ISSN: 2339-2541
JURNAL GAUSSIAN, Volume 13, Nomor 2, Tahun 2024, Halaman 319 - 327
Online di: https://ejournal3.undip.ac.id/index.php/gaussian/

ANALISIS KLASIFIKASI MENGGUNAKAN REGRESI LOGISTIK BINER DAN ALGORITMA *NAÏVE BAYES CLASSIFIER* PADA PENYAKIT HIPERTENSI

Riza Sahila[1*], Tatik Widiharih[2], Iut Tri Utami[3]
[1,2,3]Departemen Statistika, Fakultas Sains dan Matematika, Universitas Diponegoro
***e-mail** : rizasahila@gmail.com

DOI: 10.14710/j.gauss.13.2.319-327

Article Info:
Received: 2023-06-07
Accepted: 2024-11-14
Available Online: 2024-11-15

Keywords:
Hypertension; Classification; Binary Logistic Regression; Naïve Bayes Classifier; Sensitivity.

Abstract: Hypertension is a primary cause of cardiovascular disease. Approximately 60% of people with hypertension are in developing countries, including Indonesia. In this analytical study, classification will be carried out to prove the status of hypertensive patients or not hypertensive. The classification method used is Binary Logistic Regression and Naïve Bayes Classifier. Binary Logistic Regression is Logistic Regression with the response variable being binary. Naïve Bayes Classifier namely predicting future opportunities using previous data. The factors used in this study were gender, age, height, and weight. The greatest accuracy of classification results is in the proportion of 90%:10%. The accuracy of the classification produced by Binary Logistic Regression method resulted in a sensitivity of 93.33%. The classification accuracy obtained by the Naïve Bayes Classifier with a sensitivity of 63.64%. This shows that the Binary Logistic Regression method has a better sensitivity value.

1. PENDAHULUAN

Hipertensi menjadi masalah kesehatan di segala belahan dunia dan menjadi sebagai salah satu faktor utama terjadinya penyakit kardiovaskular. Sebanyak kurang lebih 60% pengidap hipertensi terletak di negara yang berkembang salah satunya negara Indonesia. Menurut Jain (2011) faktor-faktor resiko terjadinya hipertesi antara lain faktor keluarga, jenis kelamin, usia, berat badan, kebiasaan merokok stress, kurang aktivitas seperti olahraga serta konsumsi obat-obatan.

Penelitian mengenai klasifikasi status pasien penyakit hipertensi ini dilakukan guna mengetahui bagaimana mengklasifikasikan status pasien hipertensi tersebut, agar bisa diketahui secara detail yakni dengan melihat nilai sensitivitas yang baik. Klasifikasi adalah salah satu dari tata cara statistika yang mengkategorikan informasi yang cocok dengan cirinya secara teratur di dalam kelas yang sudah ditetapkan.

Metode pengklasifikasian yang pertama menggunakan Regresi Logistik Biner dan metode ke dua menggunakan Algoritma *Naïve Bayes Classifier*. Analisis regresi yang variabel dependenya bersifat dikontomus adalah Regresi Logistik Biner. Dikontomus merupakan variabel yang hanya memiliki dua kemungkinan nilai, yakni 1 menunjukan adanya ciri serta 0 tidak ada ciri. Algoritma *Naïve Bayes Classifier* memanfaatkan probabilitas sederhana dengan menghitung probabilitas kelas dan probabilitas bersyarat menggunakan data latih untuk kemudian digunakan dalam mengklasifikasi pengamatan baru. Penelitian ini diharapkan dapat mengklasifikasikan status pasien hipertensi dengan setepat mungkin dan menilai model dengan sensitivitas tinggi.

2. TİNJAUAN PUSTAKA

Metode yang menghubungkan variabel independen dengan variabel dependennya yang bersifat kategorik. Menurut Agresti (2007), pada pemodelan statistik variabel respon. Bentuk model regresi logistiknya dapat dilihat pada persamaan berikut:

$$\pi(x_i) = \frac{\exp(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi})}{1 + \exp(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi})} \quad (1)$$

dengan fungsi logitnya yaitu:

$$g(x_i) = \ln\left[\frac{\pi(x_i)}{1-\pi(x_i)}\right] = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} \quad (2)$$

Menurut Agresti (2007), metode penduga maksimum likelihood dapat digunakan untuk mengestimasi parameter dalam logistik biner. Misalnya ada sampeI dari n pengamatan independen $(x_i, y_i), 1,2, \ldots, n$ beserta y_i

Dilambangkan mulai variabeI respon biner serta x_i yakni nilai dari variabel prediktor untuk subjek ke-i, maka fungsi likelihoodnya merupakan perkalian dari setiap fungsi densitasnya.

Uji Rasio LikeIihood bertujuan untuk menentukan signifikansi koefisien β terhadap variabeI respon secara serentak.

Hipotesis:

$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$ (semua variabeI prediktor tidak ada yang mempengaruhi variabel respon)

H_1: Paling sedikit hanya ada satu $\beta_j \neq 0$ dengan j = 1,2,…, p (paling sedikit ada satu variabel prediktor yang mempengaruhi variabel respon)

Statistik Uji:

$$G = -2\ln\left[\frac{likelihood\ tanpa\ variabel\ independen}{likelihood\ dengan\ variabel\ independen}\right] \quad (3)$$

Dengan taraf signifikansi α, H_0 ditoIak jika $G > \chi^2_{(\alpha,p)}$ atau niIai *p-value* < α

Tujuan dari uji Wald untuk mengetahui pengaruh masing-masing koefisien β_j secara individual yang dibandingkan dengan standar *errornya* untuk menentukan apakah variabel prediktor dalam model memiliki pengaruh signifikan terhadap variabel respon.

Hipotesis:

H_0: β_j = 0 dengan j = 1,2,..., p (Tidak ada pengaruh antara variabel independen ke-j dengan variabel prediktor)

H_1: $\beta_j \neq 0$ dengan j = 1,2,...,p (Ada pengaruh antara variabel independen ke-j dengan variabel prediktor)

Statistik Uji:

$$W_j = \left\{\frac{\hat{\beta}_J}{SE\left(\hat{\beta}_J\right)}\right\} \quad (4)$$

Dengan taraf signifikansi α, H_0ditolak jika $|W_j| > Z_{(0.05/2)}$ atau nilai *p-value* > α

Mengemukakan tujuan dari uji kecocokan model itu untuk melihat apakah model sesuai atau tidak sesuai.

Hipotesis:

H_0: Model sesuai (Tidak ada perbedaan antara prediksi dengan hasil observasi)

H_1: Model tidak sesuai (Ada perbedaan antara prediksi dengan hasil observasi)

Statistik Uji:

$$\hat{C} = \sum_{k=1}^{g} \frac{(O_k - n'_k\hat{\pi}_k)}{(n'_k\hat{\pi}_k)(1-\hat{\pi}_k)} \quad (5)$$

Dengan taraf signifikansi α, menolak H_0 apabila $\hat{C} > \chi^2_{(\alpha,g-2)}$ atau nilai *p-value* < α .

Variabel dalam penelitian ini menggunakan satuan ukuran yang berbeda, maka harus dilakukan normalisasi. Normalisasi ini menggunakan normalisasi min-max. Data diubah antara 0 hingga 1 supaya data menjadi seimbang. Berikut persamaan normalisasi min-max.

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \qquad (6)$$

Naïve Bayes Classifier adalah suatu algoritma yang termasuk dalam metode klasifikasi. Pengklasifikasian *Naïve Bayes Classifier* mengansumsikan bahwasanya terdapat atau tidaknya fitur terpilih sejak sebuah kategori tidak terkait dengan fitur kategori lainnya.

$$P(A|B) = \frac{P(B|A).P(A)}{P(B)} \qquad (7)$$

Keterangan:
A : SampeI data yang labelnya tidak diketahui.
B : KeIas hasiI kIasifikasi
$P(A|B)$: ProbabiIitas terjadinya A jika B diketahui
$P(A)$: ProbabiIitas *prior* A

Dimisalkan bahwa kumpulan data yang berisi n kasus x_i, I =1,2,…,n, yang terbentuk dari atribut p, yaitu $x_i = (x_{1i,}x_{2i}, \dots, x_{pi})$. Diasumsikan bahwa masing-masing kasus milik satu kelas $y \in \{y_{1,}y_2, \dots, y_c\}$. Pengelompokan *Naïve Bayes Classifier* sederhana menggunakan peluang ini untuk menetapkan sebuah kasus ke kelas. Menerapkan teorema bayes (persamaan 7) dan menyederhanakan notasi, maka didapatkan persamaan:

$$P(y_j|x_i) = \frac{P(x_i|y_j)P(y_j)}{P(x_i)} \qquad (8)$$

Perhatikan bahwa pembilang pada suatu persamaan 8 merupakan probabilitas gabungan dari x_i dan y_j. x hanya akan digunakan dengan menghilangkan indeks i untuk disederhanakan, berikut ini pembilang yang dapat ditulis ulang yaitu:

$$P(x|y_j)P(y_j) = P(y_j|x)P(x)$$

$$P(x|y_j)P(y_j) = \frac{P(y_j \cap x)}{P(x)}P(x)$$

$$P(x|y_j)P(y_j) = P(x, y_j)$$

$$= P(x_{1,}x_2, \dots, x_{p,}y_j)$$

Karena $P(a,b) = P(a|b)P(b)$, maka

$$= P(x_1|x_2, \dots, x_{p,}y_j)P(x_2, \dots, x_{p,}y_j)$$

$$= P(x_1|x_2, x_{3,} \dots, x_{p,}y_j)\, P(x_2|x_3, x_{4,} \dots, x_{p,}y_j)\, P(x_3, x_{4,} \dots, x_{p,}y_j)$$

$$= P(x_1|x_2, x_{3,} \dots, x_{p,}y_j)\, P(x_2|x_3, x_{4,} \dots, x_{p,}y_j) \dots P(x_p|y_j)P(y_j)$$

Diasumsikan bahwa x_i tidak bisa bergantung antara satu dengan yang lainnya. Inilah yang memperbedakan *Bayesian* dengan *Naïve Bayes Classifier*. Asumsi ini menyiratkan bahwa $P(x_1|x_2,\dots,x_p,y_j) = P(x_1|y_j)$. Jadi probabilitas gabungan dari x dan y_j adalah:

$$P(x|y_j)P(y_j) = P(x_1|y_j).P(x_2|y_j)\dots P(x_p|y_j).P(y_j)$$
$$= \prod_{k=1}^{P} P(x_k|y_j).P(y_j) \quad (9)$$

Jika persamaan 9 dimasukkan ke suatu persamaan 10 sehingga didapatkan:

$$P(y_j|x) = \frac{\prod_{k=1}^{P} P(x_k|y_j).P(y_j)}{P(x)} \quad (10)$$

Keterangan:
$P(y_j|x)$: Peluang menggunakan *vector* x pada variabel y
$P(y_j)$: Peluang kemunculan
$\prod_{k=1}^{P} P(x_k|y_j)$: ProbabiIitas prediktor kelas y dari semua nilai *vector* x
$P(x)$: Peluang dari x, tidak terikat pada kelasnya

Penetuan kategori yang sesuai untuk sampel yang dilakukan dengan cara mempertimbangkan suatu nilai *posterior* tiap kategori dan memilih kategori dengan nilai *posterior* yang terbaik. Kategori terbaik dalam mengklasifikasikan *Naïve Bayes Classifier* ditentukan dengan menghitung *Maximum a Posterior* (MAP) kategori C_{map} pada persamaan 11.

$$C_{map} = argmax_{y \in Y} P(y) \prod_{k=1}^{n} P(x_k|y) \quad (11)$$

Keterangan:
$P(y)$: Probabilitas *prior* dari kelas j
$P(x_k|y)$: Probabilitas kata $x_k$untuk kelas j

Fungsi yang disebut $argmax$ digunakan hanya bertujuan mengambil nilai yang paling besar pada sub kelas y dari kelas Y. Berdasarkan persamaan 11 tanpa memasukan niIai $P(x)$, faktor ini dapat diabaikan karena memiliki nilai yang positif dan berlaku untuk setiap keIas, maka tidak dapat mempengaruhi perbandingan nilai *posterior*. Metode *Naïve Bayes Classifier* bisa digunakan ketika sebelumnya telah tersedia data yang dijadikan landasan untuk melakukan suatu klasifikasi.

Menurut Han dan Kamber (2006), nilai posterior dihitung berbeda ketika atribut memiliki nilai kontinu. Atribut dengan nilai yang kontinu diasumsikan memiliki distribusi Gaussian menggunakan *mean* μ serta standar deviasi σ. maka didapat persamaan 12:

$$P(x_k|C_i) = \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{(x_k-\mu_{ci,})^2}{2\sigma_{c_i^2}}} \quad (12)$$

Keterangan:
$P(x_k|C_i)$: peluang x_k bersyarat C_i
x_k : NiIai data k yang akan dikIasifikasi
K : data yang dikIasifikasi;
C_i : sub keIas C yang dicari ($C_{hipertensi}$ dan $C_{tidak\ hipertensi}$)
i : kelas yang diklasifikasikan (hipertensi dan tidak hipertensi)
$\mu_{ci,}$: Menyatakan *mean* simbol C_i

σ_{ci} : Standar deviasi atribut simbol keIas C_i

Keakuratan model algoritma yang dibuat ditentukan dengan bantuan evaluasi. Kriteria evaluasi yang digunakan yakni akurasi, sensitivitas, dan spesifisitas. *Mengukur tingkat kebenaran pada proses klasifikasi dapat dihitung menggunakan Confution matrix.*

Tabel 1. *Confution Matrix*

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

Proporsi data yang terprediksi dengan benar disebut akurasi. Akurasi adalah ukuran ketelitian model dalam memprediksi data dibandingan dengan data aktualnya dan digunakan sebagai ukuran model untuk menentukan keakurat dalam melakukan prediksi. Persamaan akurasi seperti pada persamaan berikut:

$$Accuracy = \frac{TP + TN}{Total} \times 100\% \quad (13)$$

Sensitivitas adalah bagian dari data positif yang di prediksi dengan tepat sebagai data positif. Perhitungan ini digunakan untuk menilai sebagian besar kesuksesan suatu model dalam memprediksi kelas positif yang diklasifikasikan. Persamaan ditunjukkan pada persamaan berikut.

$$Sensitivity = \frac{TP}{TP + FN} \quad (14)$$

Spesifisitas adalah bagian data negatif yang terprediksi dengan tepat sebagai data negatif. Spesifisitas digunakan untuk menilai sebagian besar kesuksesan model dalam memprediksi kelas negatif yang diklasifikasi. Persamaan Spesitifitas ditunjukkan pada persamaan berikut.

$$Specificity = \frac{TN}{TN - FP} \quad (15)$$

3. METODE PENELITIAN

Tipe data yang diaplikasikan yakni data sekunder yang diperoleh dari pasien Di Puskesmas Lamper Tengah Semarang Tahun 2022. Variabel yang digunakan terdiri dari status pasien (Y), Jenis Kelamin (x_1), Usia (x_2), Tinggi Badan (x_3), dan Berat Badan (x_4).

Software yang akan digunakan untuk mengolah data menggunakan program R Studio 4.2.2 dan *Microsoft Excel* 2019. Tahapan analisis yang dilakukan yaitu:

1. Memisahkan data menjadi data latih dan data uji. Memperbandingkan data latih dan data uji yang terpilih yakni 90%: 10%.
2. Menetapkan klasifikasi yang mempengaruhi penyakit hipertensi menggunakan metode Regresi Logistik Biner
   a. Membuat model Regresi Logistik Biner.
   b. Membuat uji serentak menggunakan uji Rasio Likelihood.
   c. Membuat uji parsial menggunakan uji Wald.
   d. Membuat uji Hosmer dan Lemeshow.
   e. Menetapkan model akhir Regresi Logistik Biner.
   f. Menghitung nilai $\pi(x_i)$, $\pi(x_i)$ yakni peluang pasien hipertensi. Apabila nilai $\pi(x_i) <$ 0,5 maka masuk kedalam kelas 0 dan dikategorikan pasien Tidak Hipertensi. Apabila nilai $\pi(x_i) \geq 0{,}5$ maka masuk kedalam kelas 1 dan dikategorikan pasien Hipertensi.
   g. Membentuk *confution matrix* dan menghitung ketepatan klasifikasinya menggunakan *accuracy*, *sensitivity*, dan *specificity*.
3. Melakukan klasifikasi menggunakan *Naïve Bayes Classifier*.
   a. Melakukkan probabilitas likelihood setiap variabel setiap kelas.

b. Menghitung probabilitas prior setiap kelas.
c. Menghitung probabilitas posterior masing-masing kelas
d. Membandingkan nilai posterior masing-masing kelas. Apabila nilai *posterior* tertinggi pada kategori Hipertensi maka status pasien diklasifikasikan ke dalam Hipertensi, dinotasikan 1. Apabila nilai *posterior* tertinggi pada kategori tidak hipertensi maka status pasien diklasifikasikan tidak hipertensi, dinotasikan 0.
e. Membentuk *confution matrix* dan menghitung ketepatan klasifikasinya menggunakan *accuracy*, *sensitivity*, dan *specificity*.

4. Memperbandingkan ketepatan klasifikasi kedua metode Regresi Logistik Biner dan *Naïve Bayes Classifier* menggunakan uji beda dua proporsi.

4. HASIL DAN PEMBAHASAN

Data yang diaplikasikan sejumlah 532 pasien, terdiri dari 351 pasien atau sebesar 59,21% pasien hipertensi dan 217 pasien atau sebesar 40,79% pasien tidak hipertensi. Sebelumnya dilakukan pemisahan data membentuk dua bagian, yang dibagi menjadi data latih dan data uji yang telah dilakukan beberapa kali percobaan. Proporsi 75%:25%. Proporsi 80%:20%, dan Proporsi 90%:10%. Model awal Regresi Logistik Biner dibentuk dengan menggunakan estimasi parameter. Dibentuk tabel hasil Estimasi Parameter dan Uji Wald sebagai berikut:

Tabel 2. Hasil Estimasi Parameter dan Uji Wald Model Awal

Variabel	Estimasi	Std. Error	W_j	p-value	Keterangan
X1(2)	0,546166	0,249923	2,185	0,02886	Signifikan
X2	0,035958	0,008700	4,133	3,58e-05	Signifikan
X3	0,013006	0,018383	0,707	0,47928	Tidak signifikan
X4	0,023044	0,008886	2,593	0,00951	Signifikan
constant	-5,524483	3,002130	-1,840	0,06574	-

Model awal Regresi Logistik Biner yang terbentuk yaitu

$$\pi(x_i) = \frac{e^{g(x_i)}}{1+e^{g(x_i)}}$$

sebagai peluang penyakit hipertensi dengan

$g(x_i)$= -5,524483 + 0,546166 (JKp) + 0,035958 (Usia) + 0,013006 (TB) + 0,023044 (BB).

Hipotesis Uji Rasio Likelihood:

H_0: $\beta_1 = \beta_2 = \cdots = \beta_4$ (Seluruh variabel prediktor tidak ada yang berpengaruh terhadap variabel status penyakit hipertensi)

H_1: Paling sedikit ada satu $\beta_j \neq 0$ dengan j = 1,2,...,4 (Paling sedikit ada satu variabel prediktor yang berpengaruh terhadap variabel status pasien penyakit hipertensi)

Statistik Uji:

$$G = -2\ln\left[\frac{likelihood\ tanpa\ variabel\ independen}{likelihood\ dengan\ variabel\ independen}\right] = 25{,}83521207$$

Pada taraf signifikansi α = 5%, Ho ditolak karena G = 25,83521207 > χ^2(4,0,05) = 9,487729. Disimpulkan bahwa paling sedikit ada satu variabel prediktor yang berpengaruh terhadap variabel status pasien penyakit hipertensi.

Hipotesis Uji Wald:

H_0: β_j = 0 dengan j = 1,2,...,4 (Tidak ada pengaruh antara variabel prediktor ke-j dengan variabel respon)

$H_1: \beta_j \neq 0$ dengan j = 1,2,...,4 (Ada pengaruh antara variabel prediktor ke-j dengan variabel respon)

Statistik Uji:

$$W_j = \left\{ \frac{\hat{\beta}_J}{SE\left(\hat{\beta}_J\right)} \right\}$$

Dengan taraf signifikan α = 5% H_0 ditolak jika $|W_j| > Z_{\propto/2}$ = 1,96 atau H_0 ditolak jika nilai *p-value* < α, variabel prediktor yang berpengaruh secara signifikan terhadap status pasien penyakit hipertensi yakni Jenis Kelamin (JK) $(x_1)(\text{perempuan})$, Usia (x_2), dan Berat Badan (x_4).

Model kedua Regresi Logistik Biner yang terbentuk yaitu

$$\pi(x_i) = \frac{e^{g(x_i)}}{1+e^{g(x_i)}} \text{ dengan}$$

$g(x_1)$ = -3,494103 + 0,454649(JKp)+ 0,034837(Usia) + 0,025100(BB).

Ketepatan hasil klasifikasi dihitung dengan menggunakan *Confution matrix*. Ketepatan hasil klasifikasi yang digunakan adalah akurasi, sensitivitas, dan spesifisitas. Berikut adalah hasil perhitungan akurasi, sensitivitas, dan spesifisitas setiap proporsi yang dapat dilihat pada Tabel 3. Hasil ketepatan klasifikasi pada setiap proporsi diolah menggunakan R *Studio*.

Tabel 3. Ukuran Kebaikan Model Regresi Logistik Biner

Nilai Sebenarnya	Prediksi		
	75% : 25%	80% : 20%	90% : 10%
Accuracy	65,44%	66,05%	75%
Sensitivity	88,89%	86,36%	93,33%
Spesificity	30,91%	34,88%	50%

Hasil perhitungan pada proporsi 90%:10% menjadi proporsi yang terbaik dibandingkan proporsi lainnya. Nilai Sensitivitas menunjukkan bahwa proporsi keberhasilan model dalam mengklasifikasikan pasien ke dalam kelas hipertensi sebesar 93,33%.

Metode *Naïve Bayes CIassifier* adalah suatu algoritma yang berada pada data mining, dalam memproses data mining harus melakukan proses normalisasi yang bertujuan agar data berada pada label yang tepat. Merubah data menjadi 0 hingga 1 agar data seimbang. Perhitungan normalisasi min max didapatkan dari hasil *output* menggunakan *software* R *Studio* sebagai berikut:

Tabel 4. Data *Min Max Normalization*

Pasien	X_1	X_2	X_3	X_4	**Y**
1	2	0.6707317	0.8113208	0.6235294	0
2	1	0.7682927	0.6226415	0.6000000	1
3	2	0.5487805	0.5471698	0.4117647	1
.	.	.	.	.	.
.	.	.	.	.	.
531	2	0.7560976	0.5849057	0.2588235	0
532	2	0.6585366	0.5660377	0.3176471	0

Penentuan klasifikasi menggunakan metode *Naïve Bayes Classifier* yakni dengan membandingkan nilai *posterior* masing-masing variabel tersebut. Jika nilai *posterior* tertinggi pada kategori hipertensi maka status pasien diklasifikasikan ke dalam hipertensi dinotasikan 1. Begitupun sebaliknya, jika nilai *posterior* tertinggi pada kategori tidak hipertensi maka status pasien diklasifikasikan tidak hipertensi, maka dinotasikan 0.

Tujuan pengujian ini yaitu untuk mengidentifikasikan hasil klasifikasi, menilai kinerja dari *Naïve Bayes Classifier* dalam mengklasifikasikan data ke kelas yang telah ditentukan.

Tabel 5. Hasil Klasifikasi *Naïve Bayes Classifier*

ID	JK	Usia	TB	BB	Status Aktual	Status Prediksi
15	P	0.5731707	0.3773585	0.5176471	Hipertensi	Hipertensi
20	L	0.7073171	0.7735849	0.5294118	Hipertensi	Hipertensi
22	P	0.5853659	0.4716981	0.5058824	Hipertensi	Hipertensi
26	P	0.9146341	0.3207547	0.1294118	Hipertensi	Tidak Hipertensi
.	.	.	.	.	.	.
.	.	.	.	.	.	.
505	P	0.6829268	0.6415094	0.4470588	Hipertensi	Hipertensi
510	P	0.7195122	0.5849057	0.3294118	Hipertensi	Hipertensi

Dari perhitungan klasifikasi pada tabel 3 terlihat sejumlah 53 pasien yang diklasifikasikan dengan benar sebanyak 38 dan sejumlah 15 pasien salah diklasifikasikan. Ketepatan hasil klasifikasi dihitung dengan menggunakan *Confution matrix*. Ketepatan hasil klasifikasi yang digunakan adalah akurasi, sensitivitas, dan spesifisitas. Hasil ketepatan klasifikasi pada setiap proporsi diolah menggunakan R *Studio*.

Tabel 6. Ukuran Kebaikan Model *Naïve Bayes Classifier*

Ukuran Hasil Kebaikan	Proporsi		
	75% : 25%	80% : 20%	90% : 10%
Accuracy	64,71%	65,14%	73,08%
Sensitivity	82,72%	80,30%	80%
Specificity	38,18%	41,86%	63,64%

Berdasarkan Tabel 6 menunjukkan bahwa pada proporsi 90% : 10% menjadi proporsi terbaik dibandingkan proporsi yang lain. Hasil perhitungan pada proporsi 90%:10% nilai Sensitivitas menunjukkan bahwa proporsi keberhasilan model dalam mengklasifikasikan pasien ke dalam kelas hipertensi sebesar 63,64%.

Ketepatan hasil klasifikasi dari kedua metode telah diketahui pada masing-masing proporsi. Langkah selanjutnya membandingkan hasil ketepatan klasifikasi, dilihat pada Tabel 3 dan Tabel 6. Tabel 3 dan Tabel 6 menunjukkan bahwa yang memiliki ketepatan hasil klasifikasi terbesar yaitu pada proporsi 90%:10%. Kasus diagnosis penyakit pasien hipertensi memperoleh nilai sensitivitas tinggi, nilai yang tinggi pada sensitivitas diartikan sistem secara handal mengenali semua data yang terkena hipertensi. Hasil ketepatan klasifikasi metode Regresi Logistik Biner menghasilkan sensitivitas 93,33% diartikan sebagai proporsi keberasilan model dalam mengklasifikasikan pasien ke dalam kelas hipertensi sebesar 93,33%. Ketepatan klasifikasi yang diperoleh *Naïve Bayes Classifier* dengan sensitivitas sebesar 63,64% yang artinya proporsi keberasilan model dalam mengklasifikasikan pasien ke dalam kelas hipertensi sebesar 63,64%. Hal ini menujukkan bahwa metode Regresi Logistik Biner memiliki sensitivitas lebih tinggi.

5. KESIMPULAN

Hasil ketepatan klasifikasi menunjukkan bahwa metode Regresi Logistik Biner menghasilkan nilai sensitivitas sebesar 93,33% yang artinya proporsi keberasilan model dalam mengklasifikasikan pasien ke dalam kelas hipertensi sebesar 93,33%. Ketepatan klasifikasi yang dihasilkan *Naïve Bayes Classifier* dengan nilai sensitivitas sebesar 63,64% yang artinya proporsi keberasilan model dalam mengklasifikasikan pasien ke dalam kelas hipertensi sebesar 63,64%. Hal ini menujukkan bahwa metode Regresi Logistik Biner memiliki nilai sensitivitas yang lebih baik.

DAFTAR PUSTAKA

Agresti, A. (2007). *An Introduction to Categorical Data Analysis Second Edition*. New Jersey: John Wiley and Sons.

Han, J., dan Kamber, M. (2006). *Data Mining: Concepts and Techniquess,*2^{th}*Annual Internasional ACM SIGIR Conference on Research and Development in Information Retrieval*, 97-104.

Hosmer, D. W, dan Lemeshow S. (2000). *Applied Logistic Regression*. United States of American: Sons Inc.

Jain, R. 2011. *Pengobatan Alternatif untuk Mengatasi Tekanan Darah*. Jakarta: PT. Gramedia Pustaka Utama.

Kowalski, R, E. 2007. *Terapi Hipertensi*. Bandung: Qanita.

Manjoer, A. 2001. *Kapita Selekta Kedokteran Edisi Ketiga*. Jakarta: Media Aesculapius Fakultas Kedokteran Universitas Indonesia.

Prasetyo, E. 2012. *Data Mining Konsep Dan Aplikasi Menggunakan Matlab*. Yogyakarta: Andi.

Prasetyo, E. 2014. *Data Miing Mengolah Data Menjadi Informasi Menggunakan Matlab*. Yogyakarta: Andi Offset.

UPT Puskesmas Lamper Tengah. 2022. *Profil Kesehatan UPT Puskesmas Lamper Tengah Dinas Kesehatan Kota Semarang Tahun 2022 (Data Tahun 2022)*. Semarang: UPT Puskesmas Lamper Tengah.

Utami, L. A. 2017. Analisis Sentimen Opini Publik Berita Kebakaran Hutan Melalui Komparasi Algoritma Support Vector Machine Dan K-Nearest Neighbor Berbasis Particle Swarm Optimization. *Jurnal Pilar Nusa Mandiri*, 13(1), 103-112.

Virmani, D., Taneja, S. dan Malhotra, G., 2015. Normalization based K means Clustering Algorithm. *International Journal of Advanced Engineering Research and Science (IJAERS)*, 2(2), pp. 36-40.

Widjaja. 2009. *Hubungan Keluarga Dengan Tingkat Kepatuhan Diet Rendah Garam*. Jakarta.

Widodo, P.P., Handayanto, R. T., dan Herlawati, H. 2013. *Penerapan Data Mining Dengan Matlab*. Bandung: Rekayasa Sains.

World Health Organization (WHO). 2018. *Global Health Estimates 2016: Deaths by Cause, Age, Sex, by country and by Region*, 2000-2016. Geneva: World Health Organization.