INOTEK, Vol.8 Agustus 2024
ISSN: 2580-3336 (Print) / 2549-7952 (Online)
Url: https://proceeding.unpkediri.ac.id/index.php/inotek/

Prediksi Kematian Akibat Gagal Jantung Menggunakan Algoritma K-Nearest Neighbors

Diterima:
10 Juni 2024
Revisi:
10 Juli 2024
Terbit:
1 Agustus 2024

[1*]Atma Agilia Triwardani, [2]Muhlishoh Husna Ulfiah, [3]Aidina Ristyawan, [4]Erna Daniati
[1-3]Universitas Nusantara PGRI Kediri
[1]atmaagilia@gmail.com, [2]muhlishohhusnaulfiah@gmail.com, [3]aidinaristi@unp.ac.id, [4]ernadaniati@unp.ac.id.

Abstrak—Artikel ini mengeksplorasi penggunaan algoritma K-Nearest Neighbors (K-NN) untuk klasifikasi pasien gagal jantung berdasarkan data klinis dari Kaggle. Proses penelitian mencakup pra-pemrosesan data, normalisasi fitur, pemilihan parameter k optimal melalui cross-validation, dan evaluasi model dengan metrik akurasi, precision, recall, dan F1-score. Hasil menunjukkan bahwa algoritma K-NN dengan parameter k=7 optimal mampu mengklasifikasikan kematian pasien dengan akurasi yang memadai sebesar 84%. Penemuan ini menunjukkan potensi besar dari penggunaan K-NN dalam mendukung pengambilan keputusan klinis dan meningkatkan diagnosis kematian akibag gagal jantung. Implementasi data mining dengan K-NN menawarkan pendekatan yang efektif untuk analisis medis, berkontribusi pada peningkatan kualitas perawatan pasien.

Kata Kunci—Data Mining;Gagal Jantung;KNN

Abstract— *This article explores the use of the K-Nearest Neighbors (K-NN) algorithm for classification of heart failure patients based on clinical data from Kaggle. The research process includes data pre-processing, feature normalization, selection of optimal k parameters through cross-validation, and model evaluation with accuracy, precision, recall, and F1-score metrics. The results show that the K-NN algorithm with optimal k=7 parameters is able to classify patient deaths with an accuracy of 84%. These findings demonstrate the great potential of using K-NN in supporting clinical resolution and improving the diagnosis of heart failure-related deaths. Implementation of data mining with K-NN offers an effective approach to medical analysis, contributing to improving the quality of patient care..*
Keywords—*data mining;heart failure;KNN*

Penulis Korespondensi:

Aidina Ristyawan,
Sistem Informasi,
Universitas Nusantara PGRI Kediri,
Email: adinaristy@unpkediri.ac.id
ID Orcid: [https://orcid.org/0009-0003-2712-1507]
Handphone: 081232624460

I. PENDAHULUAN

Gagal jantung, juga dikenal sebagai gagal jantung, merupakan salah satu penyakit yang memiliki tingkat mortalitas dan morbiditas yang paling tinggi. Menurut World Health Organisation (WHO), prevalensi penyakit gagal jantung di Amerika Serikat pada tahun 2013 mencapai kurang lebih 550.000 kasus per tahun, dan American Heart Association (AHA) melaporkan bahwa sebanyak 375.000 orang meninggal dunia setiap tahun akibat penyakit gagal jantung di Amerika Serikat. Pada tahun 2018, data yang dikumpulkan di Indonesia menunjukkan bahwa penyakit gagal jantung adalah salah satu dari 10 penyakit tidak menular yang paling umum di negara itu, dengan jumlah kasus yang diperkirakan sebanyak 229,696 atau 0,13% [1]. Dokter dan tenaga medis telah memprioritaskan prediksi gagal jantung karena pentingnya organ vital seperti jantung. Namun, sampai saat ini, prediksi gagal jantung dalam praktik klinis seringkali tidak cukup akurat [2].

Mengetahui seseorang yang berisiko tinggi terkena penyakit jantung dan memastikan mereka mendapatkan perawatan yang tepat dapat membantu mencegah kematian dini. Data mining adalah proses penggalian informasi dari sejumlah besar data untuk menghasilkan informasi tersembunyi di dalamnya. Teknik data mining banyak digunakan di berbagai industri untuk menemukan pola atau informasi dari industri tersebut. Data mining di industri kesehatan menjadi model baru dan banyak digunakan karena metode ini dapat mengekstraksi dan menemukan pola tersembunyi dari data yang dapat digunakan sebagai pendukung keputusan [3]. Dari penelitian sebelumnya penerapan algoritma KNN terhadap data penyakit gagal jantung dilakukan dengan pengujian akurasi menggunakan aplikasi RapidMiner. Dataset yang digunakan terdiri dari 299 data dengan 13 atribut. Implementasi algoritma KNN pada aplikasi RapidMiner dilakukan dengan variasi nilai k. Hasil akurasi tertinggi dicapai pada nilai k=7 dengan akurasi sebesar 94,92% tetapi hanya mendapat 68% pada phyton dengan confusion matrix. Maka dari itu penelitian ini membuat analisis klasisifikasi penyakit jantung menggunakan algoritma K-Nearest. Algoritma ini sering juga disebut metode K-NN, banyak digunakan untuk klasifikasi dan prediksi karena pengajarannya yang menyeluruh. Kekuatan metode K-NN termasuk sifatnya yang kuat, intensif, dan tidak asumsif terhadap pencilan [4]. Untuk hasil yang lebih optimal di penelitian kali ini akan ditambahkan dengan menggunakan teknik SMOTEE dan Cross Validation agar hasil yang diperoleh semakin akurat.

K-Nearest Neighbors (K-NN) dilakukan dengan mencari kelompok objek dalam data pelatihan yang paling dekat (mirip) dengan objek pada data baru atau data pengujian. Menghitung nilai akurasi adalah cara untuk menilai kinerja klasifikasi [5]. Untuk menganalisis data, Python dianggap sebagai bahasa yang memiliki banyak fitur perpustakaan standar yang luas dan

komprehensif, dan sintaks kodenya sangat jelas. Python adalah bahasa pemrograman yang cocok untuk pembuatan aplikasi berbasis kecerdasan buatan (artificial intelligence) karena memiliki banyak kelebihan, termasuk pemahaman tentang data mining, ilmu data, dan pembelajaran mesin [6]. Penelitian ini bertujuan untuk mengetahui seberapa akurat menggunakan algoritma K-NN sebagai pilihan metode klasifikasi dengan mengolah data yang sudah ada pada phyton menggunakan cross validaton untuk mengidentifikasi kematian pasien akibat penyakit gagal jantung.

II. METODE

A. Data Mining : Data mining adalah proses mengekstraksi dan menemukan informasi berguna dengan menggunakan matematika, statistik, kecerdasan buatan, dan pembelajaran mesin. Data mining adalah proses menemukan pola dalam data. Data mining dikategorikan menjadi deskripsi, estimasi, prediksi, klasifikasi, clustering, dan asosiasi berdasarkan fungsinya. Data mining tahap terdiri dari tiga langkah utama. Pertama, data dipilih, dibersihkan, dan diproses sebelum diproses. Proses ini dilakukan sesuai dengan pedoman dan pengetahuan ahli domain, yang menangkap dan mengintegrasikan data internal dan eksternal ke dalam tinjauan organisasi yang menyeluruh. Data mining algoritma digunakan untuk menggali data yang terintegrasi, yang memudahkan identifikasi informasi bernilai. Namun, waktu yang dibutuhkan untuk mengolah data menjadi lebih lama [7].

B. Machine Learning : Machine learning adalah sekumpulan algoritma yang digunakan untuk mengoptimalkan kinerja sistem atau komputer berdasarkan data sampel. Kemampuan utama machine learning adalah mengubah dan mengatur keputusan untuk menyesuaikan diri dengan perubahan [8]. Kelebihan machine learning adalah kemampuan untuk mengubah dan menyesuaikan data untuk menerima perubahan. Dalam hal kegunaannya, machine learning dapat digunakan untuk [9] :

1. Classification adalah metode yang digunakan untuk memprediksi nilai atau kelas individu dalam sebuah populasi.
2. Similarity matching adalah metode yang digunakan dalam pembelajaran mesin untuk menemukan kemiripan antar individu berdasarkan data yang ada.
3. Clustering adalah metode yang digunakan dalam pembelajaran mesin untuk membagi data ke dalam kelompok berdasarkan kriteria [10].

C. K-Nearest Neighbors : Algoritma K-Nearest Neighbor (K-NN) adalah metode klasifikasi sekumpulan data yang didasarkan pada pembelajaran data yang sudah terklasifikasikan

sebelumnya. Ini termasuk dalam pembelajaran yang diawasi, di mana hasil query instance yang baru diklasifikasikan berdasarkan mayoritas kedekatan jarak dari kategori yang ada dalam K-NN [11]. K-NN dilakukan dengan mencari kelompok k objek dalam data pelatihan yang paling dekat (mirip) dengan objek pada data baru atau pengujian. Untuk melakukan ini, suatu sistem klasifikasi harus memiliki kemampuan untuk mencari informasi [12]. Untuk menghitung nilai jarak pada metode K-NN, rumus jarak geometris dapat digunakan. Metode ini sederhana dan dapat memberikan hasil klasifikasi yang sangat akurat. Rumus Euclidean Distance adalah seperti berikut:

$$\text{d (xy)} = \sqrt{\sum_{i=1}^{n} (xi - yi)^2} \text{ (1)} \qquad (1)$$

Keterangan:
d : jarak terdekat
x : data training
y : data testing
n : jumlah atribut antara 1 s/d n
i : atribut individu antara 1 s/d n [13].

D. Cross Validation : Untuk mengevaluasi kinerja model, kesalahan prediksi diestimasi melalui cross validation. Data dibagi menjadi himpunan bagian k yang berjumlah hampir sama. Untuk setiap pengulangan, salah satu himpunan bagian akan digunakan sebagai pelatihan dan pengujian data [14].

E. Pengumpulan Data : Data kesehatan jantung yang digunakan dalam penelitian ini berasal dari platform Kaggle.com yang dapat diakses di https://www.kaggle.com/datasets/andrewmvd/heart-failure-clinical-data terdiri dari 12 atribut kategori dan 1 class, dengan masing-masing data terdiri dari 299 data [15].

III. HASIL DAN PEMBAHASAN

Dataset ini terdiri dari beberapa fitur penting yang mencakup usia, jenis kelamin, tekanan darah, tingkat kolesterol, dan beberapa parameter klinis lainnya yang relevan dengan kondisi gagal jantung. Berikut dataset yang digunakan:

	age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure	platelets	serum_creatinine	se
0	75.0	0	582	0	20	1	265000.00	1.9	13
1	55.0	0	7861	0	38	0	263358.03	1.1	13
2	65.0	0	146	0	20	0	162000.00	1.3	12
3	50.0	1	111	0	20	0	210000.00	1.9	13
4	65.0	1	160	1	20	0	327000.00	2.7	11

Gambar 3. 1 Isi dataset

Dari atribut yang ada berikut adalah tipe data yang bisa diketahui :

Tabel 3. 1 Tipe data

No	Nama Atribut	Tipe	Deskripsi
1	*Age*	Float64	Variabel ini menunjukkan tentang umur pasien
2	*Anemia*	Int64	Variabel ini menunjukkan penurunan hemoglobin
3	*Creatinine_Phosphokinase*	Int64	Variabel ini menunjukkan tingkatan enzim CPK pada darah (mcg/L)
4	*Diabetes*	Int64	Variabel ini mengindikasikan apakah pasien menderita diabetes atau tidak
5	*Ejection_Fraction*	Int64	Variabel ini menunjukkan presentaasi darah dalam kontaksi meninggalkan jantung
6	*High_Blood_Pressure*	Int64	Variabel ini mengindikasikan apakah pasien menderita hipertensi
7	*Platelets*	Int64	Variabel ini menunjukkan trombosit darah
8	*Serum_Creatinine*	Float64	Variabel yang mengukur kadar kreatinin serum dalam darah pasien
9	*Serum_Sodium*	Int64	Variabel yang mengukur kadar natrium serum dalam darah pasien
10	*Sex*	Int64	Variabel yang menunjukkan jenis kelamin pasien
11	*Smoking*	Int64	Variabel yang menunjukkan jika pasien merokok
12	*Time*	Int64	Variabel yang menunjukkan pengamatan pasien

A. Sebelum menerapkan algoritma K-NN, dilakukan beberapa langkah pra-pemrosesan data untuk memastikan kualitas data yang digunakan:

1. Penanganan Missing Values : Mengisi atau menghapus data yang hilang untuk mencegah bias dalam model.

2. Penanganan Imbalance Data: Menambahkan salah satu class yang diketahui minor jumlahnya dengan metode SMOTE

```
# Handle missing values by imputing with the mean of each column
imputer = SimpleImputer(strategy='mean')
data_imputed = pd.DataFrame(imputer.fit_transform(data), columns=data.columns)
```

```
# Memisahkan fitur (X) dan label (y)
X = data.drop('DEATH_EVENT', axis=1)
y = data['DEATH_EVENT']
```

```
smote = SMOTE(random_state=42)
X_resampled, y_resampled = smote.fit_resample(X, y)
```

Gambar 3. 2 Proses balancing

3. Normalisasi : Mengubah skala data ke rentang yang seragam untuk memastikan bahwa fitur-fitur yang berbeda memiliki dampak yang setara dalam proses klasifikasi.

4. Pembagian Data : Membagi dataset menjadi data latih (80%) dan data uji (20%) untuk mengevaluasi kinerja model secara objektif.

```
# Split data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X_resampled, y_resampled, test_size=0.2, random_state=42)
```

```
# Standardize features
scaler = StandardScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)
```

Gambar 3. 3 Proses pembagian data

B. Selanjutnya implementasi algoritma K-NN dilakukan dengan beberapa langkah sebagai berikut :

1. Pemilihan Parameter k : Menentukan jumlah tetangga terdekat (k) yang optimal sesuai dengan saran jurnal sebelumnya.

2. Pelatihan Model : Melatih model K-NN menggunakan data latih.

3. Pengujian Model : Menguji model menggunakan data uji untuk menilai kinerjanya.

```
# Initialize KNN classifier
knn = KNeighborsClassifier(n_neighbors=7)
```

```
# Perform cross-validation
cv_scores = cross_val_score(knn, X_train, y_train, cv=5, scoring='accuracy')
print(f'Cross-validation scores: {cv_scores}')
print(f'Mean CV accuracy: {np.mean(cv_scores):.2f}')
```

```
Cross-validation scores: [0.76923077 0.87692308 0.78461538 0.8        0.875     ]
Mean CV accuracy: 0.82
```

Gambar 3. 4 Pengujian

C. Kinerja model K-NN dievaluasi menggunakan metrik-metrik berikut:

1. Precision : Rasio antara prediksi positif yang benar terhadap total prediksi positif.
2. Recall : Rasio antara prediksi positif yang benar terhadap total kasus positif sebenarnya.
3. F1-Score : Harmonik rata-rata dari precision dan recall, memberikan ukuran keseimbangan antara keduanya.

D. Hasil evaluasi model dengan berbagai nilai k menunjukkan bahwa nilai k = 7 memberikan kinerja terbaik dengan metrik sebagai berikut : Akurasi : 84% untuk keseluruhan data pasien. 0 (pasien hidup) nilai precision : 83%, Recall : 85%, F1 score : 84%. 1 (pasien meninggal) nilai precision : 85%, Recall : 83%, F1 score : 84%.

```
# Print classification report
print('Classification Report:')
print(classification_report(y_test, y_pred))
```

```
Classification Report:
              precision    recall  f1-score   support

           0       0.83      0.85      0.84        41
           1       0.85      0.83      0.84        41

    accuracy                           0.84        82
   macro avg       0.84      0.84      0.84        82
weighted avg       0.84      0.84      0.84        82
```

Gambar 3. 5 Performa model

E. Perbandingan dengan jurnal sebelumnya bisa dilihat pada tabel 3.8 dibawah ini. Dengan menggunakan cross validation serta metode SMOTE akurasi, precision, dan recall meingkat secara drastis. Menggabungkan cross validation dan SMOTE dapat memberikan hasil yang lebih optimal. Cross validation memastikan bahwa model dievaluasi secara adil dan menyeluruh, sementara SMOTE memastikan bahwa kelas minoritas diwakili dengan baik dalam setiap fold. Ini membantu dalam membangun model yang tidak hanya akurat tetapi juga adil dalam prediksi

untuk semua kelas. Secara keseluruhan, cross validation dan SMOTE adalah teknik yang saling melengkapi yang sangat berguna dalam meningkatkan kualitas dan keandalan model pembelajaran mesin, terutama dalam situasi dengan dataset yang kecil atau tidak seimbang.

Tabel 3.8 Tabel perbandingan

	KNN Confusion Matrix	KNN Cross Validation + SMOTE
Akurasi	68%	84%
Precision Meninggal	75%	85%
Precision Hidup	68%	83%
Recall Meninggal	14%	83%
Recall Hidup	97%	85%

IV. KESIMPULAN

Penelitian ini telah berhasil mengaplikasikan algoritma K-Nearest Neighbor (K-NN) untuk klasifikasi pasien gagal jantung menggunakan dataset dari Kaggle. Hasil yang diperoleh menunjukkan bahwa algoritma K-NN, dengan pemilihan parameter yang tepat dan penerapan teknik Cross Validation, mampu memberikan hasil klasifikasi yang akurat dan andal. Implementasi algoritma K-NN dalam klasifikasi pasien gagal jantung dapat memberikan dukungan signifikan dalam proses diagnosis medis. Dengan model yang andal, tenaga medis dapat membuat keputusan yang lebih cepat dan tepat, yang pada akhirnya dapat meningkatkan kualitas perawatan pasien gagal jantung. Dengan hasil yang diperoleh, diharapkan dapat memberikan manfaat praktis dan menjadi landasan bagi pengembangan lebih lanjut dalam aplikasi medis berbasis data.

DAFTAR PUSTAKA

[1] D. Prihatiningsih dan T. Sudyasih, “Perawatan Diri Pada Pasien Gagal Jantung,” Des 2018, Diakses: 7 Juni 2024. [Daring]. Tersedia pada: http://localhost:8080/xmlui/handle/123456789/800

[2] Y. Pratama, A. Prayitno, D. Azrian, N. Aini, Y. Rizki, dan E. Rasywir, “Klasifikasi Penyakit Gagal Jantung Menggunakan Algoritma K-Nearest Neighbor,” *Bulletin of Computer Science Research*, vol. 3, no. 1, hlm. 52–56, Des 2022, doi: 10.47065/BULLETINCSR.V3I1.203.

[3] D. A. Firdlous, “Komparasi Algoritma Klasifikasi Data Mining Untuk Memprediksi Penyakit Jantung,” *Infoman’s : Jurnal Ilmu-ilmu Manajemen dan Informatika*, vol. 16, no. 1, hlm. 79–84, Mei 2022, Diakses: 7 Juni 2024. [Daring]. Tersedia pada: https://journal.unsap.ac.id/index.php/infomans/article/view/412

[4] A. Setiawan, R. F. Waleska, M. A. Purnama, Rahmaddeni, dan L. Efrizoni, "KOMPARASI ALGORITMA K-NEAREST NEIGHBOR (K-NN), SUPPORT VECTOR MACHINE (SVM), DAN DECISION TREE DALAM KLASIFIKASI PENYAKIT STROKE," *Jurnal Informatika dan Rekayasa Elektronik*, vol. 7, no. 1, hlm. 107–114, Apr 2024, doi: 10.36595/JIRE.V7I1.1161.

[5] A. Muhadi dan A. Octaviano, "Penerapan Data Mining Untuk Prediksi Hasil Keuntungan Lelang Mesin X-Ray Tahun 2020 Dengan Metode K-Nearest Neighbor (Studi Kasus : PT.Ramadika Mandiri)," *Jurnal Informatika Multi*, vol. 1, no. 2, hlm. 126–136, Mar 2023, Diakses: 7 Juni 2024. [Daring]. Tersedia pada: https://jurnal.publikasitecno.id/index.php/multi/article/view/19

[6] R. Setiawan dan A. Triayudi, "Klasifikasi Status Gizi Balita Menggunakan Naïve Bayes dan K-Nearest Neighbor Berbasis Web," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 6, no. 2, hlm. 777–785, Apr 2022, doi: 10.30865/MIB.V6I2.3566.

[7] M. Baharuddin, M. M. Baharuddin, H. Azis, dan T. Hasanuddin, "ANALISIS PERFORMA METODE K-NEAREST NEIGHBOR UNTUK IDENTIFIKASI JENIS KACA," *ILKOM Jurnal Ilmiah*, vol. 11, no. 3, hlm. 269–274, Des 2019, doi: 10.33096/ilkom.v11i3.489.269-274.

[8] H. K. Pambudi *dkk.*, "PREDIKSI STATUS PENGIRIMAN BARANG MENGGUNAKAN METODE MACHINE LEARNING," *Jurnal Ilmiah Teknologi Infomasi Terapan*, vol. 6, no. 2, hlm. 100–109, Apr 2020, doi: 10.33197/JITTER.VOL6.ISS2.2020.396.

[9] D. Immanuel Salintohe, I. Alwiah Musdar, T. Informatika, dan S. Kharisma Makassar, "IMPLEMENTASI MACHINE LEARNING UNTUK MENGIDENTIFIKASI TANAMAN HIAS PADA APLIKASI TIERRA," *JTRISTE*, vol. 9, no. 1, hlm. 1–15, Mar 2022, doi: 10.55645/JTRISTE.V9I1.360.

[10] D. Theodorus, S. Defit, dan G. W. Nurcahyo, "Machine Learning Rekomendasi Produk dalam Penjualan Menggunakan Metode Item-Based Collaborative Filtering," *Jurnal Informasi dan Teknologi*, hlm. 202–208, Des 2021, doi: 10.37034/JIDT.V3I4.151.

[11] F. T. Admojo dan Ahsanawati, "Klasifikasi Aroma Alkohol Menggunakan Metode KNN," *Indonesian Journal of Data and Science*, vol. 1, no. 2, hlm. 34–38, Jul 2020, doi: 10.33096/IJODAS.V1I2.12.

[12] F. T. Admojo dan Ahsanawati, "Klasifikasi Aroma Alkohol Menggunakan Metode KNN," *Indonesian Journal of Data and Science*, vol. 1, no. 2, hlm. 34–38, Jul 2020, doi: 10.33096/IJODAS.V1I2.12.

[13] S. P. Adenugraha, V. Arinal, dan D. I. Mulyana, "Klasifikasi Kematangan Buah Pisang Ambon Menggunakan Metode KNN dan PCA Berdasarkan Citra RGB dan HSV," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 6, no. 1, hlm. 9–17, Jan 2022, doi: 10.30865/MIB.V6I1.3287.

[14] L. Mardiana, D. Kusnandar, dan N. Satyahadewi, "ANALISIS DISKRIMINAN DENGAN K FOLD CROSS VALIDATION UNTUK KLASIFIKASI KUALITAS AIR DI KOTA PONTIANAK," *Bimaster : Buletin Ilmiah Matematika, Statistika dan Terapannya*, vol. 11, no. 1, hlm. 97–102, Jan 2022, doi: 10.26418/BBIMST.V11I1.51608.

[15] A. Puspita Sari, A. Nugroho Sihananto, dan D. Arman Prasetya, "Implementasi Metode K-NN dalam Klasterisasi Kasus Kesehatan Jantung," *ALINIER: Journal of Artificial Intelligence & Applications*, vol. 3, no. 2, hlm. 94–99, Des 2022, doi: 10.36040/ALINIER.V3I2.5761.

[16] Andrew MVD. (2020). Heart Failure Clinical Data. Kaggle. Retrieved from https://www.kaggle.com/datasets/andrewmvd/heart-failure-clinical-data