Fakultas Teknik,
Universitas Wijaya Kusuma Surabaya
ISSN 3090-1154 (Media Online)
Volume 2, Tahun 2025

Perbandingan Algoritma Naïve Bayes dan K-Nearest Neighbor Untuk Mengklasifikasikan Status Kesehatan

Nazhifatul Muthohharoh*[1], Lukman Fakih Lidimilah[2], Ahmad Homaidi[3].
[1,2,3] Universitas Ibrahimy
Email: [1]nsfh2502@gmail.com, [2]lukylukman7@gmail.com , [3]Ahmadhomaidi@ibrahimy.ac.id

Abstrak

Pemanfaatan algoritma klasifikasi dalam bidang kesehatan dapat membantu mengidentifikasi status kesehatan individu secara lebih akurat dan efisien. Penelitian ini bertujuan untuk membandingkan kinerja algoritma Naive Bayes dan K-Nearest Neighbor (K-NN) dalam mengklasifikasikan status kesehatan berdasarkan beberapa parameter menurut kebiasaan gaya hidup seperti kebiasaan merokok, aktifitas bekerja, aktifitas begadang, aktifitas olahraga, pola makan teratur dan penyakit bawaan. Data diolah menggunakan Google Colaboratory dengan pembagian 80% data latih dan 20% data uji. Evaluasi dilakukan menggunakan metrik akurasi, presisi, recall, dan F1-score. Hasil menunjukkan bahwa algoritma, Naïve Bayes dan K-Nearest Neighbor, keduanya menunjukkan akurasi yang sama, yaitu 0.92, dimana keduanya dapat digunakan secara efektif, tetapi, jika data memeiliki distribusi normal dan kontinu, naïve bayes bisa menjadi pilihan yang efisien. Temuan ini mengindikasikan bahwa pemilihan algoritma sebaiknya disesuaikan dengan kebutuhan sistem, apakah mengutamakan akurasi atau efisiensi. Penelitian ini diharapkan dapat menjadi referensi dalam pengembangan sistem klasifikasi kesehatan berbasis machine learning yang efektif dan adaptif.

Kata kunci: Klasifikasi, Naïve Bayes, K-NN, Data Mining, Pembelajaran mesin

Abstract (10pt, cetak tebal, dan di tengah)

The utilization of classification algorithms in the health sector can help identify individual health status more accurately and efficiently. This study aims to compare the performance of Naive Bayes and K-Nearest Neighbor (K-NN) algorithms in classifying health status based on several parameters according to lifestyle habits such as smoking habits, work activities, late night activities, sports activities, regular diet and congenital diseases. Data is processed using Google Collaboratory with a division of 80% training data and 20% test data. Evaluation is done using accuracy, precision, recall, and F1-score metrics. Results showed that the algorithms, Naïve Bayes and K-Nearest Neighbor, both showed similar accuracy of 0.92, which can be used effectively, however, if the data has a normal and continuous distribution, Naïve Bayes can be an efficient choice. This finding indicates that algorithm selection should be tailored to the needs of the system, whether it prioritizes accuracy or efficiency. This research is expected to be a reference in the development of an effective and adaptive machine learning-based health classification system.

Keywords*: Classification, Naïve Bayes, K-NN, Data Mining, Machine Learning*

I. PENDAHULUAN

Gaya hidup modern yang semakin tidak sehat telah berkontribusi pada meningkatnya berbagai permasalahan kesehatan, seperti obesitas, tekanan darah tinggi, dan diabetes. Pola makan tidak seimbang, kurangnya aktivitas fisik, serta stres yang berkepanjangan menjadi faktor utama pemicunya. Oleh karena itu, dibutuhkan suatu sistem prediksi yang dapat mendeteksi potensi risiko kesehatan berdasarkan pola hidup individu. Dengan adanya sistem prediksi ini, intervensi dini dan upaya pencegahan dapat dilakukan secara lebih tepat sasaran. Integrasi teknologi data mining dalam sistem prediksi tersebut memungkinkan analisis data kesehatan secara menyeluruh, sehingga menghasilkan rekomendasi yang akurat dan bermanfaat dalam mendukung pengambilan keputusan di bidang medis maupun kebijakan kesehatan masyarakat [1].

Karena kemampuannya dalam menangani data numerik dan kategorikal secara efisien, Naïve Bayes dan K-Nearest Neighbor (K-NN) yang merupakan dua algoritma klasifikasi yang banyak digunakan [2], [3].

Prinsip probabilistic dengan berdasar pada teorema bayes digunakan pada algoritma naïve bayes, dan untuk K-NN bekerja dengan menghitung jarak ketetanggaan antar data untuk menentukan kelas.

Beberapa penelitian sebelumnya telah mengimplementasikan kedua algoritma tersebut dalam bidang medis, seperti klasifikasi status pertumbuhan anak stunting[4], diabetes[5], dan metabolic sindrom [6]. Namun, perbandingan performa keduanya dalam konteks klasifikasi status kesehatan secara umum masih belum banyak dibahas secara komprehensif.

Tujuan penelitian ini untuk membandingkan kinerja algoritma Naïve Bayes dan K-NN dalam mengklasifikasikan status kesehatan berdasarkan variable tertentu. Identifikasi masalah yang diangkat adalah: (1) Seberapa besar tingkat akurasi diantara kedua algoritma, dan (2) Seberapa efisien pemrosesan data yang dimiliki masing-masing algoritma. Metode yang digunakan mencakup proses preprocessing data, pembagian data menjadi latih dan uji, penerapan kedua algoritma menggunakan Google Colaboratory, serta evaluasi performa dengan metrik akurasi, presisi, recall, dan F1-score. Data yang digunakan diperoleh dari repositori *Lifestyle Habits* di platform Kaggle dengan 387 entri dengan Sembilan atribut. Proses pengolahan dan pelatihan model menggunakan RapidMiner untuk visualisasi dan juga menggunakan *Python* pada *Google Colab*.

Penelitian ini diharapkan dapat memberikan kontribusi dalam pengembangan sistem klasifikasi kesehatan berbasis machine learning yang optimal, serta menjadi acuan dalam pemilihan algoritma yang sesuai dengan kebutuhan sistem. Berbeda dengan penelitian sebelumnya, penelitian ini tidak hanya menyoroti akurasi, tetapi juga mempertimbangkan efisiensi komputasi sebagai faktor penting dalam implementasi nyata.

II. METODE PENELITIAN

Penelitian ini, menggunakan pendekatan kuantitatif dengan metode eksperimen komparatif untuk membandingkan performa algoritma klasifikasi, yaitu Naïve bayes dan K-Nearest Neighbor (K-NN), dalam mengklasifikasikan status Kesehatan. Desain penelitian ini bertujuan untuk menganilisis perbedaan performa dalam melakukan klasifikasi berdasarkan parameter data yang sudah ditentukan. Dimana kedua algoritma ini diuji dan dibandingkan kinerjanya dalam mengklasifikasikan data status Kesehatan sebagai "sehat" atau "tidak sehat". Algoritma naïve bayes memiliki karakteristik berbasis probabilistic dan KNN berbasis kedekatan jarak, kedua algoritma ini dipilih karena masing-masing memiliki karakteristik berbeda dalam pendekatan klasifikasi.

Metode algoritma Naïve Bayes

Naive Bayes merupakan suatu algoritma yang dapat mengklasifikasikan suatu variable tertentu dengan menggunakan metode probabilitas dan statistic. Naive bayes menggunkan sebuah ilmu cabang matematika yang dikenal juga dengan teori probabilitas untuk mencari peluang terbesar dari kemungkinan klasifikasi, dengan cara melihat frekuensi tiap klasifikasi pada data training.

$$P(Ci\,|X) = \frac{P(X\,|Ci)\,P\,(Ci)}{P(X)}(1) \qquad [1]$$

Keterangan:
X: kriteria suatu kasus berdasarkan masukan
Ci: Kelas solusi pola ke-i, dimana i adalah jumlah label kelas
P(Ci|X): Probabilitas label kelas Ci dengan kriteria masukan X
P(X|Ci): Probalitas kriteria masukan X dengan label kelas Ci
P(Ci): Probabilitas label kelas Ci

Metode algoritma K-Nearest Neighbor

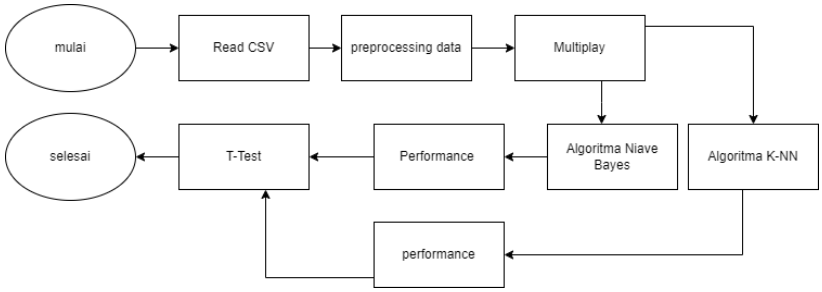
K-Nearest Neighbors (knn) merupakan algoritma yang mengklasifikasikan data berdasarkan data pembelajaran (training data set) yang diambil dari k-tetangga terdekat (nearest neighbours). dimana k adalah banyaknya tetangga terdekat. Metode K-Nearest Neighbors melakukan 14 klasifikasi dengan memproyeksikan data latih ke dalam ruang multidimensi. Area ini dibagi menjadi beberapa bagian yang mewakili data dasar pelatihan. Semua data pelatihan direpresentasikan sebagai titik c dalam ruang multidimensi.

K-Nearest Neighbors (KNN) merupakan algoritma klasifikasi yang menggunakan himpunan nilai K dari data terdekat (tetangganya) sebagai acuan untuk menentukan kelas data baru. KNN mengklasifikasikan data berdasarkan kemiripan atau kedekatannya dengan data lain. Algoritma KNN ini adalah pembelajaran yang malas. Artinya, tidak menggunakan titik data pelatihan untuk membangun model. Dengan kata lain algoritma KNN mempunyai fase pelatihan yang sangat minim. Tujuan dari algoritma ini adalah untuk mengklasifikasikan objek baru berdasarkan atribut dan sampel pada data pelatihan[7].

$$(A, B) = \sqrt{\sum_{i=1}^{n}(A_i - B_i)^2} \qquad [2]$$

Setelah menghitung jarak Euclidean langkah selanjutnya adalah menentukan K- Neighborsnya dengan cara mengurutkan dari nilai yang kecil sampai yang terbesar. Dari K neighbors terdekat tentukan label berdasarkan mayoritas dari K tetangga terdekat untuk mengevaluasi model dari algoritma K-NN tersebut, setelah mengevaluasi mode langkah selanjutnya adalah menentukan kelas dari dataset, Dari seluruh perhitungan tersebut terbagi kepada perhitungan dengan class label 'Iya' dan 'Tidak', kemudian hasil yang terbesar dari perbandingan ketiga label tersebut merupakan hasil dari prediksi algoritma K-Nearest Neighbors tersebut.

Penelitian ini dilakukan melalui beberapa tahapan utama untuk membandingkan performa algoritma naïve bayes dan K-NN dalam mengklasifikasikan status Kesehatan. Setiap tahap dirancang secara sistematis agar proses pengujian dapat berjalan dengan optimal. Adapun alur proses penelitian ditunjukkan pada gambar 1 berikut:



Gambar 1. Flowchart Tahapan Penelitian

Flowchart pada gambar 1 menunjukkan alur kerja penelitian dimulai dari *read CSV,* yaitu membaca data dalam format CSV *(Comma Separated Values)*. Dilanjutkan dengan tahap preprocessing dimana tahap tersebut meliputi membersihkan dan mempersiapkan data. Selanjutnya pada multiplay data dibagi untuk kebutuhan pelatihan dan pengujian model. Setelah itu dilakukan proses klasifikasi menggunakan dua algoritma, yaitu naïve bayes dan K-NN, kinerja masing-masing algoritma diukur melalui metrik performa meliputi akurasi, presisi, recall, dan F1-score. Hasil evaluasi kemudian dianalisis secara statistic menggunakan T-Test utnuk mengetahui seberapa signifikan perbedaan performa kedua algoritma.

1. Pengumpulan data dan Preprocessing data

Dalam penelitian ini, data diperoleh dari repositori online aplikasi Kaggle yaitu *lifestyle habits.* Berikut adalah tautan yang dapat diakses *https://www.kaggle.com/datasets/rustaas/kebiasaan-buruk-berdampak-ke-kesehatan*.

Tabel 1. Atribut Pada Data Status Kesehatan

No	Usia	Jenis kelamin	Merokok	Bekerja	Aktivitas Begadang	Aktivitas Olahraga	Pola makan teratur	Penyakit Bawaan	Hasil
1.	Muda	Pria	Aktif	Tidak	Iya	Jarang	Teratur	Tidak ada	Tidak
2.	Muda	Wanita	Pasif	Iya	Tidak	Sering	Kurang	Ada	Ya
3.	Muda	Wanita	Pasif	Iya	Tidak	Sering	Kurang	Ada	Ya
4.	Muda	Pria	Aktif	Tidak	Iya	Sering	Kurang	Tidak ada	Ya
5.	Muda	Pria	Aktif	Tidak	Iya	Jarang	Teratur	Tidak ada	Tidak
6.	Tua	Wanita	Aktif	Iya	Iya	Jarang	Teratur	Ada	Tidak

Langkah awal yang krusial dalam proses klasifikasi adalah preprocessing data, untuk memastikan bahwa data dalam kondisi siap olah dan sesuai untuk dimasukkan kedalam algoritma *machine learning*, beberapa tahapan dalam preprocessing yang diterapkan dalam penelitian ini meliputi:

a. Penanganan nilai kosong *(Missing Values)*

Pada tahap ini, data diperiksa untuk menemukan nilai kosong pada fitu-fitur tertentu. Jika ditemukan, nilai kosong diisi menggunakan Teknik *imputasi* seperti rata-rata atau modus tergantung pada jenisnya.

b. Normalisasi data

Normalisasi dilakukan untuk menyertakan skala fitur numerik agar algoritma yang berbasis jarak seperti K-NN dapat bekerja secara optimal. Algoritma K-NN menghitung jarak antar data, sehingga perbedaan skala antar fitur dapat memengaruhi hasil klasifikasi secara signifikan. Normalisasi yang digunakan adalah Min-Max Normalization:

$$x = \frac{x - x_{min}}{x_{max} - x_{min}} \quad [4]$$

sebaliknya, algoritma naïve bayes tidak bergantung pada skala fitur karena menggunakanlprinsip probabilitas dan distribusi data antar fitur. Oleh karena itu, normalisasi tidak bersifat wajib untuk naïve bayes, tetapi tetap dilakukan secara seragam untuk menjaga kesetaraan perlakuan terhadap seluruh dataset.

c. Label Encoding

Agar data dapat diproses oleh machine learning, maka proses label encoding ini adalah mengubah data kategorikal menjadi numerik, karena Sebagian besar algoritma, tidak dapat bekerja secara langsung dengan data bentuk asing[8], [9].

2. Pemisahan data

   Untuk membantu mengevaluasi kinerja machine learning, maka Kumpulan data dibagi menjadi subset training dan testing. Kemudian machine learning melatihnya pada satu bagian data dan mengujinya pada bagian yang lain [10].

3. Implementasi Naïve bayes
   a. Hitung probabilitas awal
   b. Probabilitas Kondisional
   c. Probabilitas Gabungan
4. Implementasi K-NN
   a. Data testing (sebagai target klasifikasi)
   b. Data training dan label (hasil)
   c. Euclidean distance
   d. Penentuan tetangga terdekat (k = 3)
   e. Prediksi kelas
5. Evaluasi Model

III. HASIL DAN PEMBAHASAN

Implementasi Algoritma Naïve Bayes

Naive Bayes merupakan suatu algoritma yang dapat mengklasifikasikan suatu variable tertentu dengan menggunakan metode probabilitas dan statistic. Naive bayes menggunkan sebuah ilmu cabang matematika yang dikenal juga dengan teori probabilitas untuk mencari peluang terbesar dari kemungkinan klasifikasi, dengan cara melihat frekuensi tiap klasifikasi pada data training

3.1. Hitung Probabilitas Awal

Probabilitas awal menunjukkan seberapa besar kemungkinan suatu kelas muncul dalam keseluruhan data sebelum mempertimbangkan dengan fitur lainnya.

$$P(ya) = \frac{jumlah\ data\ hasil{=}ya}{total\ data} = \frac{186}{369} = 0{,}50 \quad [5]$$

$$P(tidak) = \frac{jumlah\ data\ hasil{=}ya}{total\ data} = \frac{203}{369} = 0{,}55 \quad [6]$$

3.2. Probabilitas Kondisional

Probabilitas kondisional menghitung kemungkinan suatu fitur (atribut) yang muncul dalam kelas tertentu. Diperhitungkan untuk setiap fitur terhadap masing-masing kelas dengan menerapkan

Fakultas Teknik,
Universitas Wijaya Kusuma Surabaya
ISSN 3090-1154 (Media Online)
Volume 2, Tahun 2025

Laplace Smooting untuk menghindari hasil nol (0) jika tidak ada data yang sesuai. Probabilitas kondisional dari setiap atribut terhadap masing-masing kelas dengan persamaan berikut:

$$P(x_j|C_i) = \frac{N_{x_j \wedge C_i} + 1}{N_{C_i} + k} \quad [7]$$

Keterangan:

x_j = nilai dari fitur ke-j yang diamati

C_i = salah satu kelas target

$N_{x_j \wedge C_i}$ = jumlah data yang memiliki nilai fitur x_j dan termasuk dalam kelas C_i

N_{C_i} = jumlah total data yang termasuk dalam kelas C_i

k = jumlah kategori unik dalam fitur x_j

$+1\ dan + k =$ merupakan bagian dari laplace smoothing, untuk menghindari hasil 0 jika tidak ditemukan kombinasi fitur dan kelas di data, dari persamaan diatas,

Tabel 2. Probabilitas Kondisional Setiap Kelas

Variabel	Kategori	Hasil	Probabilitas Kondisional
Usia	Muda	Ya	0.56
	Muda	Tidak	0.39
	Tua	Ya	0.43
	Tua	Tidak	0.60
Jenis Kelamin	Wanita	Ya	0.95
	Wanita	Tidak	0.51
	Pria	Ya	0.43
	Pria	Tidak	0.57
Merokok	Aktif	Ya	0.06
	Aktif	Tidak	0.97
	Pasif	Ya	0.93
	Pasif	Tidak	0.02
Bekerja	No	Ya	0.21
	No	Tidak	0.56
	Yes	Ya	0.78
	Yes	Tidak	0.43
Aktifitas begadang	Iya	Ya	0.41
	Iya	Tidak	0.71
	Tidak	Ya	0.58
	Tidak	Tidak	0.28
Aktifitas olahraga	Jarang	Ya	0.25
	Jarang	Tidak	0.87
	Sering	Ya	0.74
	Sering	Tidak	0.12
Pola makan	Teratur	Ya	0.49
	Teratur	Tidak	0.88
	Kurang	Ya	0.50
	Kurang	Tidak	0.11
Penyakit bawaan	Ada	Ya	0.79
	Ada	Tidak	0.45
	Tidak ada	Ya	0.20
	Tidak ada	Tidak	0.54

3.3. Probabilitas Gabungan

Pada probabilitas gabungan, semua probabilitas kondisinonal dikalikan untuk setiap kelasnya, dan hasilnya dikalikan dengan probabilitas awal.

$$P(sehat|X) \propto P(sehat).P(L|sehat).P(Ya|Sehat).P(Tidak|sehat).P(Ya|sehat)$$

$P(ya\ |X)$
$= 0{,}56\ .\ 0{,}43\ .\ 0{,}96\ .\ 0{,}43\ .\ 0{,}06\ .\ 0{,}93\ .\ 0{,}21\ .\ 0{,}78\ .\ 0{,}41\ .\ 0{,}58\ .\ 0{,}25\ .\ 0{,}74.\ 0{,}49\ .0{,}50\ .\ 0{,}79\ .\ 0{,}20$
$= 1{,}53\ .\ 0{,}50 = 7{,}56$
$P(Tidak\ |X)$
$= 0{,}39\ .\ 0{,}60\ .\ 0{,}51\ .\ 0{,}57\ .\ 0{,}97\ .\ 0{,}02\ .\ 0{,}56\ .\ 0{,}43\ .\ 0{,}71\ .\ 0{,}28\ .\ 0{,}87\ .\ 0{,}12.\ 0{,}88\ .0{,}11\ .\ 0{,}45\ .\ 0{,}54$
$= 1{,}55\ .\ 0{,}55 = 8{,}53$

Implementasi Algoritma *K-Nearest Neighbor*

K-Nearest Neighbors (KNN) merupakan algoritma klasifikasi yang menggunakan himpunan nilai K dari data terdekat (tetangganya) sebagai acuan untuk menentukan kelas data baru. KNN mengklasifikasikan data berdasarkan kemiripan atau kedekatannya dengan data lain. Algoritma KNN ini adalah pembelajaran yang malas. Artinya, tidak menggunakan titik data pelatihan untuk membangun model. Dengan kata lain algoritma KNN mempunyai fase pelatihan yang sangat minim. Tujuan dari algoritma ini adalah untuk mengklasifikasikan objek baru berdasarkan atribut dan sampel pada data pelatihan. Untuk memberikan pemahaman konkret terhadap proses klasifikasi menggunakan algoritma K-NN, disajikan contoh penghitungan jarak euclidean antara satu data uji dan sejumlah data testing. Langkah ini merupakan bagian dari pemodelan dan sangat menentukan kelas dari data baru berdasarkan kedekatan terhadap data yang telah diketahui labelnya.

3.1. Data Testing (sebagai target klasifikasi)

Data testing yang akan diklasifikasikan terdiri dari delapan atribut yang telah diubah ke dalam numerik menggunakan label encoding. Nilai-nilai yang digunakan menggambarkan karakteristik seperti usia, jenis kelamin, merokok, dan sebagainya. Dengan vector data uji adalah berikut :

X = [0, 0, 0, 0, 0, 0, 0, 1]

Dimana, dengan kelas dimasing-masing atribut : usia muda, aktif merokok, tidak bekerja, sering begadang, jarang olahraga, pola makan tidak teratur, dan tidak ada penyakit bawaan.

3.2. Data Training dan Label (hasil)

Empat data training digunakan untuk membandingkan kedekatan dengan data uji. Masing – masing memiliki nilai atribut yang juga telah dinormalisasi secara label, serta dilengkapi dengan label kelas (0 atau 1).

Tabel 3. Data Training dan Label

No	Data training	Label (Hasil)
1.	[0, 1, 1, 1, 1, 1, 1, 0]	1 (ya)
2.	[0, 1, 1, 1, 1, 1, 1, 0]	1 (ya)
3.	[0, 0, 0, 0, 0, 0, 1, 0]	1 (ya)
4.	[0, 0, 0, 0, 0, 0, 1, 1]	0(Tidak)

3.3. *Euclidean Distance*

1. Jarak ke data training 1

Data training pertama memiliki banyak nilai berbeda terhadap data uji. Berikut perhitungannya :

$d_1 = \sqrt{(0-0)^2 + (0-1)^2 + (0-1)^2 + (0-1)^2 + (0-1)^2 + (0-1)^2 + (0-1)^2 + (0-1)^2 + (1-0)^2}$

$d_1 = \sqrt{0+1+1+1+1+1+1+1+1} = \sqrt{7} = 2{,}645$

2. Jarak ke data testing 2

Karena data training 2 identik dengan data training 1, hasil jaraknya juga sama :

$d_2 = \sqrt{(0-0)^2 + (0-1)^2 + (0-1)^2 + (0-1)^2 + (0-1)^2 + (0-1)^2 + (0-1)^2 + (0-1)^2 + (1-0)^2}$

$d_2 = \sqrt{0+1+1+1+1+1+1+1+1} = \sqrt{7} = 2{,}645$

3. Jarak ke data training 3

Data ini memiliki lebih banyak kesamaan dengan data testing, namun berbeda pada dua fitur terakhir :

$d_3 = \sqrt{(0-0)^2 + (0-0)^2 + (0-0)^2 + (0-0)^2 + (0-0)^2 + (0-0)^2 + (0-0)^2 + (0-1)^2 + (1-0)^2}$

$d_3 = \sqrt{0+0+0+0+0+0+0+1+1} = \sqrt{2} = 1{,}414$

4. Jarak ke data training 4

Data ini sangat mirip dengan data training 3, tetapi memiliki kecocokan penuh pada fitur ke 8:

$$d_3 = \sqrt{(0-0)^2 + (0-0)^2 + (0-0)^2 + (0-0)^2 + (0-0)^2 + (0-0)^2 + (0-0)^2 + (0-1)^2 + (1-1)^2}$$

$$d_3 = \sqrt{0+0+0+0+0+0+1+0} = \sqrt{2} = 1{,}000$$

5. Penentuan Tetangga Terdekat (k = 3)

Setelah seluruh jarak dihitung, tiga data latih dengan jarak terdekat dipilih sebagai referensi untuk klasifikasi. Berikut hasil peringkat berdasarkan kedekatan, dimana semakin kecil nilainya, maka semakin baik modelnya.

Tabel 4. Penentuan Tetangga Terdekat

Jarak	Data training	Label
1.000	y_4	0
1.414	y_3	1
2.645	y_1	1

6. Prediksi Kelas

Prediksi kelas, ditentukan dari banyaknya label yang muncul, label dari tiga tetangga terdekat yaitu, label 1 (ya) sebanyak 2 kali muncul, dan label 0 (tidak). Dengan demikian, label mayoritas adalah 1(ya) dan ditentukan sebagai prediksi akhir.

7. Evaluasi Model

Evaluasi model merupakan tahap penting dalam penelitian, guna mengukur dan membandingkan kinerja dua algoritma yang digunakan, yaitu naïve bayes dan K-NN, dalam mengklasifikasikan status Kesehatan. Evaluasi dilakukan dengan menggunakan beberapa metrik yang umum digunakan dalam klasifikasi, yaitu akurasi, presisi, recall, dan F1-score. Untuk memperoleh metrik evaluasi tersebut, penelitian ini menggunakan confusion matrix sebagai dasar penghitungan. Dimana, confusion matrix memberikan Gambaran performa klasifikasi berdsarkan empat komponen utama : True Positive (TP), True Negative (TN), False Positive (FP), dan False Negative (FN). Berdasarkan matriks ini, masing-masing metrik dihitung menggunakan rumus sebagai berikut:

Akurasi dimana metrik ini mengukur proporsi prediksi yang benar dari keseluruhan data, dengan rumus :

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

Selanjutnya presisi, metrik in menunjukkan seberapa tepat model dalam mengklasifikasikan data ke dalam kelas positif:

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

Lalu, recall atau sensitivitas yang menunjukkan seberapa banyak data positif yang berhasil diklasifikasikan dengan benar:

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

Terakhir adalah F1-Score adalah *harmonic mean* antara presisi dan recall, yang digunakan untuk mengukur keseimbangan keduanya :

$$F1 = 2.\frac{Presisi\ .Recall}{Presisi + Recall}$$

Metode evaluasi ini diterapkan pada hasil klasifikasi yang diperoleh dari kedua algoritma, baik naïve bayes maupun K-NN. Pengolahan data dan evaluasi model dilakukan dengan bantuan platform Google Colaboratory. Hasil evaluasi dari masing-masing model kemudian dianalisis untuk mengetahui algoritma mana yang memiliki performa lebih baik dalam mengklasifikasikan staus Kesehatan berdasarkan dataset yang digunakan.

Penelitian ini menggunakan dataset yang berisi atribut-atribut Kesehatan pada tabel Kesehatan diatas, serta hasil sebagai label. Dataset dibagi menjadi dua bagian, 80% sebagai data latih dan 20% sebagai data uji. Dengan menggunakan dua algoritma, yaitu naïve bayes dan K-NN serta diimplementasikan menggunakan bahasa pemrograman phyton pada platform Google Colaboratory. Hasil klasifikasi dari dua lagoritma di

evaluasi menggunakan metrik meliputi: akurasi, presisi, recall dan F1- Score. Tabel 5. Berikut menunjukkan hasil evaluasi berdasarkan confusion matrix yang dihasilkan.

Tabel 5. Hasil Evaluasi Model Naive Bayes dan K-NN

Sumber : hasil pengujian model, diolah menggunakan Google Colab

Algoritma	Akurasi	Kelas	Presisi	Recall	F1-Score
Naive Bayes	0.92	0	0.88	0.97	0.93
		1	0.97	0.87	0.92
K-NN	0.92	0	0.88	0.97	0.93
		1	0.97	0.87	0.92

Dari kedua model diatas, naive bayes dan K-NN, keduanya menunjukkan hasil akurasi yang sama, dimana presisi untuk kelas 1 sangat tinggi dikedua model, artinya prediksi positif sangat akurat. Recall untuk kelas 0 juga sangat tinggi, menunjukkan bahwa mayoritas kelas negatif dikenali dengan baik. Kelas 1 memiliki recall yang sedikit lebih rendah, menunjukkan bahwa ada beberapa data kelas 1 yang gagal dikenali.

Dari hasil evaluasi model pada tabel diatas, menunjukkan bahwa perbedaan kinerja antara kedua algoritma disebabkan oleh cara kerja masing-masing. Naive Bayes bekerja berdasarkan probabilitas dan mengasumsikan independensi antar fitur. Pendekatan ini cocok digunakan apabila fitur-fitur yang digunakan memang saling bebas secara statistik. Namun dalam praktiknya, banyak atribut kesehatan yang saling berkorelasi, sehingga dapat mempengaruhi akurasi prediksi Naive Bayes. Disisi lain, K-NN merupakan algoritma berbasis *instance* yang mengklasifikasikan data berdasarkan kemiripan *(similarity)* terhadap tetangga terdekatnya. Dengan kata lain, K-NN tidak membuat asumsi terhadap distribusi data, sehingga lebih fleksibel dalam menangani data nyata yang kompleks. Namun, kekurangan K-NN terletak pada efisiensi komputasinya, karena seluruh dataset perlu disimpan dan dihitung ulang untuk setiap prediksi.

IV. KESIMPULAN

Penelitian ini bertujuan untuk membandingkan kinerja algoritma naïve bayes dan K-NN dalam mengklasifikasikan status Kesehatan berdasarkan data yang mencakup variable usia, jenis kelamin, serta enam variable lainnya, dengan satu label berupa hasil. Tujuan tersebut telah tercapai melalui pengitungan algoritma dan hasil pengujian model yang di olah menggunakan Google Colab, dimana kedua algoritma menunjukkan akurasi yang sama. Artinya, kedua algoritma dapat digunakan secara efektif, tetapi jika data memiliki distribusi normal dan kontinu, naïve bayes bisa menjadi pilihan yang efisien. Sedangkan K-NN cocok untuk data yang tidak memerlukan asumsi distribusi dan lebih eksploratif terhadap pola-pola lokal data,

REFERENSI

[1] Fadhillah, R. A'la, and Z. Fatah, "Perbandingan Algoritma Decision Tree dan Deep Learning dalam," *Multidiciplinary Scientifict Journal*, vol. 2, 2024.

[2] N. Bhatia, "Survey of Nearest Neighbor Techniques," *Article in International Journal of Computer Science and Information Security*, vol. 8, no. 2, 2010, doi: 10.48550/arXiv.1007.0085.

[3] Powers, David M. W "Evaluation_From_Precision_Recall_and_F-Factor_to_R".

[4] M. Jannah, M. Arief, H. M. Kom, M. Al Fajar, and M. A. Hasan, "Perbandingan Metode Naïve Bayes Dan K-Nearest Neighbor Dalam Mengklasifikasi Status Pertumbuhan Anak Stunting (Studi Kasus : Posyandu Cemara)".

[5] D. Nasien *et al.*, "Perbandingan Implementasi Machine Learning Menggunakan Metode KNN, Naive Bayes, Dan Logistik Regression Untuk Mengklasifikasi Penyakit Diabetes," 2024.

[6] 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.

[7] J. Indriyanto, *Algoritma K-Nearest Neighbor untuk Prediksi Nasabah Asuransi*. Jawa Tengah: Penerbit NEM, 2021. Accessed: Jun. 03, 2025. [Online]. Available: https://www.google.co.id/books/edition/ALGORITMA_K_NEAREST_NEIGHBOR_UNTUK_PREDI/EE0tEAAAQBAJ

[8] S. Maesaroh *et al.*, *Bahasa Pemrograman Python*. Banten: PT Sada Kurnia Pustaka, 2024. Accessed: Jun. 05, 2025. [Online]. Available: https://www.google.co.id/books/edition/Bahasa_Pemrograman_Python/bOIKEQAAQBAJ

[9] G. Maulani *et al.*, *Machine Learning*. Jawa Barat: CV. Mega Press Nusantara, 2025. Accessed: Jun. 09, 2025. [Online]. Available: https://www.google.co.id/books/edition/Machine_Learning/RblPEQAAQBAJ

[10] A. Maulida Argina, "Penerapan Metode Klasifikasi K-Nearest Neighbor pada Dataset Penderita Penyakit Diabetes," *Indonesian Journal of Data and Science*, vol. 1, no. 2, Jul. 2020.