Algoritma Naïve Bayes Media Sosial X,” J. Inf. Syst. Res., vol. 5, no. 4, pp. 940–946, 2024, doi: 10.47065/josh.v5i4.5420.
- [20] T. Handayani, S. Bahri, and Kasliono, “Implementasi Metode K-Medoids Dalam Pengelompokan Kepuasan Masyarakat Terhadap Pelayanan Rumah Sakit,” J. Inf. Syst. Res., vol. 5, no. 4, pp. 1006–1017, 2024, doi: 10.47065/josh.v5i4.5331.
- [21] M. Istifarsari, L. T. Ningrum, and L. Utari, “Implementasi Algoritma Apriori Menggunakan Cross-Industry Standar Process for Data-Mining Untuk Menentukan Pola Pembelian Obat,” J. Inf. Syst. Res., vol. 5, no. 4, pp. 1063–1075, 2024, doi: 10.47065/josh.v5i4.5263.
- [22] R. Harahap, M. Irtan, M. A. Dinata, L. Efrizoni, and Rahmaddeni, “Perbandingan Algoritma Random Forest Dan Xgboost Untuk Klasifikasi Penyakit Paru-Paru Berdasarkan Data Demografi Pasien,” J. Ilm. Betrik, vol. 15, no. 02, pp. 130–141, 2024.
- [23] S. Anggraini, M. Akbar, A. Wijaya, H. Syaputra, and M. Sobri, “Klasifikasi Gejala Penyakit Coronavirus Disease 19 (COVID-19) Menggunakan Machine Learning,” J. Softw. Eng. Ampera, vol. 2, no. 1, pp. 57–68, 2021, doi: 10.51519/journalsea.v2i1.105.



- [24] A. Handika Permana, F. Rakhmat Umbara, and F. Kasyidi, “Klasifikasi Penyakit Jantung Tipe Kardiovaskular Menggunakan Adaptive Synthetic Sampling dan Algoritma Extreme Gradient Boosting,” *Build. Informatics, Technol. Sci.*, vol. 6, no. 1, pp. 499–508, 2024, doi: 10.47065/bits.v6i1.5421.
- [25] J. Sihombing, “Klasifikasi Data Antropometri Individu Menggunakan Algoritma Naïve Bayes Classifier,” *BIOS J. Teknol. Inf. dan Rekayasa Komput.*, vol. 2, no. 1, pp. 1–10, 2021, doi: 10.37148/bios.v2i1.15.
- [26] M. C. Yustina, I. N. Ichsan, and G. M. Suranegara, “Implementasi Algoritma Genetika Proses Mutasi Differential Evolution Pada Sistem Penjadwalan Mata Pelajaran,” *KLIK Kaji. Ilm. Inform. dan Komput.*, vol. 5, no. 1, pp. 116–130, 2024, doi: 10.30865/klik.v5i1.2109.