

Perbandingan Algoritma *Support Vector Machine* dan *Naïve Bayes* pada Klasifikasi Penyakit Tekanan Darah Tinggi (Studi Kasus: Klinik Polresta Samarinda)

Raka Putra Pridiptama^{1*}, Wasono Wasono¹, Fidia Deny Tisna Amijaya¹

¹Laboratorium Matematika Komputasi Program Studi Matematika Jurusan Matematika FMIPA Universitas Mulawarman

Dikirim: Desember 2023; Diterima: Maret 2024; Dipublikasi: Maret 2024

Alamat Email Korespondensi: rakapridiptama@gmail.com

Abstrak.

Klasifikasi adalah suatu proses untuk menemukan sifat-sifat yang sama dalam suatu himpunan data untuk dikelompokkan ke dalam kelas-kelas yang berbeda. Metode klasifikasi yang digunakan dalam penelitian ini adalah algoritma *support vector machine* (SVM) dan *naïve Bayes*. Algoritma SVM adalah algoritma *supervised learning* yang bekerja dengan mencari *hyperplane* atau fungsi pemisah terbaik untuk memisahkan kelas, sedangkan algoritma *naïve Bayes* adalah *supervised learning* yang didasarkan pada asumsi kemandirian (naif) antar prediktor yang dikenal dengan teorema Bayes. Penelitian ini bertujuan untuk mengetahui model dan keakuratan algoritma SVM dan *naïve Bayes* dalam melakukan klasifikasi terhadap status hipertensi dari rekam medis pasien di Klinik Polresta Samarinda tahun 2022. Berdasarkan hasil analisis akurasi pada algoritma SVM sebesar 96,67% dengan tepat mengklasifikasikan 29 dari 30 data sedangkan pada algoritma *naïve Bayes* sebesar 93,33% dengan tepat mengklasifikasikan 28 dari 30 data. Hasil perbandingan pengukuran akurasi dari kedua algoritma tersebut menunjukkan bahwa algoritma SVM memiliki tingkat akurasi yang lebih baik dibandingkan dengan algoritma *naïve Bayes*.

Kata Kunci:

Hipertensi, Klasifikasi, Naïve Bayes, SVM

PENDAHULUAN

Penyakit Tidak Menular (PTM) merupakan penyakit dengan pengobatan lama dan berbiaya mahal yang menyebabkan kematian tertinggi di Indonesia [1]. Berdasarkan data dari *World Health Organization* (WHO) menyatakan 66% penyebab kematian tertinggi di Indonesia salah satunya adalah tekanan darah tinggi (hipertensi) [2].

Hipertensi didefinisikan sebagai tekanan darah sistolik 130 mmHg atau lebih tinggi dan/atau tekanan darah diastolik 80 mmHg atau lebih tinggi. Hipertensi jarang menimbulkan gejala atau secara nyata membatasi kesehatan fungsional pasien. Hipertensi merupakan faktor risiko utama penyakit jantung koroner, gagal jantung, dan stroke [3]. Faktor penyebab terjadinya hipertensi adalah umur, jenis kelamin, genetik, kebiasaan merokok, obesitas, kurangnya aktivitas fisik, stres, dan konsumsi garam berlebihan [4].

Klasifikasi adalah salah satu teknik dalam *data mining* proses mempelajari fungsi tujuan yang digunakan untuk menghubungkan setiap himpunan atribut dari suatu objek dengan salah satu label kelas yang telah ditentukan sebelumnya [5]. Algoritma yang termasuk dalam metode klasifikasi antara lain adalah regresi logistik, pohon keputusan, *random forest*, *support vector machine* (SVM), *k-nearest neighbour* (K-NN) dan *naïve Bayes* [6].

Mulaab menyebutkan algoritma *support vector machine* (SVM) memetakan setiap *item* data sebagai titik (*vector*) dalam ruang n-dimensi, di mana setiap nilai fitur menjadi koordinat. Algoritma *naïve Bayes* adalah pengelompokan dengan mengasumsikan bahwa kejadian fitur tertentu di kelas tidak terkait dengan kehadiran fitur lainnya [7].

Terdapat beberapa penelitian yang sudah dilakukan terkait masalah mengklasifikasikan atribut menggunakan algoritma SVM dan algoritma *naïve Bayes*. Misalnya penelitian implementasi algoritma *naïve Bayes* dan SVM pada klasifikasi penyakit kardiovaskular [8]. Hasil penelitian yang diperoleh dari algoritma *naïve Bayes* sebesar 71% dan algoritma SVM sebesar 65% ketika menggunakan *K-Fold Cross Validation*. Selanjutnya [9] perbandingan performa algoritma SVM dan *naïve Bayes* pada diagnosa diabetes, diperoleh hasil bahwa akurasi SVM lebih unggul dibandingkan dengan *naïve Bayes*, yaitu pada *data split* 80:20 SVM mendapatkan akurasi sebesar 73% dibandingkan dengan *naïve Bayes* sebesar 72%.

Alasan penulis memilih algoritma SVM dan *naïve Bayes* adalah algoritma SVM memiliki kemampuan untuk menangani data non-linear dan kompleks, yang sering ditemukan dalam data medis, mampu memisahkan data dengan hyperplane yang optimal, sehingga menghasilkan klasifikasi yang akurat, dan telah terbukti efektif dalam berbagai penelitian klasifikasi penyakit, termasuk hipertensi. Dan algoritma *naïve Bayes* memiliki memiliki asumsi independensi fitur yang membuatnya mudah untuk dipelajari dan diimplementasikan serta bekerja dengan baik dengan dataset kecil, yang sering terjadi dalam penelitian medis.

Berdasarkan dari uraian yang telah dijelaskan, maka penulis tertarik untuk melakukan penelitian dengan judul "Perbandingan Algoritma *Support Vector Machine* dan *Naïve Bayes* pada Klasifikasi Penyakit Tekanan Darah Tinggi (Studi Kasus: Klinik Polresta Samarinda)".

METODE PENELITIAN

Penelitian ini termasuk ke dalam penelitian kuantitatif dengan populasi yang digunakan adalah seluruh pasien di Klinik Polresta Samarinda. Adapun sampel yang digunakan adalah pasien dengan penyakit tekanan darah tinggi dan pasien dengan tekanan darah normal di Klinik Polresta Samarinda Tahun 2022. Pada penelitian ini tahapan analisis data yang dilakukan adalah sebagai berikut:

1. Pembersihan data

Pembersihan data adalah serangkaian proses yang dimulai dari persiapan data untuk mengidentifikasi kesalahan pada data, kemudian melakukan tindakan perbaikan ataupun penghapusan data.

2. Transformasi data

Transformasi data adalah salah satu proses dari *data preprocessing* untuk mengubah atribut sesuai dengan format yang dapat diproses dalam program dan manual.

3. Pembagian data latih dan uji

Pembagian data dilakukan memperoleh data latih dan data uji yang digunakan untuk menguji dan mengevaluasi data. Pada penelitian ini proporsi data latih dan data uji yang digunakan adalah 70:30.

4. Klasifikasi *support vector machine*

Algoritma SVM diterapkan dengan langkah-langkah sebagai berikut:

- Membaca data latih.
- Menghitung nilai kernel dengan fungsi kernel polinomial sebagai berikut:

$$K(x_i, x_j) = (x_i^T \times x_j + 1)^d \tag{1}$$

dengan:

x_i = data ke-i

x_j = data ke-j

d = derajat

- Menghitung nilai *alfa* menggunakan persamaan berikut:

$$\alpha = \frac{N}{K(N * N)} \tag{2}$$

- Menghitung nilai *weight* menggunakan persamaan berikut:

$$w = \alpha \cdot y_i \cdot K_i \tag{3}$$

- Menghitung nilai *bias* menggunakan persamaan berikut:

$$b = -\frac{1}{2} (w \cdot x^+ + w \cdot x^-) \tag{4}$$

- Membangun *hyperplane* menggunakan persamaan berikut:

$$w \cdot x_i + b = 0 \tag{5}$$

- Membangun model SVM menggunakan persamaan berikut:

$$f(\Phi(X)) = \text{sign}(w \cdot \Phi(X) + b) \tag{6}$$

5. Klasifikasi *naïve* Bayes

Algoritma *naïve* Bayes diterapkan dengan langkah-langkah sebagai berikut:

- Membaca data latih.
- Menghitung nilai *prior* data latih.
- Menghitung nilai peluang setiap variabel terhadap setiap kelasnya dengan menggunakan rumus peluang bersyarat sebagai berikut

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad (7)$$

dengan:

$P(A|B)$ = peluang terjadi A dengan syarat B telah terjadi

$P(A \cap B)$ = peluang awal A dengan petunjuk B telah terjadi secara simultan

$P(B)$ = peluang terjadinya B

Apabila dataset bertipe numerik, maka digunakan perhitungan distribusi Gaussian. Untuk tipe data numerik atau kontinu, maka persamaan *naïve* Bayes menjadi:

$$P(X_i = x_i | Y = y_i) = \frac{1}{\sqrt{2\pi}\sigma_{ij}} \exp\left(-\frac{(x_i - \mu_{ij})^2}{2\sigma_{ij}^2}\right). \quad (8)$$

Adapun rumus untuk menghitung masing-masing nilai rata-rata (μ) dan standar deviasi (σ) adalah sebagai berikut:

$$\mu = \frac{\sum_{i=1}^n x_i}{n}$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^n (x_i - \mu)^2}{n - 1}}$$

- Menghitung nilai akumulasi peluang dari setiap kelas menggunakan persamaan berikut

$$P(C|F_1, \dots, F_n) = P(C) \cdot \prod_{i=1}^n P(F_i|C) \quad (9)$$

- Menghitung perkalian peluang *prior* dengan nilai akumulasi peluang variabel bebas dari masing-masing kelas.

6. Evaluasi model klasifikasi

Evaluasi model klasifikasi bertujuan untuk mengetahui tingkat akurasi model klasifikasi yang telah dibuat dengan rumus:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (10)$$

dengan keterangan sebagai berikut:

- (a) *True Positive* (TP) adalah jumlah dokumen dari kelas positif yang benar diklasifikasikan sebagai kelas positif.

- (b) *True Negative* (TN) adalah jumlah dokumen dari kelas negatif yang benar diklasifikasikan sebagai kelas negatif.
- (c) *False Positive* (FP) adalah jumlah dokumen dari kelas negatif yang salah diklasifikasikan sebagai kelas positif.
- (d) *False Negative* (FN) adalah jumlah dokumen dari kelas positif yang salah diklasifikasikan sebagai kelas negatif.

Empat keterangan di atas dapat digambarkan dengan matriks konfusi yang diilustrasikan pada Tabel 1 berikut

Tabel 1. Matriks Konfusi

Nilai Aktual	Nilai Prediksi	
	Positif	Negatif
Positif	TP	FN
Negatif	FP	TN

HASIL DAN PEMBAHASAN

Data yang digunakan dalam penelitian ini berjumlah 100 data dengan tujuh variabel yaitu usia, jenis kelamin, tekanan sistolik, tekanan diastolik, obesitas, merokok dan status hipertensi. Teknik analisis data yang dilakukan pertama kali adalah menggunakan teknik analisis deskriptif untuk mengetahui gambaran dari keseluruhan data sampel yang digunakan. Langkah selanjutnya adalah membagi data latih dan data uji sesuai dengan proporsi pembagian 70:30. Adapun data latih yang telah dilakukan pengacakan dapat dilihat pada Tabel 2.

Tabel 2. Data Latih

No.	Usia (X_1)	Jenis Kelamin (X_2)	Tekanan Sistolik (X_3)	Tekanan Diastolik (X_4)	Obesitas (X_5)	Merokok (X_6)	Status Hipertensi (Y)
1	33	L	163	133	Tidak	Tidak	Hipertensi
2	53	L	154	120	Tidak	Ya	Hipertensi
3	21	L	161	115	Tidak	Tidak	Hipertensi
⋮	⋮	⋮	⋮	⋮	⋮	⋮	
99	26	L	142	109	Tidak	Ya	Hipertensi
100	53	L	169	118	Tidak	Tidak	Hipertensi

Algoritma *Support Vector Machine*

Klasifikasi SVM dengan fungsi kernel polinomial menggunakan parameter derajat (d) yaitu 1. Pemilihan nilai parameter ini bertujuan untuk meningkatkan efisiensi dan mengurangi kompleksitas pembuatan model. Pada contoh perhitungan manual digunakan sebanyak **6 data**, yaitu data ke-1 hingga data ke-6. Adapun keenam data latih tersebut dapat dilihat pada Tabel 3 berikut.

Tabel 3. Data Ke-1 hingga Data Ke-6

No.	X_1	X_2	X_3	X_4	X_5	X_6	Y
1	33	0	163	133	0	0	1
2	53	0	154	120	0	1	1
3	21	0	161	115	0	0	1
4	37	0	115	74	0	0	-1
5	41	0	118	90	1	0	-1
6	39	0	146	118	1	0	1

Langkah-langkah yang perlu dilakukan dalam klasifikasi menggunakan SVM adalah sebagai berikut:

1. Menghitung nilai kernel latih

Langkah pertama yang perlu dilakukan dalam perhitungan klasifikasi SVM adalah dengan mencari nilai kernel. Adapun persamaan yang digunakan dalam penelitian adalah Persamaan (1). Parameter yang digunakan dalam perhitungan adalah $d = 1$, sehingga didapatkan perhitungan sebagai berikut:

$$K(x_i, x_j) = (x_i^T \times x_j + 1)^d$$

$$\begin{aligned} K(x_1, x_1) &= ((33 \cdot 33) + (0 \cdot 0) + (163 \cdot 163) + (133 \cdot 133) + (0 \cdot 0) + (0 \cdot 0) + 1)^1 \\ &= 45348 \end{aligned}$$

$$\begin{aligned} K(x_1, x_2) &= ((33 \cdot 53) + (0 \cdot 0) + (163 \cdot 154) + (133 \cdot 120) + (0 \cdot 0) + (0 \cdot 1) + 1)^1 \\ &= 42812 \end{aligned}$$

$$\begin{aligned} K(x_1, x_3) &= ((33 \cdot 21) + (0 \cdot 0) + (163 \cdot 161) + (133 \cdot 115) + (0 \cdot 0) + (0 \cdot 0) + 1)^1 \\ &= 42232 \end{aligned}$$

Perhitungan dilakukan dengan cara yang sama untuk data selanjutnya, sehingga akan mendapatkan nilai kernel membentuk sebuah matriks kernel $K(6 \times 6)$. Berdasarkan perhitungan yang telah dilakukan, maka diperoleh hasil yang ditunjukkan sebagai berikut:

$$\begin{bmatrix} 45348 & 42812 & 42232 & 29809 & 32558 & 40780 \\ 42812 & 40927 & 39708 & 28552 & 31146 & 38712 \\ 42232 & 39708 & 39588 & 27803 & 30210 & 37896 \\ 29809 & 28552 & 27803 & 20071 & 21748 & 26966 \\ 32558 & 31146 & 30210 & 21748 & 23707 & 29449 \\ 40780 & 38712 & 37896 & 26966 & 29449 & 36763 \end{bmatrix}$$

2. Menghitung nilai alfa (α)

Untuk mendapatkan nilai alfa, langkah pertama adalah menjumlahkan semua nilai dari matriks kernel dan memeriksa label pada tabel data latih. Sebelum melakukan penjumlahan nilai matriks kernel, akan dicari terlebih dahulu nilai $y_i \times y_j$.

$$Y = [1 \ 1 \ 1 \ -1 \ -1 \ 1]$$

$$Z = Y^T \times Y$$

$$Z = \begin{bmatrix} 1 & 1 & 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & -1 & -1 & 1 \\ -1 & -1 & -1 & 1 & 1 & -1 \\ -1 & -1 & -1 & 1 & 1 & -1 \\ 1 & 1 & 1 & -1 & -1 & 1 \end{bmatrix}$$

Selanjutnya, akan dilakukan *dot product* antara matriks Z dan matriks kernel K, sehingga diperoleh matriks kernel baru K:

$$\begin{bmatrix} 45348 & 42812 & 42232 & -29809 & -32558 & 40780 \\ 42812 & 40927 & 39708 & -28552 & -31146 & 38712 \\ 42232 & 39708 & 39588 & -27803 & -30210 & 37896 \\ -29809 & -28552 & -27803 & 20071 & 21748 & -26966 \\ -32558 & -31146 & -30210 & 21748 & 23707 & -29449 \\ 40780 & 38712 & 37896 & -26966 & -29449 & 36763 \end{bmatrix}$$

Kemudian menjumlahkan nilai matriks perkolom untuk mendapatkan nilai jumlah matriks kernel K(N*N):

45348	42812	42232	-29809	-32558	40780	
42812	40927	39708	-28552	-31146	38712	
42232	39708	39588	-27803	-30210	37896	
-29809	-28552	-27803	20071	21748	-26966	
-32558	-31146	-30210	21748	23707	-29449	
40780	38712	37896	-26966	-29449	36763	+
108805	102461	101411	-71311	-77908	97736	

Dengan mensubstitusikan nilai N dan jumlah matriks kernel dari setiap kolom yang telah dicari sebelumnya ke dalam Persamaan (2), diperoleh nilai alfa

$$\alpha = \frac{N}{K(N * N)}$$

$$\alpha = \frac{6}{108805 + 102461 + 101411 + (-71311) + (-77908) + 97736}$$

$$\alpha = \frac{6}{261194}$$

$$\alpha = 0,000023$$

3. Menghitung nilai *weight* (w) dan bias (b)

Setelah memperoleh nilai alfa, langkah selanjutnya yaitu dengan mencari nilai *weight* dan bias. Untuk mencari nilai *weight* (w), digunakan Persamaan (3). Dengan demikian diperoleh:

$$\begin{aligned} w_i &= \alpha \cdot y_i \cdot K_i \\ w_i &= 0,00023 \cdot [1 \quad 1 \quad 1 \quad -1 \quad -1 \quad 1] \cdot \\ &\quad [108805 \quad 102461 \quad 101411 \quad -71311 \quad -77908 \quad 97736] \\ w_i &= [2,5025 \quad 2,3566 \quad 2,3325 \quad 1,6402 \quad 1,7919 \quad 2,2479] \end{aligned}$$

Dan seterusnya untuk setiap *weight* pada kolom. Setelah didapatkan nilai w , selanjutnya memilih salah satu *support vector* dari kelas -1 dan $+1$ dari matriks kernel K yang mempunyai nilai positif ($+$) dan negatif ($-$). Pada penelitian ini, diambil pada kolom pertama dan keempat untuk menghitung nilai bias (b) menggunakan Persamaan (4).

$$\begin{aligned} b &= -\frac{1}{2} (w \cdot x^+ + w \cdot x^-) \\ b &= -\frac{1}{2} ((2,5025 \cdot 45348 + 2,3566 \cdot 42812 + \dots + 2,2479 \cdot 40780) + \\ &\quad (2,5025 \cdot 29809 + 2,3566 \cdot 28552 + \dots + 2,2479 \cdot 26966)) \\ b &= -\frac{1}{2} (511783,0332 + 339240,73) \\ b &= -425511,8816 \end{aligned}$$

4. Menentukan *hyperplane*

Berdasarkan perhitungan nilai *weight* (w) dan bias (b) yang telah dilakukan, dengan Persamaan (5) maka didapatkan persamaan *hyperplane* sebagai berikut:

$$\begin{aligned} 2,5025x_1 + 2,3566x_2 + 2,3325x_3 + 1,6402x_4 + \\ 1,7919x_5 + 2,2479x_6 - 425511,8816 = 0 \end{aligned}$$

5. Membangun persamaan klasifikasi SVM

Setelah didapatkan nilai alfa dan bias, maka selanjutnya dapat dibentuk model SVM yang digunakan dalam proses klasifikasi dengan menggunakan Persamaan (6) sebagai berikut:

$$\begin{aligned} f(\Phi(X)) &= \text{sign}(w_i \cdot \Phi(X) + b) \\ &= \text{sign}(w_i \cdot K(X_i, X_j) - 425511,8816) \end{aligned}$$

dengan w_i adalah nilai *weight* yang telah diperoleh sebelumnya dan $\Phi(X) = K(X_i, X_j)$ adalah matriks kernel dari data uji.

6. Menghitung nilai kernel uji

Langkah terakhir sebelum melakukan prediksi adalah menghitung matriks kernel uji. Sebagai contoh data uji yang digunakan adalah data rekam medis ke-1 yang dapat dilihat pada Tabel 4.

Tabel 4. Data Uji Pertama

No.	X_1	X_2	X_3	X_4	X_5	X_6	Y
1	49	L	162	110	Tdk Obesitas	Merokok	?

Perhitungan untuk mendapatkan nilai matriks kernel uji dapat dilakukan dengan cara yang serupa dengan mencari nilai matriks kernel latih. Adapun persamaan yang digunakan adalah Persamaan (1) dengan parameter $d = 1$, x_i^T adalah data dari data uji dan x_j adalah data dari data latih, sehingga diperoleh hasil yang ditunjukkan sebagai berikut:

$$[42654 \quad 40747 \quad 39762 \quad 28584 \quad 31026 \quad 38544]$$

7. Prediksi menggunakan SVM

Klasifikasi SVM untuk keseluruhan data akan dilakukan dengan menggunakan bantuan Python. Sebagai contoh, pada data uji yang pertama dengan menggunakan model yang telah dibuat sebelumnya, dapat diperoleh prediksi kelas sebagai berikut:

$$\begin{aligned} f(\Phi(X)) &= \text{sign}(w_i \cdot \Phi(X) + b) \\ &= \text{sign}(w_i \cdot K(X_i, X_j) - 425511, 8816) \\ &= \text{sign}((2, 5025 \cdot 42654) + (2, 3566 \cdot 40747) + (2, 3325 \cdot 39762) + \\ &\quad (1, 6402 \cdot 28584) + (1, 7919 \cdot 31026) + (2, 2479 \cdot 38544) - 425511, 8816) \\ &= \text{sign}(59121, 0224) = 1. \end{aligned}$$

Diperoleh hasil perhitungan *sign* rekam medis pasien pada data uji pertama adalah 59121, 0224. Berdasarkan hasil perhitungan *sign*, dapat diketahui bahwa nilai yang diperoleh oleh *sign* bernilai positif. Dengan demikian, dapat ditarik kesimpulan bahwa pada data uji pertama diklasifikasikan masuk ke dalam kelas +1 yaitu rekam medis pasien dengan status hipertensi.

Algoritma Naïve Bayes

Berikut ini adalah contoh perhitungan manual algoritma *naïve* Bayes dengan proporsi data latih sebesar 70% dan data uji sebesar 30%.

1. Perhitungan Nilai Peluang *Prior* dari Setiap Kelas

Dengan proporsi 70% data latih dan 30% data uji diperoleh jumlah data latih sebanyak 70 orang dan data uji sebanyak 30 orang. Langkah awal klasifikasi menggunakan algoritma *naïve* Bayes adalah menghitung nilai peluang *prior* pada kedua kelas dalam data latih menggunakan rumus

$$P(Y_i) = \frac{\text{Jumlah data dengan dengan kelas } Y_i}{\text{Jumlah data latih}} = \frac{n(Y_i)}{N}.$$

- (a) Kelas pertama (rekam medis dengan status tidak hipertensi)

Pada 70 data latih, terdapat 17 rekam medis dengan status tidak hipertensi, sehingga nilai peluang *prior* kelas pertama sebagai berikut:

$$P(Y_1 = TH) = \frac{n(TH)}{n(DataLatih)} = \frac{17}{70} = 0,2429$$

- (b) Kelas kedua (rekam medis dengan status hipertensi)

Pada 70 data latih, terdapat 53 rekam medis dengan status hipertensi, sehingga nilai peluang *prior* kelas kedua sebagai berikut:

$$P(Y_2 = H) = \frac{n(H)}{n(DataLatih)} = \frac{53}{70} = 0,7571$$

2. Perhitungan Nilai Peluang Setiap Variabel terhadap Setiap Kelas

Perhitungan nilai peluang setiap variabel pada kedua kelas berdasarkan masing-masing peluang dari data latih menggunakan peluang bersyarat yaitu Persamaan 7 untuk tipe data diskrit/kategorikal atau Persamaan 8 untuk tipe data kontinu/numerikal.

- Usia (X_1)

Data uji pertama menunjukkan usia 49 tahun memiliki tipe data numerikal, sehingga untuk mencari nilai peluangnya digunakan Persamaan 8. Diketahui bahwa rata-rata, variasi dan standar deviasi dari variabel usia terhadap kelasnya dapat dilihat pada Tabel 5.

Tabel 5. Parameter Variabel Usia pada Setiap Kelas

Usia	Status Hipertensi	
	Tidak Hipertensi	Hipertensi
Rata-rata	39,4706	39,5849
Variansi	96,2647	88,4013
Std. Deviasi	9,8115	9,4022

Nilai peluang variabel usia (X_1) dengan nilai "49" terhadap kelas "Tidak Hipertensi" melalui perhitungan sebagai berikut:

$$P(X_1 = 49|TH) = \frac{1}{\sqrt{2\pi} \cdot 9,8115} \exp\left(-\frac{(49 - 39,4706)^2}{2 \cdot 96,2647}\right) = 0,0427$$

sehingga nilai peluang variabel usia dengan nilai "49" pada kelas "Tidak Hipertensi" adalah 0,0427.

Nilai peluang variabel usia (X_1) dengan nilai "49" terhadap kelas "Hipertensi" melalui perhitungan sebagai berikut:

$$P(X_1 = 49|H) = \frac{1}{\sqrt{2\pi} \cdot 9,4022} \exp\left(-\frac{(49 - 39,5849)^2}{2 \cdot 88,4013}\right) = 0,0448$$

sehingga nilai peluang variabel usia dengan nilai "49" pada kelas "Hipertensi" adalah 0,0448.

- **Jenis Kelamin (X_2)**

Data uji pertama menunjukkan jenis kelamin "Laki-laki" memiliki tipe data kategorikal, sehingga untuk mencari nilai peluangnya digunakan Persamaan 7. Diketahui bahwa dari 57 rekam medis dengan kategori "Laki-laki", terdapat 15 rekam medis berada di dalam kelas "Tidak Hipertensi" dan 57 rekam medis berada di dalam kelas "Hipertensi".

Nilai peluang variabel jenis kelamin (X_2) dengan kategori "Laki-laki" terhadap kelas "Tidak Hipertensi" melalui perhitungan sebagai berikut:

$$P(X_2 = L|TH) = \frac{P(L \cap TH)}{P(TH)} = \frac{15/70}{57/70} = 0,2632$$

sehingga nilai peluang variabel jenis kelamin dengan kategori "Laki-laki" pada kelas "Tidak Hipertensi" adalah 0,2632.

Nilai peluang variabel jenis kelamin (X_2) dengan kategori "Laki-laki" terhadap kelas "Hipertensi" melalui perhitungan sebagai berikut:

$$P(X_2 = L|H) = \frac{P(L \cap H)}{P(H)} = \frac{42/70}{57/70} = 0,7368$$

sehingga nilai peluang variabel jenis kelamin dengan kategori "Laki-laki" pada kelas "Hipertensi" adalah 0,7368.

- **Tekanan Sistolik (X_3)**

Data uji pertama menunjukkan tekanan sistolik 162 mmHg memiliki tipe data numerikal, sehingga untuk mencari nilai peluangnya digunakan Persamaan 8. Diketahui bahwa rata-rata, variasi dan standar deviasi dari variabel tekanan sistolik terhadap kelasnya dapat dilihat pada Tabel 6.

Tabel 6. Parameter Variabel Tekanan Sistolik pada Setiap Kelas

Tekanan Sistolik	Status Hipertensi	
	Tidak Hipertensi	Hipertensi
Rata-rata	114,6471	154,9245
Variansi	43,6176	88,8019
Std. Deviasi	6,6044	9,4235

Nilai peluang variabel tekanan sistolik (X_3) dengan nilai "162" terhadap kelas "Tidak Hipertensi" melalui perhitungan sebagai berikut:

$$P(X_3 = 162|TH) = \frac{1}{\sqrt{2\pi} \cdot 6,6044} \exp\left(-\frac{(162 - 114,6471)^2}{2 \cdot 43,6176}\right) = 0,1039$$

sehingga nilai peluang variabel tekanan sistolik dengan nilai "162" pada kelas "Tidak Hipertensi" adalah 0,1039.

Nilai peluang variabel tekanan sistolik (X_3) dengan nilai "162" terhadap kelas "Hipertensi" melalui perhitungan sebagai berikut:

$$P(X_3 = 162|H) = \frac{1}{\sqrt{2\pi} \cdot 9,4235} \exp\left(-\frac{(162 - 154,9245)^2}{2 \cdot 88,8019}\right) = 0,0441$$

sehingga nilai peluang variabel tekanan sistolik dengan nilai "162" pada kelas "Hipertensi" adalah 0,0441.

- **Tekanan Diastolik (X_4)**

Data uji pertama menunjukkan tekanan diastolik 110 mmHg memiliki tipe data numerikal, sehingga untuk mencari nilai peluangnya digunakan Persamaan 8. Diketahui bahwa rata-rata, variasi dan standar deviasi dari variabel tekanan diastolik terhadap kelasnya dapat dilihat pada Tabel 7.

Tabel 7. Parameter Variabel Tekanan Diastolik pada Setiap Kelas

Tekanan Diastolik	Status Hipertensi	
	Tidak Hipertensi	Hipertensi
Rata-rata	78,2941	119,5660
Variansi	28,5956	89,4811
Std. Deviasi	5,3475	9,4594

Nilai peluang variabel tekanan diastolik (X_4) dengan nilai "110" terhadap kelas "Tidak Hipertensi" melalui perhitungan sebagai berikut:

$$P(X_4 = 110|TH) = \frac{1}{\sqrt{2\pi} \cdot 5,3475} \exp\left(-\frac{(110 - 78,2941)^2}{2 \cdot 28,5956}\right) = 0,1299$$

sehingga nilai peluang variabel tekanan diastolik dengan nilai "110" pada kelas "Tidak Hipertensi" adalah 0,1299.

Nilai peluang variabel tekanan diastolik (X_4) dengan nilai "110" terhadap kelas "Hipertensi" melalui perhitungan sebagai berikut:

$$P(X_4 = 110|H) = \frac{1}{\sqrt{2\pi} \cdot 9,4594} \exp\left(-\frac{(110 - 119,5660)^2}{2 \cdot 89,4811}\right) = 0,04$$

sehingga nilai peluang variabel tekanan diastolik dengan nilai "110" pada kelas "Hipertensi" adalah 0,04.

- **Obesitas (X_5)**

Data uji pertama menunjukkan obesitas "Tidak Obesitas" memiliki tipe data kategorikal, sehingga untuk mencari nilai peluangnya digunakan Persamaan 7.

Diketahui bahwa dari 64 rekam medis dengan kategori "Tidak Obesitas", terdapat 14 rekam medis berada di dalam kelas "Tidak Hipertensi" dan 50 rekam medis berada di dalam kelas "Hipertensi".

Nilai peluang variabel obesitas (X_5) dengan kategori "Tidak Obesitas" terhadap kelas "Tidak Hipertensi" melalui perhitungan sebagai berikut:

$$P(X_5 = \text{Tidak Obesitas} | TH) = \frac{P(\text{Tidak Obesitas} \cap TH)}{P(TH)} = \frac{14/70}{64/70} = 0,2188$$

sehingga nilai peluang variabel obesitas dengan kategori "Tidak Obesitas" pada kelas "Tidak Hipertensi" adalah 0,2188.

Nilai peluang variabel obesitas (X_5) dengan kategori "Tidak Obesitas" terhadap kelas "Hipertensi" melalui perhitungan sebagai berikut:

$$P(X_5 = \text{Tidak Obesitas} | H) = \frac{P(\text{Tidak Obesitas} \cap H)}{P(H)} = \frac{50/70}{64/70} = 0,7812$$

sehingga nilai peluang variabel obesitas dengan kategori "Tidak Obesitas" pada kelas "Hipertensi" adalah 0,7812.

- Merokok (X_6)

Data uji pertama menunjukkan merokok "Merokok" memiliki tipe data kategorikal, sehingga untuk mencari nilai peluangnya digunakan Persamaan 7. Diketahui bahwa dari 54 rekam medis dengan kategori "Merokok", terdapat 13 rekam medis berada di dalam kelas "Tidak Hipertensi" dan 41 rekam medis berada di dalam kelas "Hipertensi".

Nilai peluang variabel merokok (X_6) dengan kategori "Merokok" terhadap kelas "Tidak Hipertensi" melalui perhitungan sebagai berikut:

$$P(X_6 = \text{Merokok} | TH) = \frac{P(\text{Merokok} \cap TH)}{P(TH)} = \frac{13/70}{54/70} = 0,2407$$

sehingga nilai peluang variabel merokok dengan kategori "Merokok" pada kelas "Tidak Hipertensi" adalah 0,2407.

Nilai peluang variabel merokok (X_6) dengan kategori "Merokok" terhadap kelas "Hipertensi" melalui perhitungan sebagai berikut:

$$P(X_6 = \text{Merokok} | H) = \frac{P(\text{Merokok} \cap H)}{P(H)} = \frac{41/70}{54/70} = 0,7593$$

sehingga nilai peluang variabel merokok dengan kategori "Merokok" pada kelas "Hipertensi" adalah 0,7593.

3. Perhitungan Perkalian Peluang *Prior* dengan Nilai Akumulasi Peluang dari Masing-masing Kelas

Setelah mencari nilai probabilitas setiap variabel terhadap setiap kelas, maka langkah selanjutnya adalah menghitung perkalian peluang *prior* dengan nilai akumulasi dari setiap kelas dengan cara mengalikan semua peluang pada kelas. Adapun perhitungan nilai peluang *prior* dan nilai akumulasi peluang berdasarkan Persamaan 9 dapat dilihat sebagai berikut:

- Kelas pertama (rekam medis pasien dengan status tidak hipertensi)

$$\begin{aligned}
 &P(Y = -1|X_1 = 49, X_2 = L, X_3 = 162, X_4 = 110, \\
 &\quad X_5 = \text{Tidak Obesitas}, X_6 = \text{Merokok}) \\
 &= P(Y) \cdot P(X_1|Y) \cdot P(X_2|Y) \cdot P(X_3|Y) \cdot P(X_4|Y) \cdot P(X_5|Y) \cdot P(X_6|Y) \\
 &= P(Y = -1) \cdot P(X_1 = 49|-1) \cdot P(X_2 = L|-1) \cdot P(X_3 = 162|-1) \cdot \\
 &\quad P(X_4 = 110|-1) \cdot P(X_5 = \text{Tidak Obesitas}|-1) \cdot P(X_6 = \text{Merokok}|-1) \\
 &= 0,2429 \cdot 0,0427 \cdot 0,2632 \cdot 0,1039 \cdot 0,1299 \cdot 0,2188 \cdot 0,25 \\
 &= 0,000002.
 \end{aligned}$$

Diperoleh hasil dari perhitungan nilai peluang *prior* dan nilai akumulasi peluang pada kelas pertama adalah 0,000002.

- Kelas kedua (rekam medis pasien dengan status hipertensi)

$$\begin{aligned}
 &P(Y = 1|X_1 = 49, X_2 = L, X_3 = 162, X_4 = 110, \\
 &\quad X_5 = \text{Tidak Obesitas}, X_6 = \text{Merokok}) \\
 &= P(Y) \cdot P(X_1|Y) \cdot P(X_2|Y) \cdot P(X_3|Y) \cdot P(X_4|Y) \cdot P(X_5|Y) \cdot P(X_6|Y) \\
 &= P(Y = 1) \cdot P(X_1 = 49|1) \cdot P(X_2 = L|1) \cdot P(X_3 = 162|1) \cdot P(X_4 = 110|1) \cdot \\
 &\quad P(X_5 = \text{Tidak Obesitas}|1) \cdot P(X_6 = \text{Merokok}|1) \\
 &= 0,7571 \cdot 0,0448 \cdot 0,7368 \cdot 0,0441 \cdot 0,04 \cdot 0,7812 \cdot 0,75 \\
 &= 0,00003.
 \end{aligned}$$

Diperoleh hasil dari perhitungan nilai peluang *prior* dan nilai akumulasi peluang pada kelas kedua adalah 0,00003.

Berdasarkan hasil perhitungan dapat diketahui bahwa kelas yang memiliki nilai peluang terbesar adalah rekam medis medis pasien dengan kelas status 1 yaitu sebesar 0,00003, sehingga data uji pertama diklasifikasikan masuk ke dalam kelas kedua yaitu rekam medis pasien dengan status "Hipertensi".

Evaluasi Model Klasifikasi

Pada data uji selanjutnya menggunakan cara perhitungan yang sama dengan langkah-langkah sebelumnya, sehingga hasil klasifikasi dari kedua model dapat dilihat pada Tabel 8.

Tabel 8. Hasil Klasifikasi SVM dan *Naïve* Bayes

No.	X_1	X_2	X_3	X_4	X_5	X_6	Y	Prediksi SVM	Prediksi NB
1	49	0	162	110	0	1	1	1	1
2	36	0	108	81	0	0	-1	-1	-1
3	38	0	144	105	0	1	1	1	1
4	37	1	106	105	0	0	-1	-1	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
27	59	0	138	88	0	0	-1	1	1
28	33	0	168	103	0	0	1	1	1
29	37	0	120	72	0	0	-1	-1	-1
30	49	0	149	120	0	0	1	1	1

Pada Tabel 8, menunjukkan hasil klasifikasi menggunakan algoritma SVM terdapat hanya 1 data yang salah diklasifikasikan yaitu pada data ke-27, sedangkan dengan algoritma *naïve* Bayes terdapat 2 data yang salah diklasifikasikan yaitu data ke-4 dan data ke-27. Model klasifikasi yang telah dibuat kemudian dievaluasi menggunakan nilai akurasi yang dapat ditentukan dengan menggunakan Persamaan 10 yang direpresentasikan menggunakan matriks konfusi.

- ***Support Vector Machine***

Bentuk matriks konfusi dari hasil klasifikasi menggunakan algoritma SVM adalah

$$\begin{bmatrix} 20 & 0 \\ 1 & 9 \end{bmatrix}$$

sehingga nilai akurasi dari algoritma SVM adalah

$$\begin{aligned} \text{Akurasi} &= \frac{20 + 9}{20 + 0 + 1 + 9} \times 100\% \\ &= 0,9667 \times 100\% \\ &= 96,67\% \end{aligned}$$

- ***Naïve Bayes***

Bentuk matriks konfusi dari hasil klasifikasi menggunakan algoritma *naïve* Bayes adalah

$$\begin{bmatrix} 20 & 0 \\ 2 & 8 \end{bmatrix}$$

sehingga nilai akurasi dari algoritma *naïve* Bayes adalah

$$\begin{aligned} \text{Akurasi} &= \frac{20 + 8}{20 + 0 + 2 + 8} \times 100\% \\ &= 0,9333 \times 100\% \\ &= 93,33\% \end{aligned}$$

Berdasarkan kedua akurasi algoritma di atas, diketahui bahwa pada algoritma SVM menunjukkan akurasi dalam memprediksi klasifikasi sebesar 96,67% dan pada algoritma

naïve Bayes menunjukkan akurasi dalam memprediksi sebesar 93,33%. Hal ini menunjukkan bahwa algoritma SVM bekerja lebih baik dibandingkan dengan algoritma *naïve* Bayes dalam mengklasifikasikan status hipertensi rekam medis pasien Klinik Polresta Samarinda.

PENUTUP

Berdasarkan hasil analisis, diperoleh hasil klasifikasi penyakit tekanan darah tinggi pada pasien Klinik Polresta Samarinda Tahun 2022. Pada algoritma SVM memiliki akurasi sebesar 96,67% dan pada algoritma *naïve* Bayes memiliki akurasi sebesar 93,33%. Dengan demikian, algoritma SVM menunjukkan bahwa algoritma SVM memberikan ketepatan prediksi klasifikasi yang lebih baik pada kasus tersebut.

Pustaka

- [1] Mufarida, B. (2021). WHO Catat 66% Kematian di Indonesia Akibat Penyakit Tidak Menular. Dilansir dari <https://nasional.sindonews.com/read/324320/15/who-catat-66-kematian-di-indonesia-akibat-penyakit-tidak-menular>. Diakses pada tanggal 16 Januari 2023.
- [2] Rokom. (2019). Hipertensi Penyakit Paling Banyak Diidap Masyarakat. Dikutip dari <https://sehatnegeriku.kemkes.go.id/baca/umum/20190517/5130282/hipertensi-penyakit-paling-banyak-diidap-masyarakat>. Diakses pada tanggal 8 Januari 2023.
- [3] LeMone, P., Burke, K. M., dan Bauldoff, G. (2018). *Buku Ajar Keperawatan Medikal Bedah: Gangguan Kardiovaskular*. Jakarta: EGC.
- [4] Kemkes. (2018). Faktor Risiko Hipertensi. Dikutip dari <https://p2ptm.kemkes.go.id/infographic-p2ptm/hipertensi-penyakit-jantung-dan-pembuluh-darah/faktor-risiko-hipertensi>. Diakses pada tanggal 10 Februari 2023.
- [5] Nofriansyah, D., dan Nurcahyo, G.W. (2015). *Algoritma Data Mining Dan Pengujian*. Yogyakarta: Deepublish.
- [6] Gong, D. (2022). Top 6 Machine Learning Algorithms for Classification. Dikutip dari <https://towardsdatascience.com/top-machine-learning-algorithms-for-classification-2197870ff501>.
- [7] Mulaab. (2017). *Data Mining: Konsep dan Aplikasi*. Malang: Media Nusa Creative.
- [8] Desiani, A., Akbar, M., Irmeilyana, dan Amran, A. (2022). Implementasi Algoritma *Naïve* Bayes dan Support Vector Machine (SVM) Pada Klasifikasi Penyakit Kardiovaskular. *ELKOM*, 4(2). DOI: <https://doi.org/10.32528/elkom.v4i2.7691>.
- [9] Watomakin, DB. (2019). Comparison of Performance Support Vector Machine Algorithm and Naive Bayes for Diabetes Diagnosis. Proceedings of the International Conference on Science in Information Technology (ICSITech), Yogyakarta, 88-94.