Journal of Information System Research (JOSH)
Volume 6, No. 2, Januari 2025, pp 997−1005
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i2.6480

Perbandingan Algoritma Naïve Bayes dan K-Nearest Neighbor (K-NN) Untuk Klasifikasi Penyakit Gagal Jantung

Firman Zahri, Fitri Insani*, Jasril, Lola Oktavia

Sains dan Teknologi, Teknik Informatika, Universitas Islam Negeri Sultan Syarif Kasim Riau, Pekanbaru
Panam, Jl. HR. Soebrantas No.Km. 15, RW.15, Simpang Baru, Kota Pekanbaru, Riau, Indonesia
Email: [1]11950115068@students.uin-suska.ac.id, [2,*]fitri.insani@uin-suska.ac.id, [3] jasril@uin-suska.ac.id, [4]lola.oktavia@uin-suska.ac.id Email Penulis Korespondensi: fitri.insani@uin-suska.ac.id
Submitted: **19/12/2024;** Accepted: **03/01/2025;** Published: **05/01/2025**

Abstrak−Penyakit yang dikenal sebagai gagal jantung, di mana jantung tidak dapat memompa darah dalam jumlah yang cukup untuk memenuhi kebutuhan tubuh akan oksigen dan nutrisi, tidak boleh dianggap enteng. Hal ini dapat mengakibatkan sejumlah gejala, seperti kelelahan, retensi cairan, dan dispnea. Federasi Jantung Dunia memperkirakan bahwa hingga 1,8 juta orang di Asia Tenggara menderita gagal jantung pada tahun 2014. Untuk perawatan yang cepat dan efisien, gagal jantung merupakan masalah medis yang perlu diidentifikasi. Penyakit ini berpotensi semakin memburuk jika tidak segera ditangani. Beberapa metode pembelajaran mesin dapat digunakan untuk membantu diagnosis dan kategorisasi penyakit ini. Salah satunya adalah algoritma populer yaitu Naive Bayes dan K-Nearest Neighbors. Naive Bayes merupakan algoritma pembelajaran mesin berbasis probabilitas yang sederhana tetapi sangat efisien, khususnya dalam aplikasi klasifikasi. K-Nearest Neighbors adalah membandingkan data yang ingin diprediksi dengan sejumlah data terdekatnya dalam ruang fitur berdasarkan jarak tertentu, seperti jarak Euclidean, Manhattan, atau lainnya. Penelitian ini dilakukan menggunakan Confusion Matrix untuk melakukan evaluasi sekaligus perbandingan antara algoritma Naive Bayes dan K-Nearest Neighbor dalam kategorisasi penyakit gagal jantung dengan mengumpulkan data yang berjumlah 918 data pasien gagal jantung yang berasal dari kaggle. Berdasarkan temuan penelitian, metode K-Nearest Neighbor mencapai skor akurasi 76%, sedangkan pendekatan Naive Bayes yang mencapai akurasi 90% menggunakan rasio 80:20.

Kata Kunci: Akurasi; Data Mining; Klasifikasi; Naïve Bayes; K-Nearest Neighbor; Gagal Jantung

Abstract−A condition known as heart failure, where the heart is unable to pump enough blood to meet the body's needs for oxygen and nutrients, should not be taken lightly. This can result in a number of symptoms, such as fatigue, fluid retention, and dyspnea. The World Heart Federation estimates that up to 1.8 million people in Southeast Asia suffered from heart failure in 2014. For prompt and efficient treatment, heart failure is a medical problem that needs to be identified. This disease has the potential to worsen if not treated immediately. Several machine learning methods can be used to help diagnose and categorize this disease. One of them is the popular algorithm, namely Naive Bayes and K-Nearest Neighbors. Naive Bayes is a simple but very efficient probability-based machine learning algorithm, especially in classification applications. K-Nearest Neighbors is comparing the data to be predicted with a number of its closest data in a feature space based on a certain distance, such as Euclidean distance, Manhattan, or others. This study was conducted using Confusion Matrix to evaluate and compare the Naive Bayes and K-Nearest Neighbor algorithms in the categorization of heart failure disease by collecting data totaling 918 heart failure patient data from kaggle. Based on the research findings, the K-Nearest Neighbor method achieved an accuracy score of 76%, while the Naive Bayes approach achieved 90% accuracy using a ratio of 80:20.

Keywords: Accuracy; Data Mining; Classification; Naïve Bayes; K-Nearest Neighbor; Heart Failure

1. PENDAHULUAN

Dalam dunia medis, Gagal jantung adalah suatu kondisi kesehatan yang menunjukkan tren peningkatan seiring waktu, dengan angka morbiditas yang terus bertambah dan tingkat mortalitas yang semakin tinggi, sehingga menjadi perhatian serius dalam dunia medis[1]. Apabila jantung tidak mampu memompa darah secara memadai untuk memenuhi kebutuhan oksigen dan nutrisi tubuh, hal ini dapat mengakibatkan gagal jantung, kondisi ini tidak bisa disepelekan. Hal ini dapat mengakibatkan berbagai gejala, termasuk mudah kelelahan, sesak napas, dan penumpukan cairan dalam tubuh yang biasa terjadi dibagian kaki, perut, tangan dan wajah (edema)[2].

Seperti dilansir World Heart Federation, Pada tahun 2014, jumlah kematian yang disebabkan oleh gagal jantung di wilayah Asia Tenggara dilaporkan mencapai angka 1,8 juta jiwa, menunjukkan tingkat keparahan kondisi ini sebagai masalah kesehatan yang signifikan. Setidaknya 883.447 orang di Indonesia menerima diagnosis gagal jantung pada tahun 2013, mayoritas dari mereka berada dalam rentang usia 55–64 tahun. Penyakit gagal jantung jantung juga menjadi penyebab kematian yang tinggi, diperkirakan 45% dari seluruh angka kematian di Indonesia[3].

Faktor-faktor yang dapat meningkatkan gagal jantung seperti umur, nyeri data, pekerjaan, tingkat gagal jantung, resiko kematian, dan kesehatan mental[4]. Penyakit jantung iskemik dan hipertensi merupakan penyebab utama gagal jantung, yang bermanifestasi sebagai berbagai gejala klinis yang disebabkan oleh berbagai penyakit penyerta[5]. Gagal jantung dapat menyebabkan gejala fisik seperti dispnea, kelelahan, edema, dan kehilangan selera makan. Tanpa disadari, depresi dan kecemasan yang berlebihan termasuk faktor gejala penyakit gagal jantung yangdapat mempengaruhi kualtas hidup[6].

Kemajuan teknologi di bidang kecerdasan buatan dan pembelajaran mesin telah membuka peluang baru dalam menganalisis data kesehatan, termasuk dalam diagnosis penyakit. Metode pembelajaran mesin mampu

Copyright © 2025 The Author, Page 997
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Journal of Information System Research (JOSH)
Volume 6, No. 2, Januari 2025, pp 997−1005
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i2.6480

memproses data besar dan kompleks untuk menemukan pola yang relevan, sehingga dapat digunakan untuk membantu para profesional medis dalam pengambilan keputusan[7].

Dalam permasalahan tersebut, Penyakit gagal jantung membutuhkan deteksi secara cepat dan akurat. Machine learning dapat digunakan untuk mendeteksi suatu penyakit dengan efektif dan efesien. Gagal jantung melibatkan data yang sangat kompleks dan beragam, termasuk hasil tes laboratorium, citra medis, dan data riwayat kesehatan pasien. Machine learning mampu menganalisis volume data yang besar ini dengan cepat dan efisien, menentukan pola yang samar oleh analisis manual.

Tindakan menemukan data baru dengan mencari pola dalam jumlah besar data menggunakan kriteria yang telah ditentukan sebelumnya yang dapat diantisipasi untuk mengatasi keadaan tertentu dikenal sebagai penambangan data. Proses analisis data mining tujuannya untuk mencari pola dalam sejumlah besar data dan menghasilkan informasi yang dapat digunakan. Data mining serangkaian proses yang digunakan untuk mencari nilai tambahan dari data, yang merupakan pengetahuan yang sebelumnya tidak dapat diketahui dengan cara manual dan dapat dimanfaatkan[8], [9], [10]. Data mining mencakup beberapa seperti estimasi, prediksi, klasifikasi, klastering, dan asosiasi[11].

Klasifikasi adalah suatu metode dalam pengolahan data yang digunakan untuk mengelompokkan data ke dalam beberapa kategori tertentu berdasarkan kriteria yang telah ditentukan[12]. Dengan melakukan klasifikasi, dapat ditentukan tingkat keakuratan data yang diperoleh. Klasifikasi dilakukan dengan mempelajari sebuah fungsi yang menghubungkan setiap atribut atau karakteristik dengan label kelas tertentu tertentu yang telah ditetapkan. Model yang dihasilkan digunakan untuk mengelompokkan data baru [8]. Dalam kasus ini, algoritma Naïve Bayes dan K-Nearest Neighbors (KNN) sering digunakan sebagai pendekatan populer dalam klasifikasi data medis, termasuk untuk mendeteksi penyakit gagal jantung. Kedua algoritma ini memiliki keunggulan dan karakteristik unik yang dapat disesuaikan dengan kebutuhan dataset.

Naïve Bayes adalah algoritma pengklasifikasi yang mengaplikasikan metode statistik dan probabilistik sederhana berdasarkan teorema Bayes dengan asumsi adanya independensi yang kuat[13]. Algoritma ini menggunakan teorema Bayes sambil mengasumsikan bahwa setiap atribut bersifat independen atau tidak saling bergantung, yang ditentukan oleh nilai variabel kelas[12].

Teknik klasifikasi objek yang disebut K-Nearest Neighbors (KNN) menggunakan data pelatihan yang paling dekat dengan item tersebut. Properti data setiap dimensi ditampilkan saat data pelatihan diproyeksikan ke ruang multidimensi. Klasifikasi data latih kemudian digunakan untuk memisahkan area ini menjadi beberapa bagian.[14]. Algoritma K-Nearest Neighbors (K-NN) merupakan metode klasifikasi berbasis jarak yang menghitung jarak antara data uji dan data latih untuk menentukan tetangga terdekat dengan nilai terkecil. Algoritma ini digunakan untuk mengelompokkan objek yang belum diteliti berdasarkan propertinya, dengan memanfaatkan sampel data latih sebagai acuan. Parameter K pada algoritma K-NN merepresentasikan jumlah tetangga terdekat yang memiliki kemiripan paling tinggi dengan data yang dianalisis, sehingga menjadi faktor kunci dalam proses klasifikasi. Dengan banyaknya data, K-Nearest Neighbor dapat menghasilkan klasifikasi yang akurat[15]. Mencari K data terdekat dari data yang akan diprediksi merupakan langkah awal dalam pengoperasian algoritma ini. Berdasarkan mayoritas kelas atau nilai target yang ditemukan pada K data terdekat, kelas atau nilai target dari data tersebut kemudian diprediksi. K-NN termasuk dalam bidang supervisi learning, yang mana data latihnya sudah memiliki nilai target atau label kelas[16]. Algoritma ini sering digunakan dalam pengenalan pola, pengolahan citra, dan sistem rekomendasi. Biasanya pencarian terdekat menggunakan rumus jarak Euclidean[17].

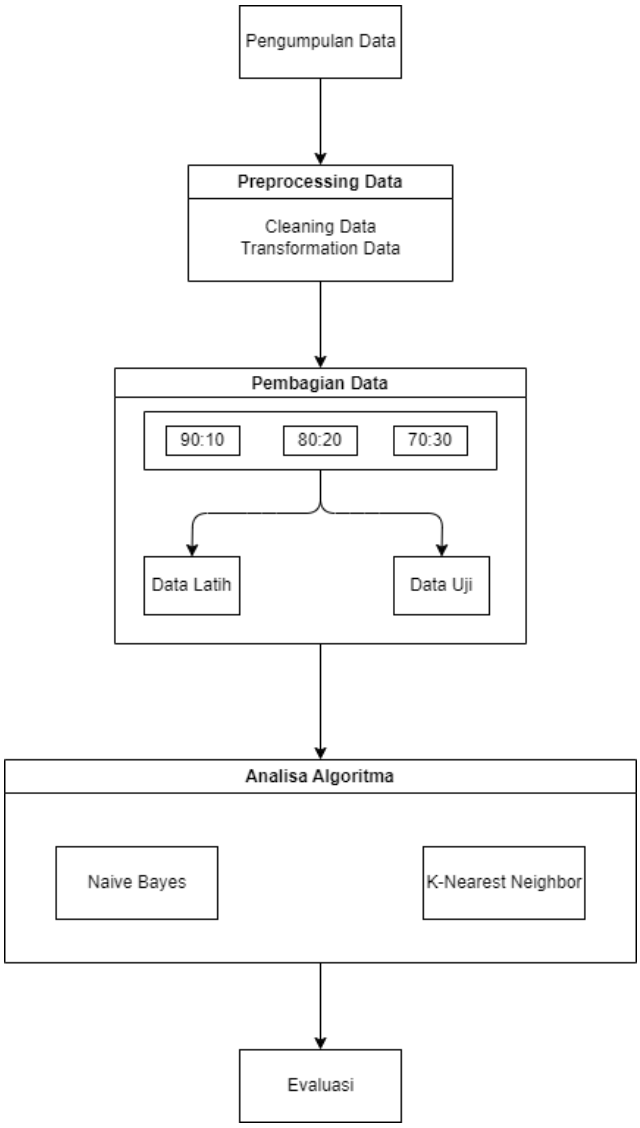
Berdasarkan penelitian terdahulu, hasil klasifikasi algoritma K-Nearest Neighbors dan Naive Bayes untuk penyakit Terhadap Resiko Diabetes Pada Ibu Hamil menunjukkan bahwa teknik K-Nearest Neighbors dengan K=25 memperoleh akurasi sebesar 74,48%, sedangkan metode Naïve Bayes memperoleh akurasi yang tinggi yaitu sebesar 75,78%[8]. Penelitian selanjutnya yaitu Perbandingan Algoritma Naïve Bayes dan K-Nearest Neighbors untuk Klasifikasi Metabolik Sindrom yang terdiri dari 505 data menunjukkan bahwa hasil yang paling efektif dengan parameter K=5 dan skema pembagian data 50:50 mendapatkan hasil akurasi mencapai 82%, sedangkan Naïve Bayes mendapatkan hasil akurasinya yaitu 79% yang mana K-Nearest Neighbors memiliki nilai tinggi dibanding Naïve Bayes[18]. Selain itu, pada penelitian yaitu membandingkan Metode Algoritma Naïve Bayes dan K-Nearest Neighbors Klasifikasi Penyakit Hati dengan data sebanyak 1025 data yang dimana K-Nearest Neighbors memiliki akurasi tertinggi dengan nilai 100% dan naïve bayes mendapatkan dengan nilai 85%[19]. Penulis menyajikan penelitian ini, yang membandingkan algoritma Naïve Bayes dan K-Nearest Neighbors untuk Klasifikasi Penyakit Gagal Jantung menggunakan total dengan jumlah 918 data, sesuai dengan latar belakang yang diberikan. Dalam peneloitian ini berfokus pada perbandingan dua algoritma yang sering digunakan dalam klasifikasi, yaitu Naïve dan K-Nearest Neighbors (KNN). Tujuan dari penelitian ini adalah untuk mengetahui cara memutuskan strategi mana yang memiliki kinerja terbaik dalam pengklasifikasian.

2. METODOLOGI PENELITIAN

Metodologi penelitian menggambarkan langkah-langkah yang diterapkan pada suatu penelitian untuk melaksanakan penelitian secara sistematis dan mengarahkan penelitian agar dilaksanakan sesuai tujuan yang telah ditetapkan. Langkah-langkah penelitian ini ditunjukkan pada Gambar 1 dibawah ini.

Copyright © 2025 The Author, Page 998
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Journal of Information System Research (JOSH)
Volume 6, No. 2, Januari 2025, pp 997−1005
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i2.6480



Gambar 1. Metodologi Penelitian

2.1 Pengumpulan Data

Data Collection adalah rangkaian cara dalam menghimpun informasi untuk pengembangan sistem atau penelitian[20]. Penelitian ini menggunakan data yang di input dari Kaggle dengan jumlah menjadi 918 data pasien gagal jantung. Data tersebut terdiri dari Umur, Sex, ChestPainType, Resting Bp, Cholesterol, Fasting BS, Resting ECG, MaxHR, Exercise Angina, Oldpeak, ST_Slope, dan HeartDisease.

2.2 Data Preprocessing

a. Data Cleaning
   Pada tahap ini, prosedur yang dilakukan antara lain membersihkan informasi duplikat (data cleaning) pada data pasien gagal jantung dan memverifikasi informasi yang tidak konsisten, seperti kesalahan usia, dan data kosong.
b. Data Transformation
   Pada tahapan ini dilakukan untuk mentransformasikan data huruf menjadi data angka. Seperti atribut sakit kepala bagian tengkuk, jika di data iya maka menjadi 1 dan jika tidak maka menjadi 0 dan atribut untuk kelas jika pasien terkena penyakit gagal jantung maka di ubah menjadi 1, jika bukan gagal jantung maka diubah menjadi 0.

2.3 Pembagian Data

Data Proses Pemisahan Data dilakukan untuk memisahkan data menjadi data pelatihan dan data pengujian. Pemisahan data digunakan untuk memastikan model dapat digeneralisasi secara efektif ke data yang sebelumnya tidak terlihat. Tiga konfigurasi dalam pemisahan data meliputi Rasio pembagian data meliputi 90% untuk data training dan 10% untuk testing, 80% untuk data training dan 20% untuk data testing, serta 70% untuk data training dan 30% untuk data testing.

Copyright © 2025 The Author, Page 999
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Journal of Information System Research (JOSH)
Volume 6, No. 2, Januari 2025, pp 997−1005
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i2.6480

2.4 Naïve Bayes

Dengan dugaan sementara yang kuat, pengklasifikasi Naive Bayes menggunakan teknik statistik dan probabilistik langsung berdasarkan teorema Bayes[12]. Mengacu pada variabel kelas sesuai nilai yang ditentukan, suatu metode yang menerapkan teorema Bayes dan membuat asumsi bahwa semua kualitas bersifat independen atau "tidak" bergantung satu sama lain[11].

Untuk mencari klasifikasi tertinggi, Naïve Bayes juga menggunakan cabang matematika yang disebut teori probabilitas. Ia melakukannya dengan memeriksa frekuensi setiap pengklasifikasi dalam data latih. Dalam metode Naïve Bayes, ketika dua peristiwa unik (seperti X dan C) terjadi, persamaan Naïve Bayes dirumuskan sebagai berikut.

$$P(C|X) = \frac{P(X|C).P(C)}{P(X)} \quad (1)$$

Pada rumus ini, P(C | X) erepresentasikan probabilitas posterior, yaitu peluang bahwa hipotesis C (kelas) benar berdasarkan data X (atribut). Selanjutnya, P(X | C) menggambarkan likelihood atau peluang data X muncul jika kelas C benar. Sementara itu, P(C) adalah probabilitas prior dari kelas C, dan P(X) merupakan probabilitas evidence atau peluang terjadinya data X.

2.5 K-Nearest Neighbor

Teknik klasifikasi objek yang disebut K-Nearest Neighbors (KNN) menggunakan data pelatihan yang dekat dengan item tersebut. Ruang multidimensi digunakan untuk data pelatihan proyek, dengan masing-masing dimensi mewakili data atribut. Berdasarkan klasifikasi data latih, ruang ini dibagi menjadi beberapa bagian[13]. Algoritma K-NN merupakan algoritma nearest neighbor yang dihitung dari nilai jarak antara data uji dengan data latih, dimulai dari nilai tetangga terkecil. Tujuan algoritma ini adalah mengkategorikan objek baru berdasarkan properti dan data latihnya. Nilai K pada K-Nearest Neighbor (K-NN), atau nilai K dari data yang paling dekat dengan data eksperimen, digunakan dalam prosedur ini. K-Nearest Neighbor dapat menghasilkan klasifikasi yang akurat ketika terdapat sejumlah besar data[14].

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

Dalam algoritma K-Nearest Neighbors (KNN), jarak antara data baru (data uji) dan data yang sudah ada (data pelatihan) dihitung menggunakan rumus tertentu, seperti Euclidean Distance. Pada rumus ini, $d(x,y)$ mewakili jarak antara titik data x dan y, x_i serta y_i adalah nilai atribut I pada data x dan y, sedangkan n adalah jumlah atribut dalam dataset.

2.6 Evaluasi

Evaluasi merupakan proses untuk menilai dan mengukur kinerja model yang dibangun menggunakan teknik data mining. Ini mencakup berbagai metode dan metrik yang digunakan untuk menentukan seberapa baik model tersebut dalam memprediksi atau mengklasifikasikan data baru[21]. Evaluasi ini bertujuan untuk membandingkan kinerja kedua algoritma menggunakan Confusion matrix berdasarkan beberapa metrik klasifikasi yang umum digunakan dalam penelitian medis. Confusion matrix memiliki yaitu Akurasi, Presisi dan Recall.

3. HASIL DAN PEMBAHASAN

3.1 Pengupulan Data

Data yang dipakai dalam penelitian ini berupa data yang berasal dari Kaggle, sebuah platform yang menyediakan berbagai dataset untuk keperluan analisis data dan pembelajaran mesin. Dataset ini diperoleh dari berbagai rumah sakit di berbagai negara yang telah mengumpulkan informasi medis terkait pasien dan menyatukannya menjadi satu dataset yang komprehensif oleh penulis asli data tersebut. Data ini berisi informasi penting mengenai gejala-gejala yang dialami oleh pasien yang mengalami gangguan kesehatan jantung, khususnya pasien dengan kondisi gagal jantung. Jumlah data yang dikumpulkan mencapai 918 catatan pasien, yang menyediakan wawasan berharga untuk analisis statistik dan pemodelan prediktif. Dataset ini mencakup 12 variabel penting, yaitu Umur, Sex, ChestPainType, Resting Bp, Cholesterol, Fasting BS, Resting ECG, MaxHR, Exercise Angina, Oldpeak, ST_Slope, dan HeartDisease.

Tabel 1. Data Pasien Gagal Jantung

No	Age	Sex	ChestPain Type	Resting BP	Cholesterol	Fasting BS	Resting ECG	Max HR	Exercise Angina	Oldpeak	STSlope	Heart Disease
1	40	M	ATA	140	289	0	Normal	172	N	0	Up	0
2	49	F	NAP	160	180	0	Normal	156	N	1	Flat	1
3	37	M	ATA	130	283	0	ST	98	N	0	Up	0
4	48	F	ASY	138	214	0	Normal	108	Y	1.5	Flat	1

Copyright © 2025 The Author, Page 1000
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Journal of Information System Research (JOSH)
Volume 6, No. 2, Januari 2025, pp 997−1005
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i2.6480

No	Age	Sex	ChestPain Type	Resting BP	Cholesterol	Fasting BS	Resting ECG	Max HR	Exercise Angina	Oldpeak	STSlope	Heart Disease
5	54	M	NAP	150	195	0	Normal	122	N	0	Up	0
6	39	M	NAP	120	339	0	Normal	170	N	0	Up	0
…	…	…	…	…	…	…	…	…	…	…	…	…
918	38	M	NAP	138	175	0	Normal	173	N	0	Up	0

Tabel 1 adalah dataset gagal jantung yang terdiri dari atribut Age, Sex, Chest Pain Type, Resting Bp, Cholesterol, Fasting BS, Resting ECG, Max HR, Exercise Angina, Oldpeak, ST_Slope, dan Heart Disease. Berikut adalah keterangan dari atribut dataset yang digunakan:

a. Age : umur pasien
b. Sex : jenis kelamin pasien (M = Pria, F = perempuan)
c. Chest Pain Type : jenis nyeri dada (TA = Typical Angina, ATA = Atypical Angina, NAP = Non-Anginal Pain, ASY = Asymptomatic)
d. Resting Bp: tekanan darah dalam kondisi istirahat [mm Hg]
e. Cholesterol : kadar kolesterol (mm/dl)
f. Fasting BS : kadar gula darah saat puasa (jika gula > 120 maka = 1, jika < 120 maka = 0
g. Resting ECG : kondisi ECG pasien saat isitrahat
h. Max HR : maksimum detak jantung
i. Exercise Angina : nyeri dada saat berolahraga (Y = Yes, N = No.)
j. Oldpeak : penurunan ST saat olahraga ST (Nilai numerik diukur dalam depresi)
k. ST Slope : latihan maksimum kemiringan segmen ST latihan maksimum (Up = upsloping, Flat = flat, Down = downsloping)
l. HeartDisease : kelas output (1 = heart disease, 0 = Normal)

3.2 Data Preprocessing

Pada tahap ini, dilakukan preprocessing data untuk memastikan kualitas data yang digunakan dalam analisis. Proses ini mencakup beberapa langkah penting, seperti memeriksa kesalahan dalam penginputan data, membersihkan data dari nilai yang tidak lengkap, tidak valid, atau atribut dengan baris kosong yang dapat mengganggu analisis. Salah satu langkah penting dalam preprocessing adalah data transformation. Data transformation dilakukan untuk mengoptimalkan struktur data sehingga lebih efektif dan efisien saat dianalisis. Proses ini melibatkan pengubahan format atau representasi data agar memudahkan dalam mengidentifikasi pola atau hubungan yang mungkin tersembunyi di dalam dataset. Misalnya, data yang semula dalam bentuk kategorikal dapat diubah menjadi data numerik menggunakan metode encoding, atau skala data dapat dinormalisasi agar lebih seragam. Selain itu, data transformation juga mencakup teknik agregasi, normalisasi, atau standarisasi untuk memastikan bahwa data yang digunakan relevan dan mendukung tujuan analisis. Langkah ini sangat penting untuk memastikan hasil analisis yang lebih akurat dan dapat diandalkan.

Tabel 2. Sebelum Dataset melakukan transformation

No	Age	Sex	ChestPain Type	Resting BP	Cholesterol	Fasting BS	Resting ECG	Max HR	Exercise Angina	Oldpeak	STSlope	Heart Disease
1	40	M	ATA	140	289	0	Normal	172	N	0	Up	0
2	49	F	NAP	160	180	0	Normal	156	N	1	Flat	1
3	37	M	ATA	130	283	0	ST	98	N	0	Up	0
4	48	F	ASY	138	214	0	Normal	108	Y	1.5	Flat	1
5	54	M	NAP	150	195	0	Normal	122	N	0	Up	0
6	39	M	NAP	120	339	0	Normal	170	N	0	Up	0
…	…	…	…	…	…	…	…	…	…	…	…	…
918	38	M	NAP	138	175	0	Normal	173	N	0	Up	0

Selanjutnya yaitu dengan melakukan transformasi data agar ketika melakukan pengklasifikasi algoritma lebih mudah. Berikut adalah dataset setelah dilakukan transformasi:

Tabel 3. Setelah Dataset melakukan Transformation

No	Age	Sex	ChestPain Type	Resting BP	Cholesterol	Fasting BS	Resting ECG	Max HR	Exercise Angina	Oldpeak	ST_Slope	HeartDisease
1	40	1	1	140	289	0	1	172	0	0	2	0
2	49	0	2	160	180	0	1	156	0	1	1	1
3	37	1	1	130	283	0	2	98	0	0	2	0
4	48	0	0	138	214	0	1	108	1	1.5	1	1
5	54	1	2	150	195	0	1	122	0	0	2	0
6	39	1	2	120	339	0	1	170	0	0	2	0

Copyright © 2025 The Author, Page 1001
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Journal of Information System Research (JOSH)
Volume 6, No. 2, Januari 2025, pp 997−1005
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i2.6480

No	Age	Sex	ChestPain Type	Resting BP	Cholesterol	Fasting BS	Resting ECG	Max HR	Exercise Angina	Oldpeak	ST_Slope	HeartDisease
918	38	1	2	138	175	0	1	173	0	0	2	0

Pada Tabel 3, atribut pada pasien gagal jantung di-transformasi dari teks kategori menjadi format numerik. Transformasi ini dilakukan dengan tujuan utama untuk mempermudah proses pengklasifikasian menggunakan algoritma Naïve Bayes dan K-Nearest Neighbors. Kedua algoritma ini bekerja lebih optimal dengan data numerik, sehingga pengubahan dari kategori teks ke numerik menjadi langkah krusial dalam preprocessing data. Atribut yang mengalami transformasi mencakup Sex, Chest Pain Type, Resting ECG, Exercise Angina, dan ST Slope. Berikut adalah detail transformasi atribut yang dilakukan:

a. Sex: Kategori jenis kelamin diubah menjadi nilai numerik, yaitu M (Male) diberikan nilai 1 dan F (Female) diberikan nilai 0.
b. Chest Pain Type: Tipe nyeri dada yang sebelumnya berupa kategori teks diubah menjadi nilai numerik dengan rincian: TA (Typical Angina) = 3, ATA (Atypical Angina) = 1, NAP (Non-Anginal Pain) = 2, dan ASY (Asymptomatic) = 0.
c. Resting ECG: Hasil elektrokardiogram saat istirahat diubah menjadi: LVH (Left Ventricular Hypertrophy) = 0, Normal = 1, dan ST = 2.
d. ExerciseAngina: Kondisi angina saat berolahraga diubah menjadi Y (Yes) = 1 dan N (No) = 0.
e. ST_Slope: Kemiringan segmen ST diubah menjadi nilai numerik: Up (Upsloping) = 2, Flat = 1, dan Down (Downsloping) = 0.

Transformasi ini tidak hanya memudahkan dalam proses analisis, tetapi juga meningkatkan efisiensi algoritma karena atribut yang lebih seragam meminimalkan kesalahan interpretasi dalam perhitungan algoritma. Dengan data numerik yang telah disiapkan, algoritma Naïve Bayes dan K-Nearest Neighbors dapat lebih mudah mengenali pola yang relevan, sehingga hasil klasifikasi menjadi lebih akurat dan konsisten. Transformasi ini adalah bagian integral dalam mempersiapkan data untuk tahap analisis berikutnya.

3.3 Pembagian Data

Pada penelitian ini, penulis melakukan pembagian data (data splitting) ke dalam tiga skenario berbeda untuk mengevaluasi dan mencari nilai akurasi terbaik dalam proses klasifikasi. Ketiga pembagian data tersebut adalah 90:10, 80:20, dan 70:30. Setiap skenario dilakukan untuk memastikan algoritma dapat dilatih secara optimal menggunakan data yang tersedia, serta diuji dengan data yang cukup untuk menghasilkan model yang akurat dan dapat diandalkan.

Pada skenario pertama, pembagian data dilakukan dengan rasio 90:10, di mana 90% data digunakan untuk training dan 10% untuk testing. Skenario kedua menggunakan rasio 80:20, di mana 80% data dialokasikan untuk training dan 20% untuk testing. Skenario terakhir adalah 70:30, dengan 70% data digunakan untuk training dan 30% untuk testing. Langkah ini memungkinkan model memiliki lebih banyak data untuk dipelajari, sehingga meningkatkan kemungkinan mendeteksi pola yang lebih kompleks. Dalam skenario ini, nilai K yang digunakan adalah 9, yang mewakili jumlah tetangga terdekat dalam algoritma K-Nearest Neighbors (KNN).

Hasil dari ketiga skenario ini kemudian dibandingkan untuk menentukan rasio pembagian data terbaik yang memberikan akurasi tertinggi. Dengan pendekatan ini, penelitian dapat mengidentifikasi konfigurasi yang paling optimal untuk mencapai performa model yang maksimal dalam proses klasifikasi.

3.4 Naïve Bayes

Pada penelitian ini, algoritma Naïve Bayes diterapkan untuk membandingkan data dengan tiga rasio berbeda, yaitu 90:10, 80:20, dan 70:30. Perbandingan ini bertujuan untuk menentukan pembagian data yang paling efektif dalam menghasilkan akurasi prediksi yang optimal. Pada pembagian 90:10, 90% data digunakan untuk melatih model (training) dan 10% data digunakan untuk menguji model (testing). Pembagian ini memberikan lebih banyak data untuk proses pembelajaran, sehingga diharapkan model dapat menangkap pola dengan lebih baik.

Tabel 4. Confusion Matrix Naive Bayes

	Parameter		
	Akurasi	**Recall**	**Precision**
90:10	88%	89%	91%
80:20	90%	90%	93%
70:30	89%	91%	90%

Hasil pada Tabel 4 menunjukkan bahwa dari tiga perbandingan data yang dilakukan, rasio 80:20 memberikan nilai akurasi tertinggi, yaitu sebesar 90%. Dapat disimpulkan bahwa pemisahan data menggunakan proporsi 80% untuk training dan 20% untuk testing mampu memberikan keseimbangan optimal antara data pelatihan yang cukup banyak dan data pengujian yang representatif. Dibandingkan dengan rasio 90:10, yang

This Journal is licensed under a Creative Commons Attribution 4.0 International License

Journal of Information System Research (JOSH)
Volume 6, No. 2, Januari 2025, pp 997−1005
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i2.6480

mendapatkan nilai akurasi sebesar 88%, terlihat bahwa meskipun rasio ini memberikan lebih banyak data untuk pelatihan, pengujian model menjadi kurang optimal karena jumlah data testing yang lebih kecil.

Sementara itu, pada rasio 70:30, nilai akurasi yang diperoleh adalah 89%, sedikit lebih rendah dibandingkan dengan rasio 80:20. Rasio ini memiliki data pengujian yang lebih besar, namun jumlah data pelatihan yang lebih kecil tampaknya memengaruhi kemampuan model untuk mempelajari pola secara mendalam. Dengan demikian, dapat disimpulkan bahwa rasio 80:20 memberikan keseimbangan terbaik, sehingga mampu menghasilkan akurasi tertinggi dalam pengujian model.

3.5 K-Nearest Neighbors

Dalam penelitian ini, algoritma K-Nearest Nighbors digunakan untuk melakukan tiga perbandingan data, yaitu dengan rasio 90:10, 80:20, dan 70:30. Perbandingan ini bertujuan untuk menentukan pembagian data yang paling efektif dalam menghasilkan akurasi prediksi yang optimal. Pada pembagian 90:10, 90% data digunakan untuk melatih model (training) dan 10% data digunakan untuk menguji model (testing). Pembagian ini memberikan lebih banyak data untuk proses pembelajaran, sehingga diharapkan model dapat menangkap pola dengan lebih baik.

Tabel 5. Confusion Matrix K-Nearest Nighbors

	Parameter		
	Akurasi	**Recall**	**Precision**
90:10	73%	75%	82%
80:20	76%	78%	84%
70:30	71%	71%	76%

Hasil pada Tabel 5 menunjukkan bahwa dari tiga perbandingan data yang dilakukan, rasio 80:20 menghasilkan nilai akurasi tertinggi, yaitu sebesar 76%. Rasio ini menunjukkan bahwa pembagian data dengan proporsi 80% untuk training dan 20% untuk testing memberikan hasil yang lebih optimal dibandingkan dua rasio lainnya. Rasio 90:10, yang menghasilkan akurasi sebesar 73%, dan pada rasio 70:30 mendapatkan nilai akurasi sebesar 71%.

3.6 Evaluasi

Pada penelitian ini, metode Naive Bayes dan K-Nearest Neighbors dengan menggunakan 3 rasio data yaitu 90:10, 80:20 dan 70:30. Hasil akurasi Naive Bayes dan K-Nearest Neighbors dapat dilihat Tabel 6.

Tabel 6. Perbandingan Naive Bayes dan K-Nearest Neighbors

Algoritma		**Spliting Data**		
		90:10	**80:20**	**70:30**
Naïve Bayes	Akurasi	88%	90%	89%
	Recall	89%	90%	91%
	Precision	91%	93%	90%
K-Nearest Neighbors	Akurasi	73%	76%	71%
	Recall	75%	78%	71%
	Precision	82%	84%	76%

Hasil pada Tabel 6 menunjukkan bahwa pembagian data menggunakan rasio 80:20 memberikan hasil akurasi tertinggi untuk kedua algoritma, baik Naïve Bayes maupun K-Nearest Neighbors (KNN). Pada rasio ini, algoritma Naïve Bayes mencapai akurasi sebesar 90%, sementara algoritma KNN memperoleh nilai akurasi 76%. Hal ini membuktikan bahwa algoritma Naïve Bayes mampu mengklasifikasikan data dengan lebih efektif dibandingkan dengan KNN, terutama dalam menangkap pola yang relevan dari dataset yang digunakan.

Keunggulan Naïve Bayes dalam penelitian ini dapat dikaitkan dengan asumsi independensinya, yang memungkinkan algoritma ini tetap bekerja dengan baik meskipun terdapat banyak variabel dalam dataset. Sementara itu, performa KNN yang sedikit lebih rendah dapat disebabkan oleh sensitivitasnya terhadap distribusi data dan pemilihan nilai parameter K. Dengan demikian, dapat disimpulkan bahwa algoritma Naïve Bayes memberikan performa yang lebih unggul dibandingkan KNN dalam penelitian ini. Hasil ini juga menegaskan pentingnya pemilihan algoritma yang sesuai dengan karakteristik dataset untuk mendapatkan hasil yang optimal dalam proses klasifikasi.

4. KESIMPULAN

Dalam penelitian ini, digunakan dataset yang diambil dari platform Kaggle, yang mencakup data pasien gagal jantung dari berbagai rumah sakit di seluruh dunia. Dataset tersebut terdiri dari total 918 data pasien, masing-masing dilengkapi dengan 12 variabel yang relevan untuk menganalisis kondisi kesehatan pasien. Dengan menerapkan algoritma Naïve Bayes dan K-Nearest Neighbors (KNN), dilakukan evaluasi akurasi melalui tiga

Copyright © 2025 The Author, Page 1003
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Journal of Information System Research (JOSH)
Volume 6, No. 2, Januari 2025, pp 997−1005
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i2.6480

skema pembagian data, yaitu 90:10, 80:20, dan 70:30. Dari hasil analisis, skema pembagian data 80:20 menghasilkan performa terbaik, di mana algoritma Naïve Bayes mencapai akurasi sebesar 90%, sementara K-Nearest Neighbors memperoleh akurasi sebesar 76%. Temuan ini menunjukkan bahwa algoritma Naïve Bayes lebih unggul dalam mengklasifikasikan data pasien gagal jantung dibandingkan KNN, terutama dalam skema pembagian data yang seimbang antara pelatihan dan pengujian. Keunggulan Naïve Bayes dapat dikaitkan dengan pendekatannya yang sederhana namun efektif, yang mengasumsikan independensi antar variabel, sehingga dapat menangkap pola data dengan baik meskipun terdapat kompleksitas dalam atribut dataset.

REFERENCES

[1] F. Novaldy and A. Herliana, "Penerapan Pso Pada Naïve Bayes untuk Prediksi Harapan Hidup Pasien Gagal Jantung," Jurnal Responsif, vol. 3, no. 1, pp. 37–43, 2021, [Online]. Available: https://doi.org/10.51977/jti.v3i1.396
[2] D. Purnama Sari, M. Mustain, and M. Maksum, "Gambaran Pengelolaan Hipervolemia pada Gagal Jantung Kongestif di Rumah Sakit," Jurnal Keperawatan Berbudaya Sehat, vol. 1, no. 1, pp. 9–15, Jan. 2023, doi: 10.35473/jkbs.v1i1.2155.
[3] J. Triani, Y. Pratama, and E. Yanti, "Komparasi dalam Prediksi Gagal Jantung dengan Menggunakan Metode C4.5 dan Naïve Bayes," JAKAKOM, 2023. [Online]. Available: http://ejournal.unama.ac.id/index.php/jakakom
[4] G. Evelyn, R. Feradwiyanti, and R. Rismayanti, "Faktor – Faktor yang Mempengaruhi Kualitas Hidup Pasien Gagal Jantung Kronik Dirsud Karawang," Jurnal Inovasi Penelitian, vol. 2, no. 2, pp. 775–784, Jul. 2021, doi: 10.47492/JIP.V2I2.2803.
[5] A. Khoeruddin, F. Andriansyah Sudrajat, G. Purnama, I. Kuwangid, K. Kurnia, and R. Firmansyah, "Optimasi Fitur Seleksi Random Forest Menggunakan GA Dalam Klasifikasi Data Penyakit Gagal Jantung," Jurnal Penelitian Teknologi Informasi dan Sains, vol. 1, no. 2, pp. 01–09, Jun. 2023, doi: https://doi.org/10.54066/jptis.v1i2.323
[6] H. Nursita and A. Pratiwi, "Peningkatan Kualitas Hidup pada Pasien Gagal Jantung: A Narrative Review Article (Improved Quality of Life in Heart Failure Patients: A Narrative Review Article)," Jurnal Berita Ilmu Keperawatan, vol. 13, no. 1, pp. 10–21, Jan. 2020, doi: 10.23917/bik.v13i1.11916.
[7] P. Aisyiyah and R. Devi, "Klasifikasi Penyakit Gagal Ginjal Kronis dengan Metode Knn (Studi Kasus RS di Kab Gresik)," JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika), vol. 9, no. 3, pp. 1739–1748, Sep. 2024, doi: 10.29100/jipi.v9i3.6226.
[8] J. Homepage, B. Delvika, S. Nurhidayarnis, P. D. Rinada, N. Abror, and A. Hidayat, "Perbandingan Klasifikasi Antara Naive Bayes dan K-Nearest Neighbor Terhadap Resiko Diabetes Pada Ibu Hamil," MALCOM: Indonesian Journal of Machine Learning and Computer Science, vol. 2, pp. 68–75, 2022, doi: 10.23917/bik.v13i1.11916.
[9] F. Kana, M. Ramadhan, and R. Mahyuni, "Implementasi Data Mining Menganalisa Pola Penjualan Rempah-Rempah Menggunakan Metode Fp-Growth," Jurnal Sistem Informasi Triguna Dharma (JURSI TGD), vol. 1, no. 4, p. 557, Jul. 2022, doi: 10.53513/jursi.v1i4.5586.
[10] M. Y. Putra and D. I. Putri, "Pemanfaatan Algoritma Naïve Bayes dan K-Nearest Neighbor Untuk Klasifikasi Jurusan Siswa Kelas XI," Jurnal Tekno Kompak, vol. 16, no. 2, p. 176, Aug. 2022, doi: 10.33365/jtk.v16i2.2002.
[11] S. Bahri, D. Marisa Midyanti, R. Hidayati, J. Sistem Komputer, and F. Mipa, "Perbandingan Algoritma Naive Bayes dan C4.5 Untuk Klasifikasi Penyakit Anak," Yogyakarta, Aug. 2018. Accessed: Dec. 31, 2024. [Online]. Available: https://journal.uii.ac.id/Snati/article/view/11152
[12] K. Abdul Khalim, U. Hayati, and A. Bahtiar, "Perbandingan Prediksi Penyakit Hipertensi Menggunakan Metode Random Forest dan Naïve Bayes," JATI (Jurnal Mahasiswa Teknik Informatika), vol. 7, no. 1, pp. 498–504, Mar. 2023, doi: 10.36040/jati.v7i1.6376.
[13] I. Lishania, R. Goejantoro, and Y. N. Nasution, "Perbandingan Klasifikasi Metode Naive Bayes dan Metode Decision Tree Algoritma (J48) pada Pasien Penderita Penyakit Stroke di RSUD Abdul Wahab Sjahranie Samarinda Hospital," Jurnal EKSPONENSIAL, vol. 10, no. 2, Nov 2019. Tersedia pada: https://jurnal.fmipa.unmul.ac.id/index.php/exponensial/article/view/571
[14] A. R. Oktavyani et al., "Sistem Informasi, dan Teknik Informatika Perbandingan Metode Naive Bayes, K-NN dan Decision Tree Terhadap Dataset Healthcare Stroke," SNESTIK Seminar Nasional Teknik Elektro, Sistem Informasi, dan Teknik Informatika, pp. 276–281, 2023, doi: 10.31284/p.snestik.2023.4067.
[15] I. L. F. Amien, W. Astuti, and K. M. Lhaksamana, "Perbandingan Metode Naïve Bayes dan KNN (K-Nearest Neighbor) dalam Klasifikasi Penyakit Diabetes," eProceedings of Engineering, vol. 10, no. 2, May 2023, Accessed: Dec. 11, 2024. [Online]. Available: https://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/view/20039
[16] M. Bansal, A. Goyal, and A. Choudhary, "A comparative analysis of K-Nearest Neighbor, Genetic, Support Vector Machine, Decision Tree, and Long Short Term Memory algorithms in machine learning," Decision Analytics Journal, vol. 3, p. 100071, Jun. 2022, doi: 10.1016/j.dajour.2022.100071.
[17] D. Ulfatul, M. Rachmad, H. Oktavianto, and M. Rahman, "Perbandingan Metode K-Nearest Neighbor Dan Gaussian Naive Bayes Untuk Klasifikasi Penyakit Stroke Comparison Of K-Nearest Neighbor And Gaussian Naive Bayes Methods For Stroke Disease Classification," Jurnal Smart Teknologi, vol. 3, no 4, pp. 405-412, Mei. 2022. [Online]. Available: http://jurnal.unmuhjember.ac.id/index.php/JST/article/view/7601
[18] F. Sholekhah, A. D. Putri, R. Rahmaddeni, and L. Efrizoni, "Perbandingan Algoritma Naïve Bayes dan K-Nearest Neighbors untuk Klasifikasi Metabolik Sindrom," MALCOM: Indonesian Journal of Machine Learning and Computer Science, vol. 4, no. 2, pp. 507–514, Feb. 2024, doi: 10.57152/malcom.v4i2.1249.
[19] A. Desiani, "Perbandingan Implementasi Algoritma Naïve Bayes dan K-Nearest Neighbor Pada Klasifikasi Penyakit Hati," SIMKOM, vol. 7, no. 2, pp. 104–110, Jul. 2022, doi: 10.51717/simkom.v7i2.96.
[20] Dewi Nasien et al., "Perbandingan Implementasi Machine Learning Menggunakan Metode KNN, Naive Bayes, dan Logistik Regression Untuk Mengklasifikasi Penyakit Diabetes," JEKIN - Jurnal Teknik Informatika, vol. 4, no. 1, pp. 10–17, Feb. 2024, doi: 10.58794/jekin.v4i1.640.

This Journal is licensed under a Creative Commons Attribution 4.0 International License

Journal of Information System Research (JOSH)
Volume 6, No. 2, Januari 2025, pp 997−1005
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i2.6480

[21] A. Damuri, U. Riyanto, H. Rusdianto, and M. Aminudin, "Implementasi Data Mining dengan Algoritma Naïve Bayes Untuk Klasifikasi Kelayakan Penerima Bantuan Sembako," Jurnal Riset Komputer), vol. 8, no. 6, pp. 2407–389, 2021, doi: 10.30865/jurikom.v8i6.3655.

Copyright © 2025 The Author, Page 1005
This Journal is licensed under a Creative Commons Attribution 4.0 International License