

Klasifikasi Penyakit Jantung Menggunakan Perbandingan Implementasi Algoritma *Naïve Bayes* dan *K-Nearest Neighbor*

Valentino Simamora¹, Anita Desiani², Irmeilyana³
^{1,2,3} Universitas Sriwijaya, Indonesia

Info Artikel

Riwayat Artikel

Diterima: 18-01-2024

Disetujui: 30-04-2024

Kata Kunci

Jantung;

K-NN;

Naïve Bayes;

Perbandingan;

anita_desiani@unsri.ac.id

ABSTRAK

Jantung merupakan organ penting pada manusia. Jantung memiliki peran aktif dalam memompa darah ke seluruh tubuh. Selama menjalankan tugasnya, jantung dapat mengalami berbagai macam penyakit. Kita dapat melakukan klasifikasi terhadap penyakit jantung untuk mengetahui jumlah rata-rata manusia yang terserang penyakit jantung. Untuk mendapatkan algoritma yang terbaik, kita perlu melakukan perbandingan. Pada penelitian ini, algoritma yang digunakan adalah *K-Nearest Neighbor* dan *Naïve Bayes*. Dari hasil penelitian pada *K-NN* diperoleh hasil akurasi, presisi, dan recall sebesar 100% sedangkan *Naïve Bayes* mendapat hasil akurasi, presisi, dan recall sebesar 85%-85,5%. Maka algoritma yang terbaik adalah *Naïve Bayes*.

1. PENDAHULUAN

Salah satu organ yang sangat penting pada manusia adalah jantung [1]. Jantung berfungsi untuk memompa darah keseluruh tubuh [2]. Jantung memiliki ruang yang dibatasi beberapa katup, diantaranya adalah katup atrioventrikular dan semilunar [3]. Jantung dapat mengalami berbagai macam masalah hingga terserang penyakit. Beberapa penyakit pada jantung yaitu penyakit jantung katup, jantung koroner hingga gagal jantung [4]. Penyebab kematian terbanyak di dunia adalah penyakit jantung koroner [5]. Sedangkan mengenai penyakit gagal jantung, data yang terkumpul masih sangat sedikit [6]. Klasifikasi dapat diterapkan untuk melakukan deteksi terhadap penyakit jantung dengan menggunakan data mining.

Penelitian sebelumnya pernah dilakukan Sahar yaitu menganalisis perbandingan metode *K-NN* dan *Naïve Bayes* dalam klasifikasi pada dataset penyakit jantung [3]. Hasil yang diperoleh adalah *K-NN* mendapatkan akurasi sebesar 67%, presisi 65%, dan recall sebesar 73%. Sedangkan *Naïve Bayes* mendapatkan akurasi sebesar 58%, presisi 90%, dan recall sebesar 55%. Penelitian yang serupa juga telah dilakukan oleh Riski Annisa [7]. Hasil yang diperoleh adalah *K-NN* mendapatkan akurasi sebesar 60%, presisi 64%, dan recall sebesar 65%. Sedangkan algoritma *Naïve Bayes* mendapatkan akurasi sebesar 78%, presisi 79%, dan recall sebesar 89%. Berdasarkan dua penelitian yang telah dilakukan dapat dilihat bahwa penggunaan algoritma *K-NN* memberikan hasil yang cukup seimbang antara nilai akurasi, presisi, dan recall walaupun masih relatif bernilai kecil. Pada dasarnya algoritma tau metode yang dapat digunakan sangat banyak, tergantung dengan kebutuhan dan kecocokan algoritma yang digunakan terhadap dataset yang dimiliki. Kita ambil contoh algoritma misalnya *Naïve Bayes*, kelebihan dari algoritma *Naïve Bayes* ini adalah dapat memberikan akurasi dengan kecepatan yang cukup tinggi ketika digunakan pada database yang besar dan hanya

mempunyai data training dengan jumlah yang kecil untuk menentukan estimasi parameter yang diperlukan dalam proses pengklasifikasian. Adapun kelemahan dari penggunaan algoritma *Naïve Bayes* yaitu jika probabilitas kondisionalnya adalah nol mengakibatkan probabilitas prediksi akan bernilai nol juga, sehingga algoritma *Naïve Bayes* tidak berlaku. Selain *Naïve Bayes*, algoritma lainnya seperti algoritma *K-NN* (*K-Nearest Neighbor*), Kelebihan algoritma *K-NN* ini ialah mudah dipelajari karena sederhana, proses pelatihan yang sangat cepat, tetap dapat digunakan pada data yang memiliki bias, dan tetap efektif untuk digunakan pada data berjumlah besar. Kekurangan dari algoritma *K-NN* adalah keterbatasan memori, biasanya nilai k , dipengaruhi oleh data-data yang tidak relevan dan komputasi kompleks.

Penelitian kali ini menggunakan 2 algoritma yaitu *Naïve Bayes* dan algoritma *K-NN* untuk melakukan klasifikasi pada dataset mengenai seseorang sedang terserang penyakit jantung atau tidak. Kita akan menghitung nilai-nilai presisi, recall serta akurasi dari dataset tersebut. Kedua algoritma yaitu algoritma *Naïve Bayes* dan algoritma *K-NN* akan digunakan dan setiap hasilnya akan dibandingkan untuk memperoleh kesimpulan mengenai algoritma yang paling cocok untuk dataset tersebut ketika melakukan klasifikasi.

2. METODE

Penelitian ini menggunakan dua algoritma yaitu *Naïve Bayes* dan *K-NN*. Pertama cari dan tentukan dataset mengenai penyakit jantung.

2.1. Deskripsi Data

Dataset yang akan kita gunakan untuk melakukan penelitian ini diambil dari situs Kaggle (<https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset>) dalam bentuk file csv. Ada 14 atribut diantaranya 13 atribut sebagai ciri dan sisa satu sebagai atribut target. Target juga dibagi menjadi 2 label, yaitu pertama 0 = tidak ada penyakit dan yang kedua 1 = ada penyakit. Didalam dataset ini memiliki 1025 data. Atribut yang digunakan pada dataset ini yaitu *age*, *sex*, *cp*, *trestbps*, *chol*, *fbs*, *restecg*, *thalach*, *exang*, *oldpeak*, *slope*, *ca*, *thal*, *target*.

Tabel 1. Penjelasan dan Arti dari Atribut

Attribute	Feature Meaning
<i>Target/Target</i>	0 = tidak ada penyakit, 1 = ada penyakit
<i>Age/Umur</i>	
<i>Sex/Jenis Kelamin</i>	0 = female, 1 = male
<i>Chest Pain/ Nyeri Dada</i>	0 = tidak nyeri, 1 = nyeri ringan, 2 = nyeri, 3 = sangat nyeri
<i>Trestbps/ Tekanan Darah Rendah</i>	
<i>Cholestrol/ Kolestrol</i>	
<i>Fbs / Gula Darah Normal</i>	0 = salah, 1 = benar
<i>Restecg/ Elektrokardiografi</i>	0 = tidak sehat, 1 = cukup sehat, 2 = sehat
<i>Thalach/ Detak Jantung Maksimum Tercapai</i>	

<i>Exang</i> / Kestabilan Induksi Angina	0 = tidak stabil, 1 = stabil
<i>Oldpeak</i> / Depresi ST yang diinduksi oleh olahraga relatif terhadap istirahat	
<i>Slope</i> / Kemiringan segmen	
<i>Ca</i> / nomor pembuluh darah utama	
<i>Thal</i> /	0 = normal, 1 = cacat tetap, 2 = cacat reversibel

2.2. Processing Data

Pada tahap ini kita akan melakukan beberapa hal seperti membuang data yang terduplikasi, melakukan pemeriksaan terhadap data yang tidak konsisten, memperbaiki setiap kesalahan yang ada didalam data, salah satu contohnya kesalahan cetak [8]. Didalam dataset ini tidak terdapat atribut yang harus dibuang atau dihapus, dengan begitu dapat dinyatakan bahwa seluruh atribut yaitu sebanyak 14 atribut akan digunakan dalam klasifikasi pada penelitian ini.

Kita akan menggunakan teknik *presentase split* dengan perbandingan 8:2. Kita membutuhkan data yang digunakan dalam proses pembelajaran atau latih yaitu data *training*. Selain itu kita juga membutuhkan data yang akan diuji yaitu data *testing*. Dengan menggunakan perbandingan 8:2 maka artinya sebanyak 80% dari data akan dijadikan sebagai data *training* (data latih) dan sebanyak 20% dari data akan dijadikan data *testing* (data uji).

2.3. Algoritma Naïve Bayes

Naïve Bayes diperkenalkan oleh Thomas Bayes yang merupakan seorang ilmuwan dari inggris. Ini merupakan pengklasifikasian dengan metode statistic dan metode probabilitas [9]. Berikut ini merupakan langkah-langkah melakukan klasifikasi dengan menggunakan algoritma *Naïve Bayes*:

1. Melakukan perhitungan terhadap jumlah kategori pada setiap variable yang ada
2. Melakukan perhitungan terhadap peluang dari setiap kategori
3. Menentukan jumlah kemunculan (frekuensi) dari setiap kategori
4. Menentukan kategori menggunakan nilai maksimal

Rumus perhitungan untuk *Naïve Bayes* sebagai berikut [10]:

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)}$$

dimana:

- X : data dengan kelas yang belum diketahui
- H : hipotesis data X merupakan suatu kelas spesifik
- $P(H|X)$: probabilitas hipotesis H berdasarkan kondisi X
- $P(H)$: probabilitas hipotesis H
- $P(X|H)$: probabilitas X berdasarkan kondisi pada hipotesis H
- $P(X)$: probabilitas hipotesis X

2.4. Algoritma K-NN (*K-Nearest Neighbor*)

Algoritma *K-Nearest Neighbor* (*K-NN*) digunakan untuk melakukan klasifikasi pada objek yang didasari oleh data pembelajaran dengan jarak paling dekat terhadap objek tersebut [11]. Fitur-fitur yang relevan sangat memengaruhi ketepatan algoritma ini [12]. Tahapan dalam menggunakan algoritma *K-NN*, sebagai berikut:

1. Menentukan terlebih dahulu banyaknya tetangga k (sebaiknya ganjil)
2. Menghitung jarak pada data untuk dilakukannya perbandingan dengan dataset *training* menggunakan persamaan jarak Euclidean

$$dist(p, q) = \sqrt{\sum_{i=1}^n (p_i - q_i)^2}$$

dimana:

$dist(p, q)$: jarak antara p dan q

p_i : nilai ke-i dari data p

q_i : nilai ke-i dari data q

3. Mengatur urutan berdasarkan kecil ke besar kemudian pilih himpunan k yang paling sedikit pada dataset terkecil
4. Menentukan bahwa jawaban dengan data yang akan diprediksi adalah kelompok data yang memiliki jumlah k pertama dari kumpulan data terbesar
5. Gunakan titik pertimbangan untuk menetapkan kelas-kelas terdekat

2.5. Evaluasi Hasil

Confusion matrix merupakan tabel yang menyatakan jumlah suatu data uji yang diklasifikasikan secara salah maupun secara benar. *Confusion matrix* adalah sebuah matriks yang menampilkan visualisasi kinerja dari algoritma klasifikasi menggunakan data dalam matriks. Data dalam matriks membagi klasifikasi prediksi dalam empat bentuk yaitu *True Positif* (TP), *True Negatif* (TN), *False Positif* (FP), dan *False Negatif* (FN). Bentuk *confusion matrix* untuk klasifikasi dengan dua kelas, terdapat pada Tabel 2 berikut ini [13].

Tabel 2. *Confusion Matrix*

Kelas	Prediksi YES	Prediksi NO	Total
Aktual YES	<i>True Positif</i> (TP)	<i>False Negatif</i> (FN)	<i>Positif</i> (P)
Aktual NO	<i>False Positif</i> (FP)	<i>True Negatif</i> (TN)	<i>Negatif</i> (N)
Total	P'	N'	P+N

Rumus mencari Presisi

$$presisi = \frac{TP}{TP + FP}$$

Rumus mencari Recall

$$recall = \frac{TP}{TP + FN}$$

Rumus mencari Akurasi

$$akurasi = \frac{TP + TN}{TP + FP + TN + FN}$$

3. HASIL DAN PEMBAHASAN

3.1. Hasil

Klasifikasi dataset pada penyakit *jantung* dengan menerapkan 2 algoritma yaitu *Naïve Bayes* dan algoritma *K-NN* memberikan hasil yang tidak sama. Berikut hasil *Confusion Matrix* nya.

Tabel 3. *Naïve Bayes*

Kelas		Nilai Aktual	
		Ada Penyakit	Tidak Ada Penyakit
Nilai Prediksi	Ada Penyakit	390	109
	Tidak Ada Penyakit	74	452

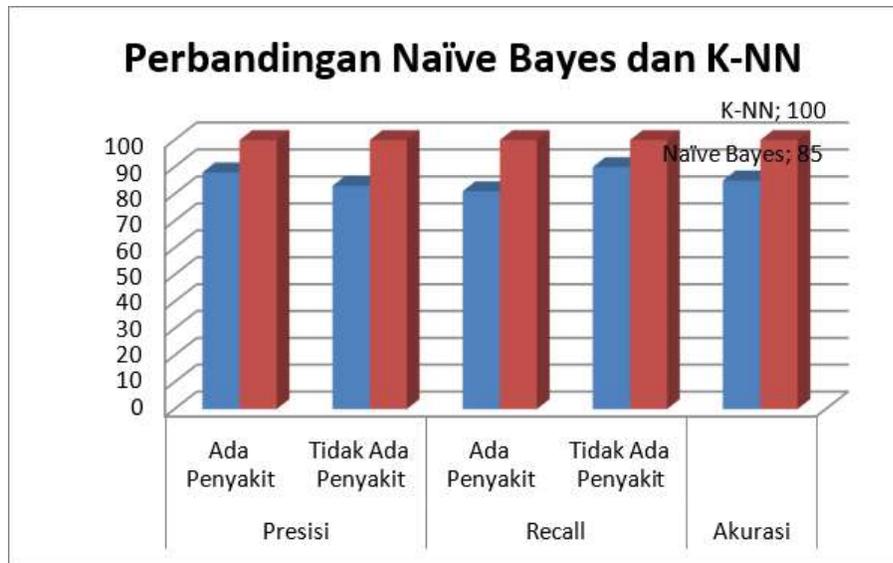
Berdasarkan Tabel 3 di atas dapat dilihat bahwa algoritma *Naïve Bayes* menyatakan bahwa sebanyak 390 orang terkena penyakit diprediksi sebagai terkena penyakit, sebanyak 109 orang tidak terkena penyakit diprediksi sebagai terkena penyakit, sebanyak 74 orang terkena penyakit diprediksi sebagai tidak terkena penyakit, dan sebanyak 452 orang tidak terkena penyakit diprediksi sebagai tidak terkena penyakit.

Tabel 4. *K-Nearest Neighbor*

Kelas		Nilai Aktual	
		Ada Penyakit	Tidak Ada Penyakit
Nilai Prediksi	Ada Penyakit	429	70
	Tidak Ada Penyakit	95	431

Sedangkan berdasarkan algoritma *K-Nearest Neighbor* pada table 4 menyatakan bahwa sebanyak 429 orang terkena penyakit diprediksi sebagai terkena penyakit, sebanyak 70 orang tidak terkena penyakit diprediksi sebagai terkena penyakit, sebanyak 95 orang terkena penyakit diprediksi sebagai tidak terkena penyakit, dan sebanyak 431 orang tidak terkena penyakit diprediksi sebagai tidak terkena penyakit.

Pada penerapan algoritma *Naïve Bayes* diperoleh nilai akurasi mencapai sebesar 85% sedangkan algoritma *K-Nearest Neighbor* memperoleh nilai akurasi sebesar 100%. Untuk nilai presisi terkena penyakit pada algoritma *Naïve Bayes* mencapai sebesar 88% dan untuk tidak terkena penyakit sebesar 83% sedangkan algoritma *K-Nearest Neighbor* untuk terkena penyakit mencapai sebesar 100% dan untuk tidak terkena penyakit sebesar 100%. Nilai *recall* untuk terkena penyakit pada algoritma *Naïve Bayes* mencapai sebesar 81% dan untuk tidak terkena penyakit sebesar 90% sedangkan algoritma *K-Nearest Neighbor* untuk terkena penyakit sebesar 100% dan untuk tidak terkena penyakit sebesar 100%.



Gambar 1. Nilai-Nilai Presisi, Recall dan Akurasi *Naives Bayes* dan *K-NN*

3.1. Perbandingan Hasil Kerja Kedua Metode

Hasil kerja atau prediksi yang didapatkan pada penggunaan 2 algoritma yaitu *Naïve Bayes* dan algoritma *K- Nearest Neighbor* membuktikan bahwa dua algoritma tersebut baik dalam melakukan klasifikasi penyakit *jantung*. Berikut ini ditampilkan perbandingan hasil rata-rata nilai presisi, recall dan akurasi.

Tabel 4. Perbandingan Rata-rata Nilai Presisi, Recall dan Akurasi Kedua algoritma

Algoritma	Presisi	Recall	Akurasi
Naïve Bayes	85,5%	85,5%	85%
K-Nearest Neighbor	100%	100%	100%



Gambar 2. Perbandingan Rata-rata Presisi, Recall dan Akurasi

Berdasarkan Gambar 2 terlihat bahwa nilai Presisi dan Recall yang diperoleh algoritma *Naïve Bayes* dan algoritma *K-Nearest Neighbor* terdapat perbedaan yang cukup stabil sebagai prediksi penyakit *jantung* karena algoritma *Naïve Bayes* 85% keatas sedangkan algoritma *K-Nearest Neighbor* mencapai 100% yang artinya memiliki perbedaan atau selisih sekitar sebesar 15%.

4. KESIMPULAN DAN SARAN

Pada penelitian ini dapat diketahui bahwa pengklasifikasian dataset penyakit *jantung* dengan menggunakan algoritma *Naïve Bayes* dan *K-NN* cukup baik. Terlihat dari nilai akurasi, presisi dan *recall* dari kedua algoritma yang melebihi nilai 80%. Algoritma *Naïve Bayes* memberikan rata-rata akurasi sebesar 85% yang bernilai lebih kecil dibandingkan dengan *K-NN* yang dapat memberikan rata-rata bernilai sempurna yaitu 100%. Rata-rata nilai presisi dan *recall* dari *Naïve Bayes* bernilai sama yaitu 85,5%, sedangkan rata-rata nilai presisi dan *recall* dari *K-NN* bernilai sempurna atau 100%. Algoritma *K-NN* memberikan hasil sempurna yang memiliki indikasi bahwa program tersebut bisa saja tidak belajar sehingga tidak dapat mendeteksi dengan benar. Dapat disimpulkan bahwa penggunaan algoritma *Naïve Bayes* pada klasifikasi penyakit *jantung* akan sangat lebih baik dibandingkan dengan menggunakan algoritma *K-NN*.

5. DAFTAR PUSTAKA

- [1] T. Rusdiana, N. A. Putriana, I. Sopyan, D. Gozali, and P. Husni, "Pemberian Pemahaman Mengenai Sediaan herbal yang Berfungsi untuk Pemeliharaan Kesehatan jantung dan Ginjal di Desa Cibeusi, Sumedang, Jawa Barat," *J. Pengabd. Kpd. Masy.*, vol. 4, no. 6, pp. 139–141, 2019.
- [2] T. L. M. J. (2012), "Pelatihan Aerobik untuk Kebugaran Paru Jantung Bagi Lansia," *Budid. AYAM RAS PETELUR (Gallus sp.)*, vol. 21, no. 58, pp. 99–104, 1990, [Online]. Available: <https://www.unhcr.org/publications/manuals/4d9352319/unhcr-protection-training-manual-european-border-entry-officials-2-legal.html?query=excom> 1989
- [3] S. Sahar, "Analisis Perbandingan Metode K-Nearest Neighbor dan Naïve Bayes Classifier Pada Dataset Penyakit Jantung," *Indones. J. Data Sci.*, vol. 1, no. 3, pp. 79– 86, 2020, doi: 10.33096/ijodas.v1i3.20.
- [4] S. A., "Penyakit Jantung Katup di Indonesia : masalah yang hampir terlupakan. J Kardiologi Indones," *J Kardiologi Indones*, vol. 33, no. 4, pp. 205–208, 2012.
- [5] A. A. Setiawan, M. Panggabean, M. Yamin, and S. Setiati, "Kesintasan Lima Tahun Pasien Penyakit Jantung Koroner Tiga Pembuluh Darah dengan Diabetes Melitus yang Menjalani Bedah Pintas Koroner, Intervensi Koroner Perkutan atau Medikamentosa di Rumah Sakit dr. Cipto Mangunkusumo," *J. Penyakit Dalam Indones.*, vol. 3, no. 2, p. 60, 2017, doi: 10.7454/jpdi.v3i2.10.
- [6] N. Pudiarianti, I. D. Pramantara, and Z. Ikawati, "Faktor-Faktor yang Mempengaruhi Kualitas Hidup Pasien Gagal Jantung Kronik," *J. Manaj. dan Pelayanan Farm.*, vol. 5, no. 4, pp. 259–266, 2015, [Online]. Available: <https://jurnal.ugm.ac.id/jmpf/article/view/29453>
- [7] A. Riski, "Analisis Komparasi Algoritma Klasifikasi Data Mining Untuk Prediksi Penderita Penyakit Jantung," *J. Tek. Inform. Kaputama*, vol. 3, no. 1, pp. 22–28, 2019, [Online]. Available: <https://jurnal.kaputama.ac.id/index.php/JTIK/article/view/141/156>
- [8] Y. Mardi, "Data Mining : Klasifikasi Menggunakan Algoritma C4.5," *Edik Inform.*, vol. 2, no. 2, pp. 213–219, 2017, doi: 10.22202/ei.2016.v2i2.1465.
- [9] Syarli and A. A. Muin, "Metode Naive Bayes Untuk Prediksi Kelulusan (Studi Kasus: Data Mahasiswa Baru Perguruan Tinggi)," *J. Ilm. Ilmu Komput.*, vol. 2, no. 1, pp. 22– 26, 2016, [Online]. Available: <http://ejournal.fikom-unasman.ac.id>
- [10] U. O. R. Permatasari, W. J. Shudiq and M. Jasri, "Prediksi Kelayakan Mahasiswa sebagai Penerima Beasiswa Bank Indonesia pada Tahap Seleksi Administrasi di Universitas Nurul Jadid Menggunakan Algoritma K Nearest Neighbor," *Journal of Electrical Engineering and Computer (JEECOM)*, vol. 6, no. 1, pp. 252-260, 2024.
- [11] A. hudawi, W. J. Shudiq and M. F. Rahman, "Penentuan Metode Terbaik Dalam Menentukan Jenis Pohon Pisang Menurut Tekstur Daun (Metode K-NN dan SVM)," *Jurnal Teknologi dan Manajemen Informatika*, vol. 6, no. 2, pp. 128-136, 2020.
- [12]
- [13] W. Yustanti, "Algoritma K-Nearest Neighbour untuk Memprediksi Harga Jual Tanah," *J. Mat. Stat. dan komputasi*, vol. 9, no. 1, pp. 57–68, 2012.
- [14] S. Dewi, "Komparasi 5 Metode Algoritma Klasifikasi Data Mining Pada Prediksi Keberhasilan Pemasaran Produk Layanan Perbankan," *Techno Nusa Mandiri*, vol. 13, no. 1, pp. 60–66, 2016.
- [15] S. Hendrian, "Algoritma Klasifikasi Data Mining Untuk Memprediksi Siswa Dalam Memperoleh Bantuan Dana Pendidikan," *Fakt. Exacta*, vol. 11, no. 3, pp. 266–274, 2018, doi: 10.30998/faktorexacta.v11i3.2777.