PERBANDINGAN KLASIFIKASI STATUS PENDONOR DARAH DENGAN MENGGUNAKAN REGRESI LOGISTIK DAN K-NEAREST NEIGHBOR

Iut Tri Utami[1], Fadjryani[2], Diah Daniaty[3]

Program Studi Statistika Jurusan Matematika FMIPA Universitas Tadulako[1,2], Badan Pusat Statistik[3]
e-mail: [1]triutami.iut@gmail.com, [2]fadjryani_mipauntad@yahoo.com, [3]diahdaniaty@gmail.com

Abstrak

Donor Darah Sukarela (DDS) adalah orang yang dengan sukarela mentranfusikan darahnya kepada orang lain. Seseorang dapat menjadi pendonor darah jika memenuhi kriteria dari PMI dan lolos dalam pemeriksaan dokter. Syarat yang diberlakukan PMI menyebabkan calon pendonor darah dapat diklasifikasikan menjadi layak dan tidak layak dalam mendonorkan darahnya. Salah satu cara untuk menentukan pola prediksi status kelayakan calon pendonor darah di PMI adalah menggunakan regresi logistik biner dan k-Nearest Neighbor (kNN). Peubah yang signifikan mempengaruhi kelayakan calon pendonor darah adalah kadar Haemoglobin. Akurasi yang dihasilkan oleh metode regresi logistik biner dan kNN pada penelitian ini adalah 93% dan 79%.

Kata kunci: DDS, regresi logistik biner, k-Nearest Neighbor

Abstract

Voluntary Blood Donors are people who voluntarily transfer their blood to others. A person can become a blood donor if he meets the criteria of PMI and passes the doctor's examination The requirements imposed by PMI can cause prospective blood donors to be classified as feasible and improper in donating blood. One way to determine the pattern of predicting the eligibility status of prospective blood donors at PMI is to use k-Nearest Neighbor (kNN) and binary logistic regression. The significant variable influencing the eligibility of prospective blood donors is hemoglobin levels. The accuracy produced by the binary logistic regression and kNN method in this study was 93% and 79%.
.

***Keywords**: Voluntary blood donors, binary logistic regression, k-Nearest Neighbor*

PENDAHULUAN

Latar Belakang

Donor darah mempunyai manfaat yang baik bagi kesehatan tubuh. Sayangnya, banyak orang takut donor darah dengan beragam alasan mulai dari takut jarum suntik, lemas dan kehabisan darah. Manfaat menyumbangkan darah tidak hanya dirasakan bagi penerima darah tetapi juga bagi pendonor darah. Ada berbagai hal positif yang bisa didapatkan ketika mendonorkan darah, mulai dari membantu membakar kalori, menjaga kesehatan jantung, meningkatkan produksi sel darah, hingga menurunkan risiko kanker. Sayangnya, jumlah pendonor darah masih belum banyak dan belum memenuhi target kebutuhan darah nasional.

Demi kesehatan tubuh pendonor dan penerima darah, tidak semua orang dapat mendonorkan darahnya. Beberapa hal yang disyaratkan PMI saat mau mendonorkan darahnya yaitu umur 17-60 tahun, berat minimal 45 kg, temperatur tubuh 36,6°C-37,5°C yang diukur secara oral, tekanan darah baik (Sistole = 110-160 mm Hg dan Diastole = 70-100 mm Hg), denyut nadi teratur 50-100 kali per menit, dan Hemoglobin wanita minimal = 12 gr % sedangkan pria minimal = 12,5 gr % (Depkes RI, 2009). Syarat yang diberlakukan PMI menyebabkan calon pendonor darah dapat diklasifikasikan menjadi layak dan tidak layak dalam mendonorkan darahnya (PMI,2009). Salah satu cara untuk menentukan pola prediksi calon pendonor darah di PMI adalah dengan menggunakan regresi logistik dan *k-Nearest Neighbor* (kNN).

Hosmer dan Lemeshow (2000) menjelaskan bahwa metode regresi logistik adalah suatu metode analisis statistika yang menganalisis hubungan antara peubah respon yang memiliki dua kategori atau lebih dengan satu atau lebih peubah penjelas. Salah satu model regresi logistik adalah regresi logistik biner. Metode regresi logistik biner juga dapat digunakan untuk menganalisis nilai ketepatan klasifikasi.

kNN merupakan suatu metode mengelompokkan suatu objek dengan mempertimbangkan kelas terdekat dari objek tersebut. kNN merupakan metode klasifikasi yang sangat sederhana, efisien dan efektif dalam bidang pengenalan pola, kategori teks, pengolahan objek dan mampu melakukan training data dalam jumlah yang besar (Bathia,2010). Meskipun sederhana, kNN dianggap menjadi salah satu dari sepuluh algoritma klasifikasi data mining yang terbaik (Wu et al, 2008). Salah satu masalah dari algoritma ini adalah efek yang sama terjadi dari semua atribut yang terdapat pada data baru dan data lama dalam data set pelatihan (Moradian dan Baraani, 2009).

Penelitian terdahulu tentang pengklasifikasian calon pendonor darah telah banyak dilakukan. Nugroho, dkk (2018) melakukan penelitian tentang klasifikasi pendonor darah menggunakan metode Support Vector Machine (SVM) pada dataset RFMTC yang menghasilkan akurasi sebesar 72.64%. Penelitian lain dilakukan Bayususetyo,dkk (2017) dengan mengklasifikasi calon pendonor darah dengan menggunakan metode Naive Bayes *Classifier* dengan studi kasus PMI di Kota Semarang. Sapriana (2017) menerapkan algoritma Naive Bayes *Classifier* pada klasifikasi status kelayakan pendonor darah di Unit Transfusi Darah Palang Merah Indonesia Kota Makassar dengan akurasi yang dihasilkan sebesar 90%.

Pada penelitian ini akan mengkaji ketepatan klasifikasi dan faktor-faktor yang mempengaruhi status kelayakan calon pendonor darah dengan menggunakan metode regresi logistik biner dan kNN. Perbandingan kedua metode dapat dilihat dari nilai akurasi nya. Semakin tinggi akurasi yang dihasilkan maka model semakin tepat dalam mengklasifikasikan. Adapun peubah penjelas yang akan digunakan adalah umur, jenis kelamin, berat badan, kadar HB, dan tekanan darah.

METODOLOGI

Tinjauan Referensi

Status Kelayakan Calon Pendonor Darah

Donor darah tidak hanya menguntungkan bagi penerima darah, tapi juga bermanfaat untuk pendonor. Tidak semua orang bisa mendonorkan darahnya, ada beberapa syarat donor darah yang perlu diketahui.

Syarat donor darah yang paling utama adalah kondisi fisik harus sehat. Usia juga menjadi syarat donor darah yaitu 17-60 tahun. Namun, untuk remaja usia 17 tahun diperbolehkan menjadi donor darah apabila mendapat izin tertulis dari orangtua. Calon pendonor baru dikatakan layak jika lolos pemeriksaan kesehatan sebelum mendonorkan darah.

Pendonor harus memiliki berat badan minimal 45 kilogram dan dalam kondisi sehat, baik jasmani maupun rohani. Selain itu, pendonor harus memiliki suhu tubuh 36, 6oC - 37,5oC. Tekanan darah pendonor harus berada pada angka 100-160 untuk *sistole* dan 70-100 untuk *diastole*. Denyut nadi saat pemeriksaan juga harus sekitar 50-100 kali per menit. Sementara itu, kadar haemoglobin pendonor harus minimal 12 gr/dL untuk wanita, dan minimal 12,5 gr/dL untuk pria (PMI, 2019).

Pendonor dapat mendonorkan darahnya paling banyak lima kali setahun dengan jangka waktu sekurang-kurangnya tiga bulan. Calon pendonor darah menjalani pemeriksaan pendahuluan, seperti kondisi berat badan, kadar HB, golongan darah, dan dilanjutkan dengan pemeriksaan dokter.

1. Sumber Data

Data yang akan digunakan pada penelitian ini adalah data sekunder yang diambil dari Unit Donor Darah (UDD) PMI Kabupaten Sigi, Provinsi Sulawesi Tengah. Jumlah data sebanyak 101 orang. Data akan dibagi menjadi dua yaitu data *training* dan data *testing*. Data *training* yang digunakan pada penelitian ini adalah 70% dari data keseluruhan, sisanya sebagai data *testing*. *Software* yang digunakan untuk mengolah data pada penelitian ini adalah R dengan paket *Class*, *Caret* dan *MASS*. Peubah penelitian yang akan digunakan pada penelitian ini adalah :

Tabel 1. Data Atribut

No	Atribut	Keterangan
1	Status Donor (Y)	1 = Layak Donor 2 = Tidak Layak
2	Jenis Kelamin (X_1)	1 = Laki-Laki 2 = Perempuan
3	Usia (X_2)	Numerikal
4	Berat Badan (X_3)	Numerikal
5	Tinggi Badan (X_4)	Numerikal
6	*Sistole* (X_5)	Tekanan darah atas
7	*Diastole* (X_6)	Tekanan darah bawah
8	Kadar HB(X_7)	Numerikal

2. Analisis Data

Tahapan analisis data yang dilakukan dalam penelitian ini yaitu :

a) Pengambilan data yang dilakukan di UDD PMI Kabupaten Sigi, Provinsi Sulawesi Tengah
b) Membagi data menjadi dua yaitu data *training* dan data *testing*. Data *training* digunakan untuk pembentukan model dan data *testing* digunakan untuk validasi model.
c) Pembentukan model kNN dan regresi logistik biner dengan menggunakan data *training*.
d) Menentukan prediksi klasifikasi calon pendonor darah dengan menggunakan data *testing*.
e) Membandingkan tingkat akurasi yang dihasilkan dari metode kNN dan regresi logistik biner.
f) Menentukan metode terbaik untuk mengklasifikasi calon pendonor darah.
g) Menarik kesimpulan

Langkah-langkah analisis data dengan menggunakan regresi logistik biner adalah :

a. Melakukan analisis deskriptif karakteristik status pendonor dan faktor-faktor yang mempengaruhi status kelayakan calon pendonor darah.
b. Menentukan model awal regresi logistik biner.

Menurut Agresti (2002) untuk menentukan estimasi parameter regresi logistik biner dapat digunakan metode

Maximum Likelihood (kemungkinan maksimum) yang membutuhkan turunan pertama dan turunan kedua dari fungsi Likelihood. Secara umum, model regresi logistik biner dengan E(Y=1|x) dapat dituliskan dengan:

$$\pi(X) = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\sum_{k=1}^{n}\beta_{ik}D_{i+\beta_p x_p}}}{1+e^{\beta_0+\beta_1 x_1+\cdots+\sum_{k=1}^{n}\beta_{ik}D_{i+\beta_p x_p}}}$$

dimana $\pi(X)$ adalah peluang sukses suatu kejadian, x_i (untuk i =1,2,..., p) adalah faktor-faktor yang mempengaruhi peubah respon, p adalah banyaknya peubah penjelas yang digunakan, D adalah variabel *dummy* dan *k* adalah banyaknya variabel *dummy* yang digunakan. Dengan menggunakan transformasi logit, model tersebut dapat dituliskan dengan:

$$g(x) = ln\left(\frac{\pi(X)}{1-\pi(X)}\right)$$

$$= \beta_0 + \beta_1 x_1 + \cdots + \sum_{k=1}^{n}\beta_{ik}\,D_i + \beta_p x_p + \varepsilon$$

dimana $\beta_i,\ i = 1,2,\ldots p$

c. Melakukan pengujian koefisien parameter model secara simultan.

Hipotesis yang digunakan adalah:

$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$

$H_1 =$ minimal ada satu $\beta_i \neq 0$

Statistik uji yang digunakan yaitu :

$$G = -2\ln\left[\frac{\left(\frac{n_1}{n}\right)^{n_1}\left(\frac{n_0}{n}\right)^{n_0}}{\prod_{i=1}^{n}\pi_i^{y_i}(1-\pi_i)^{1-y_i}}\right]$$

dengan y_i adalah variabel respon, n_1 adalah $\sum y_i$, n_0 adalah $\sum(1-y_i)$ dan n adalah $n_0 + n_1$. Statistik uji G mengikuti sebaran χ^2 dengan derajat bebas $p-1$, dimana p adalah jumlah parameter yang digunakan.

d. Melakukan pengujian koefisien parameter model secara parsial.

Hipotesis yang digunakan :

$H_0 : \beta_i = 0$ (peubah penjelas ke-i tidak berpengaruh terhadap peubah respon)

$H_1 : \beta_i \neq 0$ (peubah penjelas ke-i berpengaruh terhadap peubah respon)

Statistik uji yang digunakan :

$$W = \frac{\widehat{\beta_\iota}}{SE(\widehat{\beta_\iota})}$$

dengan $\widehat{\beta_\iota}$: penduga parameter β_i dan $SE(\widehat{\beta_\iota})$: standard error dari $\widehat{\beta_\iota}$. Statistik uji Wald mengikuti distribusi chi-kuadrat dengan derajat bebas satu dimana H_0 ditolak saat $|W| > \chi^2_{(\alpha,1)}$.

d. Melakukan pengujian koefisien parameter model secara parsial.

Hipotesis yang digunakan :

H_0 : β_i=0 (peubah penjelas ke-i tidak berpengaruh terhadap peubah respon)

H_1 : β_i≠0 (peubah penjelas ke-i berpengaruh terhadap peubah respon)

Statistik uji yang digunakan :

W=(β_i)ˆ/SE((β_i)ˆ)

dengan $\widehat{\beta_\iota}$: penduga parameter β_i dan $SE(\widehat{\beta_\iota})$: standard error dari $\widehat{\beta_\iota}$. Statistik uji Wald mengikuti distribusi chi-kuadrat dengan derajat bebas satu dimana H_0 ditolak saat $|W| > \chi^2_{(\alpha,1)}$.

e. Menguji kelayakan model.

Pengujian kelayakan (*goodness of fit*) pada model regresi logistik menggunakan uji Hosmer-Lemeshow. Uji Hosmer-Lemeshow didasarkan pada pengelompokan pada nilai dugaan peluangnya yang menyebar Chi- Kuadrat (Hosmer & Lemeshow, 2000).

Hipotesis yang digunakan :

H_0 : model yang dibangun layak

H_1 : model yang dibangun tidak layak

Statistik uji yang digunakan :

$$\hat{C} = \sum_{k=1}^{g}\frac{(O_k - n'_k\overline{\pi_k})^2}{n'_k\overline{\pi_k}(1-\overline{\pi_k})}$$

dengan $\hat{C}$ adalah statistik Hosmer-Lemeshow, $O_k \sum_{j=1}^{c_k} y_i$ adalah jumlah nilai peubah respon pada kelompok ke-k, g adalah banyaknya amatan dalam kelompok ke-k, n'_k adalah jumlah sampel pada kelompok ke- k, c_k adalah banyaknya kombinasi peubah bebas pada kelompok ke-k dan $\overline{\pi_k}$ adalah rata-rata dari $\hat{\pi}$ untuk kelompok ke-k. Statistik $\hat{C}$ menyebar

mengikuti sebaran Chi-Kuadrat dengan derajat bebas $g-2$ (Hosmer dan Lemeshow, 2000). Kesimpulan menolak hipotesis nol jika nilai $C_{\text{Hitung}} > \chi^2_{\alpha(g-2)}$.

e. Menguji kelayakan model.

Pengukuran kinerja klasifikasi dilakukan dengan matriks konfusi (*confusion matrix*). Matriks konfusi merupakan tabel yang mencatat hasil kinerja klasifikasi.

Tabel 2. Matriks Konfusi

Hasil observasi	Taksiran	
	y_1	y_2
y_1	n_{11}	y_1
y_2	n_{21}	y_2

Keterangan :

n_{11} : jumlah subjek dari y_1 tepat diklasifikasikan sebagai y_1

n_{12} : jumlah subjek dari y_1 salah diklasifikasikan sebagai y_2

n_{12} : jumlah subjek dari y_2 salah diklasifikasikan sebagai y_1

n_{11} : jumlah subjek dari y_2 tepat diklasifikasikan sebagai y_2.

Perhitungan nilai akurasi merupakan proporsi observasi yang diprediksi benar oleh fungsi klasifikasi, digunakan rumus :

$$\text{Akurasi} = \frac{n_{11} + n_{22}}{n_{11} + n_{12} + n_{21} + n_{22}}$$

Adapun langkah- langkah dari algoritma kNN adalah:

a) Tentukan parameter *k*

Penentuan *k* terbaik menggunakan *k-fold cross validation.*

b) Hitung jarak antara data testing dan data training.

Jika data berbentuk numerik maka menggunakan jarak Euclid seperti pada persamaan berikut:

$$D(x_1, y_1) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Keterangan :

x_i : data training

y_i : data testing

n : dimensi data

Perhitungan jarak antar observasi berupa data campuran kuantitatif dan kualitatif dapat dilakukan dengan menggunakan koefisien kemiripan umum Gower. Koefisien kemiripan Gower dapat digunakan untuk melihat kemiripan antar observasi dengan melakukan perhitungan jarak pada setiap peubah acak yang ada sesuai dengan skala pengukuran perubah acak tersebut (Gower, 1971). Secara umum, persamaan kemiripan Gower ditunjukkan pada persamaan berikut :

$$s(x_i, x_j) = \frac{\sum_{k=1}^{p} s_k(x_{ik}, x_{jk}) \delta(x_{ik}, x_{jk}) w_k}{\sum_{k=1}^{p} \delta(x_{ik}, x_{jk}) w_k}$$

$$\delta(x_{ik}, x_{jk}) = \begin{cases} 1; x_{ik}, x_{jk} \in \mathbb{R} \\ 0; lainnya \end{cases}$$

Keterangan :

$s(x_i, x_j)$: koefisien kemiripan Gower antara observasi ke – i dan j

$s_k(x_{ik}, x_{jk})$: koefisien kemiripan antara observasi ke-i dan j pada peubah k

$\delta(x_{ik}, x_{jk})$: kemungkinan perbandingan peubah k obsevasi ke-i dan j

w_k : bobot pilihan yang menyatakan kepentingan variabel, $w_k = 1$

x_{ik}, x_{jk} : nilai observasi ke-i dan j pada peubah ke-k

Koefisien kemiripan s_k pada setiap variabel dihitung berdasarkan skala pengukuran peubah *k*. Perhitungan nilai s_k pada skala nominal, ordinal, interval dan rasio secara berurutan ditunjukkan pada persamaan berikut :

Nominal : $s_k(x_{ik}, x_{jk}) = \begin{cases} 1; x_{ik} = x_{jk} \\ 0; x_{ik} \neq x_{jk} \end{cases}$

Ordinal : $s_k(x_{ik}, x_{jk}) = 1 - \frac{|r_k(x_{ik}) - r_k(x_{jk})|}{max_m\{r_k(x_{mk})\} - min_m\{r_k(x_{mk})\}}$

Interval ; Rasio : $s_k(x_{ik}, x_{jk}) = 1 - \frac{|x_{ik} - x_{jk}|}{max_m\{x_{mk}\} - min_m\{x_{mk}\}}$

Keterangan :

x_{mk} : nilai observasi ke-*m* peubah *k*

$r_k(x_{mk})$: rank dari nilai observasi ke-*m* peubah ordinal *k*

$max_m\{x_{mk}\}$:nilai maksimum dari seluruh nilai peubah *k*

$min_m\{x_{mk}\}$: nilai minimum dari seluruh nilai peubah *k*

$max_m\{r_k(x_{mk})\}$: rank maksimum dari seluruh nilai peubah ordinal *k*
$min_m\{r_k(x_{mk})\}$: rank minimum dari seluruh nilai peubah ordinal *k*

Penentuan tetangga terdekat kNN diperoleh berdasarkan nilai jarak ketakmiripan antar observasi sehingga nilai koefisien kemiripan Gower perlu ditransformasi menjadi nilai koefisien ketakmiripan dengan menggunakan persamaan :

$$d_k(x_i, x_j) = 1 - s_k(x_i, x_j)$$

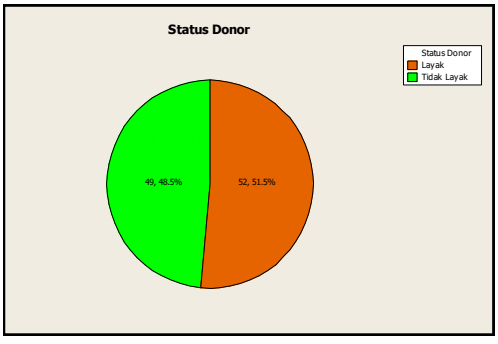
Keterangan :
$d_k(x_i, x_j)$: koefisien ketakmiripan Gower observasi ke–*i* dan *j* pada peubah k.

c) Memilih jarak terdekat sampai pada parameter k
d) Memilih jumlah kelas terbanyak sebanyak k lalu diklasifikasikan
e) Menentukan nilai akurasi klasifikasi.

HASIL DAN PEMBAHASAN

Gambaran Karakteristik Status Pendonor Darah

Informasi masing-masing peubah dapat diperoleh dengan melakukan eksplorasi data. Status pendonor sebagai peubah respon dengan kategori (1) layak mendonorkan darahnya dan (0) tidak layak mendonorkan darahnya. Berdasarkan Gambar 1, status pendonor yang layak mendonorkan darahnya sebesar 51.5% yaitu sebanyak 52 orang dan yang tidak layak sebesar 48.5% yaitu sebanyak 49 orang.

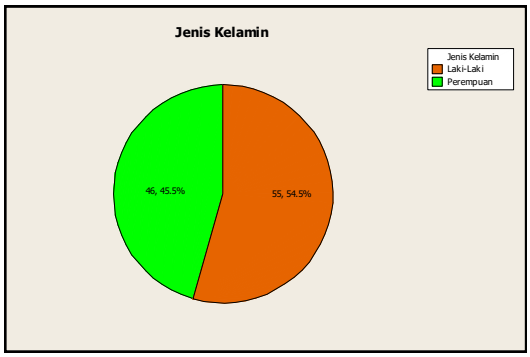


Gambar 1. Karakteristik Status Pendonor Darah

Gambaran Karakteristik Jenis Kelamin

Jenis Kelamin mempunyai skala pengukuran kategori dengan label (1) laki-laki dan (2) perempuan. Jumlah pendonor laki-laki sebanyak 55 orang dan perempuan sebanyak 46 orang. Informasi untuk peubah jenis kelamin ditampilkan seperti gambar berikut:

Berdasarkan Gambar 2 di atas, dari 101 responden diperoleh hasil persentase calon pendonor laki-laki sebanyak 54.5% dan pendonor perempuan sebanyak 45.5%.



Gambar 2. Karakateristik Jenis Kelamin

Gambaran Karakteristik Usia, Berat Badan, Tinggi Badan, Sistole, Diastole dan Kadar Haemoglobin

Peubah usia, berat badan, tinggi badan, sistole, diastole dan kadar HB bertipe numerikal sehingga informasi tentang karakteristik datanya dapat dilihat dari statistik deskriptif. Berikut adalah gambaran karakteristik peubah usia, berat badan, tinggi badan, sistole, diastole dan kadar HB :

Tabel 3. Statistik Deskriptif

	Min	Max	Mean	Std Dev
Usia	19	54	31.89	8.233
Berat badan	50	89	65.01	7.349
Tinggi badan	152	180	164.65	5.317
Sistole	100	160	118.42	20.530
Diastole	50	96	69.81	12.844
Kadar HB	46	83	61.14	9.881

Berdasarkan Tabel 3 dapat dilihat bahwa untuk peubah usia memiliki rata-rata usia calon pendonor sekitar 31 tahun dengan standar deviasi sebesar 8.233. Rata-rata berat badan dan tinggi badan pendonor

adalah 65 kg dan 164 cm dengan standar deviasi 7.349 dan 5.317. Tekanan darah terdiri dari dua yaitu tekanan darah atas (*sistole*) dan tekanan darah bawah (*diastole*). Rata-rata *sistole* dan *diastole* pendonor adalah 118.42 mm Hg dan 69.81 mm Hg dengan standar deviasi 20.53 dan 12.844. Peubah kadar_HB memiliki nilai rata-rata 61.14 g/dL dan standar deviasi 9.881 g/dL dengan kadar Hb tertinggi yaitu 83 g/dL dan terendah yaitu 46 g/dL.

1. Regresi Logistik

Regresi logistik biner merupakan suatu metode statistik yang digunakan untuk menganalisis hubungan antara peubah penjelas (X) dengan peubah respon (Y) yang berupa data kategori dikotomi. Nilai variabel Y=1 menyatakan adanya suatu karakteristik dan Y=0 menyatakan tidak adanya suatu karakteristik. Tahapan analisis data pada regresi logistik adalah :

Penentuan Model Awal Regresi Logistik Biner

Langkah pertama dalam analisis regresi logistik biner adalah menentukan model awal regresi logistik biner (Model 1). Pada penelitian ini regresi logistik biner dilakukan dengan meregresikan peubah status kelayakan calon pendonor darah dengan peubah jenis kelamin, usia, berat badan, tinggi badan, s*istole*, *diastole*, dan kadar Haemoglobin. Secara umum, model awal regresi logistik biner yang diperoleh yaitu :

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

dengan nilai

$$g(x) = 1.7532 - 0.4779X_{12} + 0.0670X_2 - 0.0313X_3 + 0.0928X_4 - 0.0016X_5 - 0.0024X_6 - 0.2817X_7$$

Pengujian Koefisien Parameter secara Simultan

Hipotesis yang digunakan adalah

$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$

$H_1 =$ minimal ada satu $\beta_i \neq 0$

Nilai G yang diperoleh pada pengujian secara simultan adalah 52.943. Karena nilai G = 52.943 > $\chi^2_{(0.05,7)}$=14,0671 maka H_0 ditolak. Berdasarkan keputusan bahwa H_0 ditolak maka disimpulkan bahwa peubah penjelas yang terdapat pada model berpengaruh nyata secara simultan.

Pengujian Koefisien Parameter secara Parsial

Hipotesis yang digunakan :

$H_0 : \beta_i = 0$ (peubah penjelas ke-*i* tidak berpengaruh terhadap peubah respon)

$H_1 : \beta_i \neq 0$ (peubah penjelas ke-*i* berpengaruh terhadap peubah respon)

Tabel 4. Nilai Wald Model 1

Peubah	**Wald**	***Sig***	**Keputusan**
Jenis Kelamin Perempuan	0.282	0.595	Tidak signifikan
Usia	1.168	0.280	Tidak signifikan
Berat Badan	0.284	0.594	Tidak signifikan
Tinggi Badan	1.788	0.181	Tidak signifikan
Sistole	0.010	0.919	Tidak signifikan
Diastole	0.003	0.954	Tidak signifikan
Kadar HB	11.577	0.001	Signifikan

Berdasarkan Tabel 4 diperoleh bahwa peubah yang berpengaruh nyata terhadap status kelayakan calon pendonor darah adalah kadar Haemoglobin, hal ini ditunjukkan dengan nilai *sig* = 0.001 < α = 0.05. Model 1 masih didapatkan peubah yang tidak berpengaruh signifikan terhadap model, karena memiliki nilai *Sig* > α = 5% sehingga peubah yang tidak berpengaruh harus dihilangkan. Langkah selanjutnya adalah menghilangkan peubah jenis kelamin perempuan, usia, berat badan, tinggi badan, *sistole*, dan *diastole* dari model regresi logistik biner.

Metode AIC adalah metode yang dapat digunakan untuk memilih model regresi terbaik yang ditemukan oleh Akaike dan Schwarz (Grasa, 1989). Menurut metode AIC, model regresi terbaik adalah model regresi yang mempunyai nilai AIC terkecil. Adapun nilai AIC untuk setiap model 1 dan 2 adalah :

Tabel 5. Nilai AIC Kedua Model

	Model	AIC
1	Model 1	62.81449
2	Model 2	55.60601

Dari hasil output diatas dapat dilihat bahwa Model 2 merupakan model yang terbaik karena memiliki nilai AIC terkecil yaitu 55.60601.

Model kedua diperoleh dari menghilangkan peubah-peubah yang tidak signifikan pada model pertama. Hasil yang diperoleh sebagai berikut :

$$g(x) = -15.176 + 0.257X_1$$

Pengujian koefisien regresi secara simultan didapatkan nilai G sebesar 48.152. Nilai tersebut lebih besar daripada $\chi^2_{(0.05,1)}$=3.84 sehingga peubah penjelas yang terdapat pada model berpengaruh nyata secara serentak.

Nilai Wald yang diperoleh pada model kedua sebesar 22.058 dengan sig = 0.000. Nilai sig model kedua < α = 5% sehingga peubah kadar haemoglobin berpengaruh nyata terhadap status kelayakan calon pendonor darah.

Pengujian Kelayakan Model

Hipotesis yang digunakan :
H_0 : model yang dibangun layak
H_1: model yang dibangun tidak layak

Berdasarkan uji kesesuaian model diperoleh nilai $\hat{C}$ = 11.270 < $\chi^2_{(0.05,8)}$ = 15.5073 sehingga H_0 diterima. Jadi, model regresi logistik biner yang terbentuk layak atau tidak ada perbedaan antara observasi dengan kemungkinan hasil prediksi.

Setelah dilakukan uji signifikansi terhadap model, baik secara keseluruhan maupun individual serta dilakukan uji kesesuaian model maka diperoleh model akhir sebagai berikut:

$$\pi(x) = \frac{\exp(-15.176 + 0.257)}{1 + \exp(-15.176 + 0.257)}$$

Odds Ratio untuk kadar Haemoglobin (exp(0.257)) = 1.293 artinya semakin tinggi kadar Haemoglobin calon pendonor darah maka kecenderungannya semakin layak untuk mendonorkan darahnya.

Ketepatan Klasifikasi

Hasil prediksi pada data *testing* selanjutnya dibandingkan dengan data sebenarnya pada respon data testing. Nilai akurasi diperoleh dari fungsi confusion Matrix () dengan menggunakan paket caret pada software R. Berikut adalah hasil ketepatan klasifikasi status kelayakan calon pendonor darah :

Tabel 6. Matriks Konfusi

Aktual	Prediksi	
	Layak	Tidak
Layak	15	0
Tidak	2	12

Berdasarkan Tabel 6, dapat dihitung nilai akurasi sebesar 93% yang didapat dari membandingkan jumlah dari calon pendonor darah yang tepat diklasifikasikan dengan jumlah seluruh observasi.

Output yang dihasilkan R untuk analisis regresi logistik biner adalah sebagai berikut:

```
Call:
glm(formula = Status.Donor ~ ., family = binomial(link = "logit"),
    data = training)

Deviance Residuals:
     Min       1Q   Median       3Q      Max
-2.32715  -0.34092  -0.05505   0.50735   2.33770

Coefficients:
                Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)     1.753221  12.313344   0.142    0.886777
JKPerempuan    -0.477929   0.899929  -0.531    0.595368
Usia            0.067030   0.062017   1.081    0.279769
Bb             -0.031323   0.058774  -0.533    0.594079
Tb              0.092768   0.069381   1.337    0.181196
Diastole       -0.001653   0.016309  -0.101    0.919273
Sistole        -0.002425   0.042318  -0.057    0.954300
Kadar          -0.281751   0.082807  -3.402    0.000668

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 99.758  on 71  degrees of freedom
Residual deviance: 46.814  on 64  degrees of freedom
```

AIC: 62.814

Number of Fisher Scoring iterations: 6

Output Confusion Matrix yang dihasilkan software R yaitu :

Confusion Matrix and Statistics

Reference
Prediction Layak tidak
Layak 15 2
tidak 0 12

Accuracy : 0.931
95% CI : (0.7723, 0.9915)
No Information Rate : 0.5172
P-Value [Acc > NIR] : 1.901e-06
Kappa : 0.8612
Mcnemar's Test P-Value : 0.4795

Sensitivity : 1.0000
Specificity : 0.8571
Pos Pred Value : 0.8824
Neg Pred Value : 1.0000
Prevalence : 0.5172
Detection Rate : 0.5172
Detection Prevalence : 0.5862
Balanced Accuracy : 0.9286

'Positive' Class: Layak

2. k-Nearest Neighbor (KNN)

Menurut Prasetyo (2012) k-Nearest Neighbor (k-NN) adalah metode yang melakukan klasifikasi berdasarkan kedekatan lokasi (jarak) suatu data dengan data lain. Nilai k pada kNN berarti k-data terdekat dari data testing.

Metode k-NN cukup sederhana, tidak ada asumsi mengenai distribusi data dan mudah diaplikasikan. Pemilihan nilai k (jumlah data/tetangga terdekat) ditentukan dengan menggunakan cross validation. Cross validation adalah sebuah teknik validasi model untuk menilai bagaimana hasil statistik analisis akan menggeneralisasi kumpulan data independen. Teknik ini utamanya digunakan untuk melakukan prediksi model dan memperkirakan seberapa akurat sebuah model prediktif ketika dijalankan dalam praktiknya. Salah satu teknik dari validasi silang adalah k-fold cross validation, yang mana memecah data menjadi k bagian set data dengan ukuran yang sama. Penggunaan k-fold cross validation untuk menghilangkan bias pada data. Pemilihan nilai k ini bisa mempengaruhi tingkat akurasi prediksi yang dikerjakan (Santosa, 2007).

Langkah-langkah dari perhitungan kNN adalah sebagai berikut:

a) Normalisasi data calon pendonor darah.

Normalisasi data linier adalah proses penskalaan nilai atribut data sehingga bisa jatuh pada range tertentu. Tujuan dari normalisasi data adalah untuk mempersempit atau mengecilkan nilai range pada data tersebut. Keuntungan dari metode ini adalah keseimbangan nilai perbandingan antara data saat sebelum dan sesudah nilai normalisasi. Kekurangannya adalah jika ada data baru metode ini akan memungkinkan terjebak pada *out of bound error*. Normalisasi dihitung menggunakan persamaan berikut:

$$Nor\ (X^*) = \frac{X - \min X}{(maks\ X - \min X)}$$

b) Menghitung jarak data training ke data acuan (data testing) calon pendonor darah menggunakan *euclidian distance*.

c) Selanjutnya mengurutkan hasil jarak *Euclidean* dari yang terkecil ke terbesar.

d) Proses menentukan nilai k.

e) Proses mencari label atau kelas mayoritas sebanyak nilai K sesudah normalisasi.

Output yang dihasilkan *software* R untuk mendapatkan *k* terbaik adalah sebagai berikut :

k-Nearest Neighbors

72 samples

7 predictor

2 classes: 'Layak', 'tidak'

Pre-processing: centered (7), scaled (7)

Resampling: Cross-Validated (10 fold, repeated 3 times)

Summary of sample sizes: 66, 65, 64, 64, 66, 64, ...

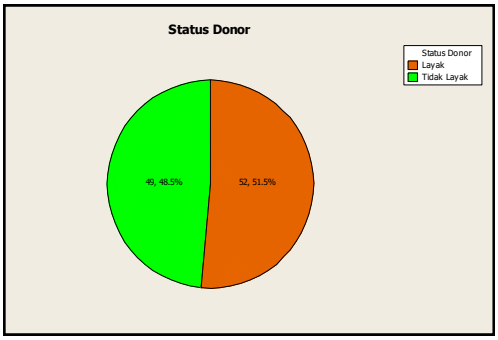
Resampling results across tuning parameters:

k	Accuracy	Kappa
5	0.8261905	0.6512836
7	0.8103175	0.6173531
9	0.8067460	0.6107975
11	0.8158730	0.6301212
13	0.8075397	0.6134545
15	0.8035714	0.6066169
17	0.8049603	0.6086768
19	0.7946429	0.5903862
21	0.7912698	0.5830190
23	0.8160714	0.6308053
25	0.8222222	0.6445058
27	0.8263889	0.6528392
29	0.8305556	0.6611725
31	0.8472222	0.6936942
33	**0.8486111**	**0.6964720**
35	0.8438492	0.6863869
37	0.8299603	0.6600450
39	0.8450397	0.6899165
41	0.8347222	0.6687541
43	0.8305556	0.6604207

Accuracy was used to select the optimal model using the largest value.

The final value used for the model was k = 33.

Akurasi digunakan untuk memilih model yang optimal dengan menggunakan nilai yang tertinggi. Berdasarkan output di atas, nilai k terbaik berdasarkan akurasi yang tertinggi adalah 33. Pada k = 33 nilai akurasi yang dihasilkan adalah 84.86%, artinya kemampuan model dalam menebak status kelayakan calon pendonor darah adalah sebesar 84.86%. Pemilihan k terbaik juga dapat dilihat dari plot yang dihasilkan dari *Cross Validation* yang berulang-ulang. Berikut adalah tampilan plot yang menununjukkan nilai k terbaik :



Gambar 3. Plot Akurasi berdasarkan *k-fold Cross validation*

Berdasarkan Gambar 3 banyaknya *k* terbaik yang dihasilkan dari plot akurasi adalah 33. Hasil *k* yang sama didapatkan dengan menggunakan k-*fold Cross Validation*. Iterasi *k* yang akan digunakan pada model dengan berbagai nilai tingkat akurasi serta nilai parameter yang lain dapat dilihat pada tabel *confusion Matrix*. Berikut adalah output yang dihasilkan *software* R :

fusion Matrix and Statistics

Prediction	Reference: Layak	Reference: tidak
Layak	14	5
tidak	1	9

Accuracy : 0.7931

95% CI : (0.6028, 0.9201)

No Information Rate : 0.5172

P-Value [Acc > NIR] : 0.002088

Kappa : 0.5817

Mcnemar's Test P-Value : 0.220671

Sensitivity : 0.9333

Specificity : 0.6429

Pos Pred Value : 0.7368

Neg Pred Value : 0.9000

Prevalence : 0.5172

Detection Rate : 0.4828

Detection Prevalence : 0.6552

Balanced Accuracy : 0.7881

'Positive' Class : Layak

Berdasarkan output di atas didapatkan nilai akurasi sebesar 79.31%, artinya kemampuan model kNN menebak status pendonor darah sebesar 79.31%. Sedangkan berdasarkan data aktual orang yang memiliki status layak mendonorkan darahnya, model dapat menebak dengan benar sebesar 93.3%. Berdasarkan data aktual orang yang memiliki status pendonor tidak layak mendonorkan darahnya, model dapat menebak dengan benar sebesar 64.29%.

3. Perbandingan Akurasi Klasifikasi kNN dan Regresi Logistik Biner

Berdasarkan hasil analisis yang telah dilakukan yaitu klasifikasi menggunakan kNN dan regresi logistik biner didapatkan hasil akurasi masing-masing metode, hasil akurasi tersebut digunakan untuk membandingkan kedua metode tersebut. Semakin tinggi nilai akurasinya maka semakin tinggi ketepatan klasifikasi suatu model. Berikut adalah perbandingan akurasi dari kedua metode:

Tabel 7. Perbandingan Akurasi

Model	Akurasi
kNN	79.31%
Regresi logistik biner	93.1%

Jika dilihat dari perbandingan nilai akurasi metode regresi logistik biner dan kNN, maka metode terbaik untuk mengklasifikasikan status kelayakan pendonor darah adalah metode regresi logistik biner. Metode regresi logistik biner memiliki kemampuan dalam memprediksi benar dari data aktual pendonor yang layak mendonorkan darahnya lebih baik daripada metode kNN karena memiliki nilai akurasi lebih tinggi dari pada metode kNN.

KESIMPULAN

proksi dari pendapatan per kapita Berdasarkan hasil dan pembahasan yang telah dipaparkan sebelumnya, maka dapat diambil kesimpulan bahwa akurasi yang dihasilkan metode kNN sebesar 79.31% sedangkan metode regresi logistik biner adalah 93.1%. Hasil perbandingan ketepatan klasifikasi antara kNN dan regresi logistik biner dapat dilihat dari nilai akurasinya. Semakin tinggi tingkat akurasi maka model akan semakin baik dalam mengklasifikasikan. Nilai akurasi klasifikasi yang diperoleh pada metode regresi logistik biner lebih tinggi daripada metode kNN. Sehingga metode terbaik untuk mengklasifikasikan status calon pendonor darah adalah regresi logistik biner karena memiliki keakuratan klasifikasi yang lebih baik daripada kNN.

Hasil lain yang diperoleh dari penelitian ini adalah peubah yang berpengaruh nyata pada status kelayakan calon pendonor darah adalah peubah kadar Haemoglobin. Hasil ini didapatkan dari pengujian pada regresi logistik biner.

DAFTAR PUSTAKA

Agresti, A. 2002. *Categorical Data Analysis*. New York : John Wiley&Sons.

Bayususetyo, D., Santoso, R., dan Tarno. 2017. Klasifikasi Calon Pendonor Darah Menggunakan Metode Naive Bayes Classifier (Studi Kasus : Calon Pendonor Darah di Kota Semarang). *Jurnal Gaussian* 6(2) : 193-200.

Bhatia, M., Vandana., 2010. Survey of Nearest Neighbor Techniques. *International Journal of Computer Science and Information Security* 8, 1947-5500.

Departemen Kesehatan RI. 2009. Donor Darah. Tersedia pada http://kemenkes.go.id/

Gower, JC. 1971. A General Coefficient of Similarity and Some of Its Properties. *Biometrics* 27(4): 857-871.

Grasa, A. 1989 Econometric Model Selection: A New. Approach. Kluwer. Springer Science and Bussiness Media.

Hosmer, D.W. and Lemeshow, S. 2000. *Applied Logistic Regression*. New York: John Wiley & Son, Inc.

Moradian, M., and Baraani, A. 2009. K-Nearest Neighbor Based Association Algorithm. *Journal of Theoretical and Applied Information Technology* 6, 123 – 129.

Nugroho, EB., Furqon, MT. dan Hidayat, N. 2018. Klasifikasi Pendonor Darah Menggunakan Metode Support Vector Machine (SVM) pada Dataset RFMTC. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer* 2(10) : 3860-3864.

Palang Merah Indonesia. (2019, September 11). Pelayanan Donor Darah. Tersedia pada http://www.pmi.or.id/

Prasetyo, E. 2012. Data Mining Konsep dan Aplikasi Menggunakan MATLAB. Yogyakarta: Penerbit ANDI Yogyakarta.

Santosa, B. 2007. *Data Mining Terapan*. Yogyakarta : Graha Ilmu.

Sapriana, B.M. 2017. Penerapan Algoritma Naive Bayes Classifier Pada Klasifikasi Status Kelayakan Pendonor Darah di Unit Transfusi Darah Palang Merah Indonesia (UTD PMI) Kota Makassar. Skripsi. Makassar: Universitas Negeri Makassar.

Wu X, Kumar V, Quinlan JR, Ghosh J, Yang Q, Motoda H, McLachlan GJ, Ng A, Liu B, Yu PS *et al*. 2008. Top 10 Algorithms in Data Mining. *Journal of Knowledge and Information Systems* 14(1): 1-37.