Jurnal Sistem Informasi dan Sistem Komputer, Vol. 7, No. 2, Juli 2022
ISSN: 2715-906X (Online)
10.51717/simkom.v7i2.96

PERBANDINGAN IMPLEMENTASI ALGORITMA NAÏVE BAYES DAN K-NEAREST NEIGHBOR PADA KLASIFIKASI PENYAKIT HATI

Anita Desiani
anita_desiani@unsri.ac.id
Universitas Sriwijaya

Abstrak

Hati adalah organ kelenjar terbesar dengan berat kira-kira 1200-1500 gram. Hati dapat terkena berbagai macam penyakit. Untuk mengetahui jumlah rerataan manusia yang terkena penyakit hati, maka kita dapat melakukan klasifikasi terhadap penyakit hati. Penelitian ini bertujuan untuk membandingkan kemudian menyimpulkan algoritma terbaik yang dapat digunakan dalam melakukan klasifikasi penyakit hati. Adapun algoritma yang dibandingkan adalah Naïve Bayes dan K-Nearest Neighbor (K-NN). Hasil dari penelitian ini menyatakan algoritma K-NN dan Naïve Bayes memperoleh nilai lebih dari 80% baik nilai akurasi, presisi maupun recall. Algoritma K-NN memberikan nilai akurasi, presisi, serta recall yang lebih tinggi dibandingkan dengan algoritma Naïve Bayes. Maka algoritma terbaik yang dapat digunakan adalah K-NN.

Kata kunci: Hati, Naïve Bayes, K-Nearest Neighbor, Perbandingan, Penyakit.

Abstract

The liver is the largest glandular organ weighing about 1200-1500 grams. The liver can be affected by various diseases. To find out the average number of people affected by liver disease, we can classify liver disease. This study aims to compare and then conclude the best algorithm that can be used in classifying liver disease. The algorithms compared are Naïve Bayes and K-Nearest Neighbor (K-NN). The results of this study stated that the K-NN and Naïve Bayes algorithms obtained a value of more than 80% both accuracy, precision and recall values. The K-NN algorithm provides higher accuracy, precision, and recall values than the Naïve Bayes algorithm. Then the best algorithm that can be used is K-NN.

Keywords: *Liver, Naïve Bayes, K-Nearest Neighbor, Perbandingan, Diseases.*

1. Pendahuluan

Hati adalah organ kelenjar terbesar dengan berat kira-kira 1200-1500 gram [1]. Hati memiliki banyak fungsi yang kompleks dan beragam, fungsi hati adalah sebagai filter semua darah yang datang dari usus melalui vena porta, kemudian menyimpan dan mengubah bahan-bahan makanan yang diterima vena porta [2]. Hati dapat meregenerasi dirinya sendiri dengan cara meningkatkan kecepatan mitosis hepatosit dan meningkatkan diferensiasi sel punca menjadi hepatosit atau kolangiosit [3]. Hati juga dapat terkena penyakit, jenis-jenis penyakit hati yang umum antara lain yaitu hepatitis, sirosis, kanker hati atau hepatoma, abses hati, kolesistitis dan perlemakan hati non alkoholik [4]. Salah satu cara untuk mendeteksi penyakit hati adalah dengan melakukan pengklasifiksian. Penklasifikasian yang diterapkan dalam penelitian ini adalah data mining.

Pengklasifikasian mengenai penyakit hati sebelumnya pernah dilakukan oleh Institut Teknologi Adhi Tama Surabaya [5] yang melakukan klasifikasi pada dataset penyakit hati menggunakan algoritma SVM (Support Vector Machines), K-NN dan Naïve Bayes. Hasilnya, algoritma yang mencapai tingkat akurasi tertinggi adalah SVM yaitu sebesar 84,62% diikuti Naïve Bayes sebesar 82,42% dan K-NN sebesar 63,74-68,13%. Ada begitu banyak metode atau algoritma yang dapat digunakan, dan

Jurnal Sistem Informasi dan Sistem Komputer, Vol. 7, No. 2, Juli 2022
ISSN: 2715-906X (Online)
10.51717/simkom.v7i2.96

algoritma Naïve Bayes merupakan salah satu algoritma pembelajaran induktif yang paling efektif dan efisien dalam machine learning dan data mining. Naïve Bayes memiliki kelebihan antara lain, sederhana, cepat, dan berakurasi tinggi [6]. Adapun kelemahan dari penggunaan algortima Naïve Bayes yaitu lamanya waktu yang digunakan untuk melakukan prediksi [7]. Selain Naïve Bayes ada juga K-NN (K-Nearest Neighbor), Kelebihan algoritma K-NN, yaitu sederhana dan mudah dipelajari, dapat memberikan hasil prosentase yang cukup bagus [8]. Kekurangan dari K-NN masih perlu penentuan nilai k dan untuk pemilihan atribut terbaik [9], keterbatasan memori, komputasi kompleks, dan terpengaruh akan data-data yang tidak relevan.

Penelitian ini menggunakan algoritma Naïve Bayes dan K-NN dalam melakukan pengklasifikasian terhadap dataset yang menentukan apakah seseorang terserang penyakit hati atau tidak. Dalam penelitian ini akan dihitung nilai akurasi, recall, serta presisinya. Kemudian hasil yang diperoleh akan dibandingkan satu sama lain untuk menentukan algoritma terbaik yang dapat digunakan untuk melakukan klasifikasi terhadap dataset penyakit hati.

2. Metode

Penelitian ini menggunakan dua algoritma untuk melakukan klasifikasi dataset penyakit hati yaitu algoritma Naïve Bayes dan algoritma K-NN. Pertama kita harus mencari terlebih dahulu dataset yang akan digunakan.

2.1 Deskripsi Data

Data yang digunakan untuk penelitian ini adalah dataset yang diperoleh dari situs Kaggle (https://www.kaggle.com/datasets/) dengan format csv. Terdapat 14 atribut dimana 13 atribut merupakan ciri dan satu lainnya atribut target. Target terdiri atas 2 label, yaitu 0 = tidak ada penyakit dan 1 = ada penyakit. Di dalam dataset ini terdapat sebanyak 1025 data. Atribut-atribut yang digunakan adalah age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal, target. Dari total 13 atribut ini, terdiri atas 2 jenis tipe data, yaitu int64 dan float64. Atribut yang bertipe data float64 adalah oldpeak dan untuk 12 atribut lainnya bertipe data int64.

Tabel 1. Penjelasan dan Arti dari Atribut

Attribute	*Feature Meaning*
Target/Target	0 = tidak ada penyakit, 1 = ada penyakit
Age/Umur	
Sex/Jenis Kelamin	0 = female, 1 = male
Chest Pain/ Nyeri Dada	0 = tidak, 1 = nyeri ringan, 2 = nyeri, 3 = sangat nyeri
Trestbps/ Tekanan Darah Rendah	(94 – 200)
Cholestrol/ Kolestrol	(126 – 564)
Fbs / Gula Darah Normal	0 = salah, 1 = benar
Restecg/ Elektrokardiografi	0 = tidak sehat, 1 = cukup sehat, 2 = sehat
Thalach/ Detak Jantung Maksimum Tercapai	(71 – 202)

Jurnal Sistem Informasi dan Sistem Komputer, Vol. 7, No. 2, Juli 2022
ISSN: 2715-906X (Online)
10.51717/simkom.v7i2.96

Exang/ Kestabilan Induksi Angina	0 = tidak stabil, 1 = stabil
Oldpeak / Depresi ST yang diinduksi oleh olahraga relatif terhadap istirahat	(0 – 6.2)
Slope/ Kemiringan segmen	(0 – 2)
Ca / nomor pembuluh darah utama	(0 – 4)
Thal /	0 = normal, 1 = cacat tetap, 2 = cacat reversibel

2.2 *Preprocessing* Data

Tahap ini merupakan salah satu tahapan yang sangat penting. Proses cleaning mencakup antara lain membuang duplikasi data, memeriksa data yang inkonsisten, dan memperbaiki kesalahan pada data, seperti kesalahan cetak [10]. Pada penelitian ini tidak di dapati atribut yang harus dihapus atau dibuang. Sehingga dapat ditentukan bahwa penelitian ini menggunakan 14 atribut dalam melakukan klasifikasi.

Teknik yang digunakan dalam pengklasifikasian ini adalah presentase split dengan ratio 8:2. Data dibagi menjadi 2 bagian yaitu data training dan data testing. Data training merupakan data yang digunakan dalam proses latih atau pembelajaran sedangkan data testing merupakan data yang akan diuji. Sebanyak 80% data akan menjadi data training dan 20% data akan menjadi data testing.

2.3 Algoritma *Naïve Bayes*

Naïve Bayes merupakan pengklasifikasian dengan metode probabilitas dan statistik yang dikemukakan oleh ilmuwan Inggris Thomas Bayes [11]. Langkah-langkah untuk melakukan klasifikasi menggunakan algoritma Naïve Bayes adalah sebagai berikut:

1. Menghitung jumlah kategori dari setiap variabel
2. Menghitung peluang pada setiap kategori
3. Menentukan frekuensi atau jumlah kemunculan pada setiap kategori
4. Menentukan kategori dengan nilai maksimal

Perhitungan *Naïve Bayes* dapat dilakukan dengan rumus [12]:

$$P(H \mid X) = \frac{P(X \mid H)\, P(H)}{P(X)} \quad \ldots(1)$$

dimana:

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

Jurnal Sistem Informasi dan Sistem Komputer, Vol. 7, No. 2, Juli 2022
ISSN: 2715-906X (Online)
10.51717/simkom.v7i2.96

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

Algoritma *K-Nearest Neighbor* (KNN) adalah sebuah metode untuk melakukan klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut. Ketepatan algoritma KNN sangat dipengaruhi oleh ada atau tidaknya fitur-fitur yang relevan [13]. Tahap perhitungan dengan algoritma K-NN [14], sebagai berikut:

1. Tentukan banyaknya tetangga k (sebaiknya ganjil)
2. Hitung arak dari data untuk dibandingkan dengan dataset *training*, dapat dihitung dengan persamaan jarak Euclidean

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

dimana:
dist(p,q) = jarak antara p dan 1
p_i = nilai ke-i pada data p
q_i = nilai ke-i pada data q

3. Atur urutan menaik dari jarak (urutkan dari kecil ke besar) dan pilih himpunan k paling sedikit dari dataset terkecil.
4. Tentukan bahwa jawaban dengan data yang akan diprediksi adalah kelompok data yang memiliki jumlah k pertama dari kumpulan data terbesar.
5. Tetapkan kelas kelas terdekat dengan titik pertimbangan.

2.5 Evaluasi Hasil

Confusion matrix adalah sebuah tabel yang menyatakan jumlah data uji yang diklasifikasikan secara benar dan jumlah data uji yang diklasifikasikan secara salah. *Confusion matrix* merupakan sebuah matriks yang menampilkan visualisasi kinerja dari algoritma klasifikasi menggunakan data dalam matriks yang membagi klasifikasi prediksi dalam bentuk *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN). Adapun bentuk confusion matriks untuk klasifikasi dua kelas, dapat dilihat pada Tabel 2 [15].

Tabel 2. *Confusion Matrix*

Kelas	Prediksi Yes	Prediksi No	Total
Aktual *Yes*	*True Positive* (TP)	*False Negative* (FN)	*Positive* (P)
Aktual *No*	*False Positive* (FP)	*True Negative* (TN)	*Negative* (N)
Total	P’	N’	P+N

Rumus Mencari Akurasi adalah :

$$akurasi = \frac{TP+TN}{TP+FP+TN+FN} \quad \ldots(3)$$

Rumus Mencari Presisi adalah :

$$presisi = \frac{TP}{TP+FP} \quad \ldots(4)$$

Rumus Mencari *Recall* adalah :

$$recall = \frac{TP}{TP+FN} \quad \ldots(5)$$

Jurnal Sistem Informasi dan Sistem Komputer, Vol. 7, No. 2, Juli 2022
ISSN: 2715-906X (Online)
10.51717/simkom.v7i2.96

3. Hasil dan Pembahasan

3.1 Hasil

Pada klasifikasi penyakit hati dengan menerapkan algoritma *Naïve Bayes* dan KNN menunjukkan hasil yang berbeda. Berikut ini *Confusion Matrix* yang dihasilkan.

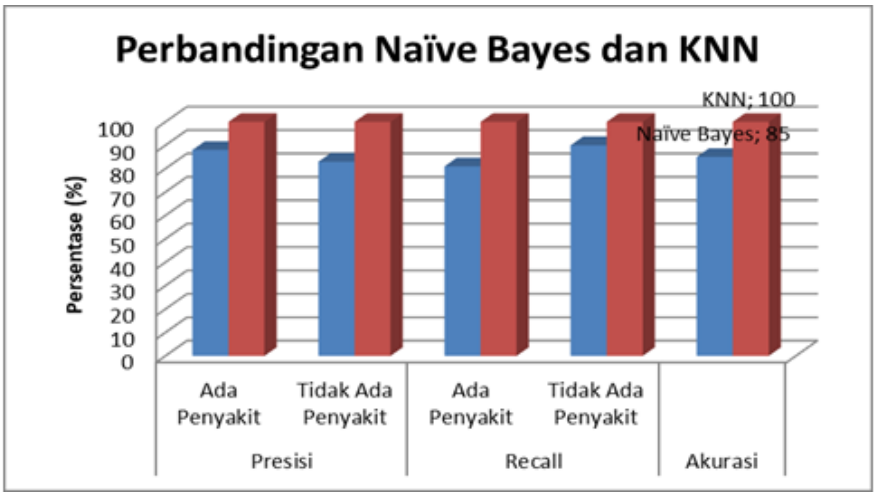
Tabel 3. *Confusion Matrix Naïve Bayes* dan *K-Nearest Neighbor*

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

K-Nearest Neighbor			
Kelas		Nilai Aktual	
		Ada Penyakit	Tidak Ada Penyakit
Nilai Prediksi	Ada Penyakit	429	70
	Tidak Ada Penyakit	95	431

Dari Tabel 3 dapat dilihat bahwa algoritma Naïve Bayes memprediksi 390 orang terkena penyakit sebagai terkena penyakit, 74 orang terkena penyakit sebagai tidak terkena penyakit, 109 orang tidak terkena penyakit sebagai terkena penyakit, 452 orang tidak terkena penyakit sebagai tidak terkena penyakit. Sedangkan berdasarkan algoritma K-Nearest Neigbhor memprediksi 429 orang terkena penyakit sebagai terkena penyakit, 95 orang terkena penyakit sebagai tidak terkena penyakit, 70 orang tidak terkena penyakit sebagai terkena penyakit, 431 orang tidak terkena penyakit sebagai tidak terkena penyakit.

Pada penerapan algoritma Naïve Bayes memperoleh akurasi sebesar 85%. Nilai presisi untuk terkena penyakit sebesar 88% dan untuk tidak terkena penyakit sebesar 83%. Nilai recall untuk terkena penyakit sebesar 81% dan untuk tidak terkena penyakit sebesar 90%. Pada penerapan algoritma K-Nearest Neigbhor memperoleh akurasi sebesar 100%. Nilai presisi untuk terkena penyakit sebesar 100% dan untuk tidak terkena penyakit sebesar 100%. Nilai recall untuk terkena penyakit sebesar 100% dan untuk tidak terkena penyakit sebesar 100%.



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

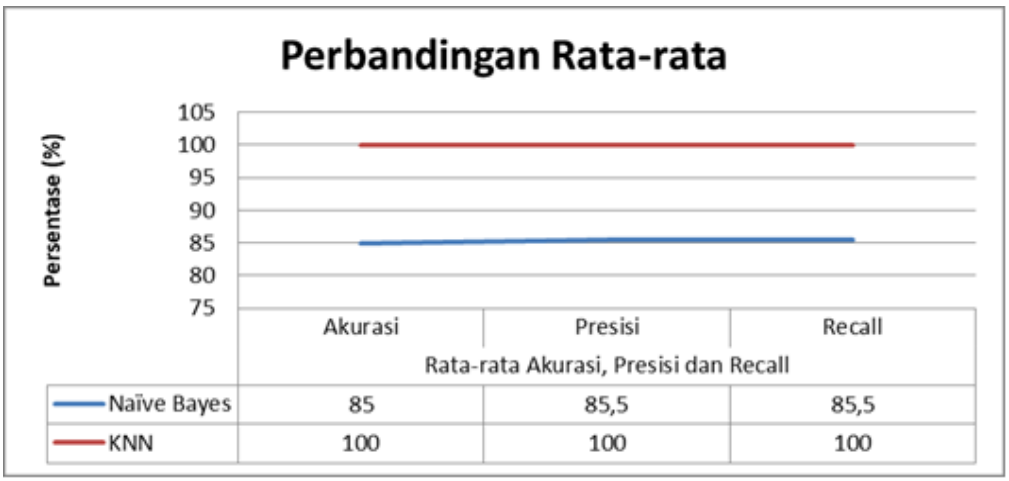
Jurnal Sistem Informasi dan Sistem Komputer, Vol. 7, No. 2, Juli 2022
ISSN: 2715-906X (Online)
10.51717/simkom.v7i2.96

3.2 Perbandingan Hasil Kerja Kedua Metode

Hasil prediksi yang diperoleh pada penggunaan algoritma Naïve Bayes dan K-Nearest Neighbor meunjukkan bahwa kedua algoritma tersebut bekerja dengan baik dalam mengklasifikasikan penyakit hati. Pada algoritma Naïve Bayes, terlihat variabel yang paling berpengaruh karena algoritma ini menghitung peluang setiap kejadian. Berikut ini perbandingan rata-rata nilai rata-rata akurasi, presisi, dan recall.

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

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



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

Dari Gambar 2 dapat dilihat bahwa nilai presisi dan recall yang dihasilkan algoritma Naïve Bayes dan K-Nearest Neighbor memiliki perbedaan yang cukup stabil dan lumayan jauh sebagai prediksi penyakit hati karena K-NN mencapai 100% sedangkan Naïve Bayes 85% keatas yang artinya memiliki perbedaan atau selisih sekitar 15%.

4. Kesimpulan dan Saran

Hasil dari penelitian ini menyatakan algoritma K-NN dan *Naïve Bayes* memperoleh nilai lebih dari 80% baik nilai akurasi, presisi maupun *recall*. Algoritma K-NN memberikan nilai akurasi, presisi, serta *recall* yang lebih tinggi dibandingkan dengan algoritma *Naïve Bayes*. Maka algoritma terbaik yang dapat digunakan adalah K-NN.

Daftar Pustaka

[1] Rosida, A., "Pemeriksaan Laboratorium Penyakit Hati", Berkala Kedokteran, Vol.12, No.1, pp.123, 2016, doi: 10.20527/jbk.v12i1.364. https://doi.org/10.20527/jbk.v12i1.364

[2] Pujiyanta, A., & Pujiantoro, A., "Sistem Pakar Penentuan Jenis Penyakit Hati dengan Metode Inferensi Fuzzy Tsukamoto", Jurnal Informatika, Vol.6, No.1, pp.617–629, Januari 2012. http://journal.uad.ac.id/index.php/JIFO/article/view/2787/1698

[3] Safithri., "Mekanisme Regenerasi Hati secara Endogen pada Fibrosis Hati", Mekanisme Regenerasi Hati Secara Endogen Pada Fibrosis Hati, Vol.2, No.4, pp.9–26, Februari 2018.

[4] Falatehan, A. I., Hidayat, N., & Brata, K. C., "Sistem Pakar Diagnosis Penyakit Hati Menggunakan Metode Fuzzy Tsukamoto Berbasis Android", Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer (J-PTIIK) Universitas Brawijaya, Vol.2, No.8, pp.2373–2381, Agustus 2018.
[5] Prabiantissa, C. N., "Klasifikasi pada Dataset Penyakit Hati Menggunakan Algoritma Support Vector Machine, K-NN, dan Naïve Bayes", Seminar Nasional Teknik Elektro, Sistem Informasi, Dan Teknik Informatika, Vol., No., pp.219–224, Juni 2021, doi: 10.31284/p.snestik.2021.1818.
[6] Muslim, A. M., Syarifah, A., "Pemanfaatan Naïve Bayes Untuk Merespon Emosi Dari Kalimat Berbahasa Indonesia", Unnes Journal of Mathematics, Vol. 4, No. 2, pp. 147-156, Jan 2015.
[7] Rosandy, T. "PERBANDINGAN METODE NAIVE BAYES CLASSIFIER DENGAN METODE DECISION TREE (C4.5) UNTUK MENGANALISA KELANCARAN PEMBIAYAAN (Study Kasus : KSPPS / BMT AL-FADHILA", Jurnal Teknologi Informasi Magister Darmajaya, Vol.2, No.1, pp.52-62, Mei 2016.
[8] Hidayanti, W. P., Yahya, "Penerapan Algoritma K-Nearest Neighbor Untuk Klasifikasi Efektivitas Penjualan Vape (Rokok Elektrik) pada "Lombok Vape On"", Infotek : Jurnal Informatika dan Teknologi, Vol.3, No.2, pp.104-114, Juli 2020.
[9] Bode, A., "K-Nearest Neighbor Dengan Feature Selection Menggunakan Backward Elimination Untuk Prediksi Harga Komoditi Kopi Arabika", ILKOM Jurnal Ilmiah, Vol.9, No.2, pp.188-195, Agustus 2017, doi: 10.33096/ilkom.v9i2.139.188-195.
[10] Yuli Mardi, "Data Mining : Klasifikasi Menggunakan Algoritma C4 . 5 Data mining merupakan bagian dari tahapan proses Knowledge Discovery in Database (KDD)" , Jurnal Edik Informatika, Vol. 2, No.2, pp.213–219, 2019.
[11] Syarli, & Muin, A. A., "Metode Naive Bayes Untuk Prediksi Kelulusan", Jurnal Ilmiah Ilmu Komputer, Vol.2, No.1, pp.22–26, April 2016. https://media.neliti.com/media/publications/283828-metode-naive-bayes-untuk-prediksi-kelulu-139fcfea.pdf
[12] Bustami, "Penerapan Algoritma Naive Bayes", Jurnal Informatika, Vol.8, No.1, pp.884–898, Januari 2014.
[13] Yustanti, W., "Algoritma K-Nearest Neighbour untuk Memprediksi Harga Jual Tanah", Jurnal Matematika Statistika Dan Komputasi, Vol.9, No.1, pp.57–68, Juli 2012.
[14] Dewi, S. "Pada Prediksi Keberhasilan Pemasaran Produk Layanan Perbankan", Techno Nusa Mandiri, Vol.13, No.1, pp.60–66, Maret 2016.
[15] Hendrian, S. "Algoritma Klasifikasi Data Mining Untuk Memprediksi Siswa Dalam Memperoleh Bantuan Dana Pendidikan", Faktor Exacta, Vol.11, No.3, pp.266–274, 2018, doi:https://doi.org/10.30998/faktorexacta.v11i3.2777