METODE REGRESI LOGISTIK BINER DAN METODE K-NEAREST NEIGHBOR PADA KLASIFIKASI MENOPAUSE DINI WANITA DISTRIK ORANSBARI PROVINSI PAPUA BARAT

(*Binary Logistic Regression Method and K-Nearest Neighbor Method on Classification of Early Menopausal Women In Oransbari District West Papua Province*)

Indah Ratih Anggriyani[1)], Eka Dewi Kusumawati [2)], Elda Irma Jeanne Joice Kawulur [3*)]

[1, 2, 3)]Universitas Papua, Jl. Gunung Salju Amban, Manokwari Papua Barat
e-mail: i.anggriyani@unipa.ac.id, ekhadewik10@gmail.com, e.kawulur@unipa.ac.id
[*)]*penulis korespondensi*

Abstract. Machine learning is a developing part of artificial intelligence. One part of that is classification. Two classification methods in this study are binary logistic regression and k-nearest neighbor. Both methods were applied to cases of women with early menopause in Oransbari district West Papua Province. The aim is to determine the effectiveness of the two methods in several conditions of training and testing data. The data with the proportion of 80% training and 20% testing resulted in the best level of effectiveness. In general, the binary logistic regression method produces a higher model accuracy than the kNN method. The accuracy of predicting women with early menopause is higher than the binary logistic regression method.

Keywords: Binary Logistics Regression, K-Nearest Neighbor, Classification Method, Early Menopause

1. Pendahuluan

Salah satu cabang dari teknologi kecerdasan buatan (*artificial intelligence*) adalah *machine learning*. *Machine learning* adalah mesin yang dikembangkan untuk bisa belajar dengan sendirinya tanpa arahan dari pengguna. Hal ini dikembangkan berdasarkan disiplin ilmu seperti statistika, matematika dan data mining sehingga mesin dapat belajar dengan menganalisa data tanpa perlu diprogram ulang atau diperintah [7].

Salah satu tipe dari *machine learning* adalah *supervised. Supervised learning* merupakan algoritma klasifikasi yang membuat model prediktif berdasarkan input (peubah bebas) dan *output* (peubah respon) data. Terdapat beberapa metode pada tipe ini yaitu regresi linear, k-*nearest neighbor*, *support vector machine*, naive bayes, *random forest* dan *neural networks*. Pada penelitian ini hanya menggunakan dua metode yaitu regresi linear dan k-*nearest neighbor* [2].

Klasifikasi pada metode regresi dapat dilakukan jika peubah respon berbentuk kategorik atau yang dikenal dengan regresi logistik. Pada saat peubah respon berbentuk dikotomi maka digunakan regresi logistik biner. Model regresi logistik biner dihasilkan dengan menggunakan transformasi logit. Amatan diklasifikasikan berdasarkan nilai ambang

probabilitas.

Metode K-Nearest Neighbor (KNN) merupakan metode nonparametrik untuk klasifikasi [2]. Klasifikasi dilakukan terhadap amatan berdasarkan data training yang jaraknya paling dekat dengan amatan tersebut. Dekat atau jauhnya tetangga dihitung menggunakan jarak euclidean. Proses KNN dimulai dengan menentukan parameter k (jumlah tetangga terdekat). Semakin besar nilai k maka semakin kecil batas klasifikasi dan semakin kecil nilai k maka semakin berbelit-belit batasnya. Kelebihan dari metode ini adalah dapat mengatasi pencilan, mudah diimplementasikan dan dapat menanggulangi data yang jumlahnya besar.

Saat membuat model, data terlebih dahulu dibagi menjadi dua yaitu training dan testing. Data training digunakan untuk membentuk sebuah model sedangkan data testing digunakan untuk memvalidasi model yang telah dihasilkan. Tujuan dari validasi adalah menghasilkan sebuah model yang representatif terhadap sistem kenyataannya dan meningkatkan kredibilitas. Persentase data yang digunakan sebagai training lebih besar dibandingkan testing, walaupun tidak ada ukuran persentase yang tepat untuk masing-masing jenis data tersebut.

Menopause adalah berhentinya menstruasi secara permanen akibat hilangnya aktivitas folikel ovarium [6,8]. Seorang wanita ketika memasuki fase menopause dini ditandai dengan tidak mengalami periode menstruasi selama 12 bulan dan biasanya terjadi pada usia rata-rata 50 tahun. Terdapat 5% wanita yang mengalami menopause dini pada usia 40-45 tahun dan sekitar 1% pada usia sebelum usia 40 tahun [9]. Menopause dini yang terjadi secara ilmiah pada perempuan di usia antara 40-45 tahun termasuk kejadian yang jarang dialami. Kami menemukan kejadian tersebut dialami oleh wanita di Distrik Oransbari dan faktor-faktor yang mempengaruhinya [4].

Tujuan dari penelitian ini adalah mengetahui efektifitas kedua metode pada beberapa kondisi data training dan testing. Efektifitas yang dimaksud meliputi tingkat keakuratan dan sensitifitas yang dihasilkan. Sebagai penerapannya, kedua metode diaplikasikan pada data menopause dini wanita jawa Distrik Oransbari Provinsi Papua Barat.

2. Metodologi

Peubah respon pada penelitian ini terdiri dari dua kategori yaitu menopause dini sebagai klasifikasi positif “1” dan menopause normal sebagai klasifikasi negatif “0”. Peubah bebas yang digunakan adalah usia lahiran pertama, pendidikan dan kontrasepsi [4]. Secara umum terdapat empat tahap yang digunakan pada penelitian ini yaitu pembagian data, penentuan dugaan validasi, klasifikasi amatan dengan menggunakan metode regresi logistik biner dan KNN serta menghitung dan membandingkan efektifitas tiap metode.

Tahapan yang digunakan dalam penelitian ini adalah sebagai berikut:

2.1 Tahap pembagian data

Tahap pertama yang dilakukan dalam penelitian ini adalah membagi data menjadi dua yaitu training dan testing, dengan persentase data training lebih besar dibandingkan testing. Empat kondisi yang diujikan pada penelitian ini yaitu 70% dan 30%; 75% dan 25%; 80% dan 20% serta 85% dan 15%.

2.2 Tahap penentuan dugaan validasi

Metode validasi yang digunakan pada tahap ini adalah k-*fold cross validation.* Merupakan metode validasi silang yang membagi data menjadi k-lipatan, kemudian melatih data pada lipatan k-1 dan menguji pada lipatan sisanya. Metode ini memungkinkan untuk semua data berperan sebagai data training dan data testing. Kelebihan dari metode ini adalah menghasilkan bias dan varians yang sedang [1]. Penelitian ini menggunakan jumlah lipatan (k) sebesar 5.

2.3 Klasifikasi amatan

Tahap ini terdiri atas dua bagian karena menggunakan dua metode yaitu regresi logistik biner dan KNN. Adapun tahapan dari setiap metode adalah sebagai berikut:

a. Metode regresi logistik biner

Langkah-langkah yang digunakan seperti tahapan pada metode regresi logistik biner [3], yaitu

i. Menduga parameter

Parameter dalam model regresi logistik dilakukan dengan menggunakan metode kemungkinan maksimum. Jika antara amatan yang satu dengan yang lain diasumsikan bebas maka fungsi kemungkinan maksimum yang diperoleh adalah

$$l(\beta) = \prod_{i=1}^{p} \pi(x_i)^{y_i}[1 - \pi(x_i)]^{1-y_i} \quad (1)$$

dengan i adalah $1,2,..p$; y_i adalah pengamatan pada peubah respon ke-i dan $\pi(x_i)$ adalah peluang untuk peubah penjelas ke-i. Parameter β_i diduga dengan memaksimumkan persamaan (1) menggunakan pendekatan logaritma, sehingga fungsi log-likelihoodnya sebagai berikut:

$$L(\beta) = \sum_{i=1}^{p} \{y_i \, ln \, ln \, [\pi(x_i)] + (1 - y_i) \, ln \, ln \, [1 - \pi(x_i)]\}$$

Nilai dugaan β_i diperoleh dengan membuat turunan pertama $L(\beta)$ terhadap $\beta_i = 0$. Secara umum jika sebuah peubah berskala nominal atau ordinal mempunyai k kemungkinan nilai, maka diperlukan k-1 peubah boneka (*dummy variable*). Dengan demikian model transformasi logitnya menjadi

$$g(x) = \beta_0 + \beta_1 X_1 + \ldots + \sum_{u}^{k_j - 1} \beta_{ju} D_{ju} + \beta_p X_p$$

dengan $k_j - 1$ adalah jumlah peubah boneka; β_{ju} adalah koefisien peubah boneka dan D_{ju} adalah $k_j - 1$ peubah boneka.

ii. Menguji signifikansi parameter
Pengujian signifikan parameter pada penelitian ini tidak dilakukan karena sudah dilakukan pada penelitian sebelumnya [4].

iii. Menentukan peluang
Peluang terjadinya suatu kejadian diperoleh dengan menggunakan rumus sebagai berikut:

$$\pi(x) = \frac{exp\ exp\ (g(x))}{1 + exp\ exp\ (g(x))} \tag{2}$$

Jika nilai $\pi(x) \geq 0.5$ maka dibulatkan menjadi 1 yang berarti pengklasifikasian amatan adalah dikelas yang menjadi pusat perhatian. Sebaliknya, jika nilai $\pi(x) < 0.5$ maka dibulatkan menjadi 0 yang berarti pengklasifikasian amatan adalah dikelas yang bukan menjadi pusat perhatian [3].

b. Metode KNN
Langkah-langkah yang digunakan seperti tahapan pada metode KNN [5], yaitu

i. Menentukan parameter k (jumlah tetangga paling dekat). Nilai k berupa bilangan bulat berapa saja.

ii. Menghitung kuadrat jarak Euclid masing-masing objek terhadap data sampel yang diberikan

$$d(P_1, P_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \tag{3}$$

iii. Mengurutkan objek-objek kedalam kelompok yang memiliki jarak terkecil

iv. Mengumpulkan kategori Y (klasifikasi tetangga terdekat)

v. Dengan kategori tetangga terdekat yang paling banyak, maka dapat ditetapkan sebuah objek masuk pada klasifikasi tertentu.

2.4 Perhitungan efektifitas model

Efektifitas model yang diukur meliputi ukuran akurasi, sensitivitas dan spesifisitas. Klasifikasi pengamatan yang digunakan dapat dilihat pada Tabel 1.

Tabel 1. Klasifikasi Pengamatan

Prediksi	Akurasi	
	Positif	**Negatif**
Positif	True Positive (TP)	False Positif (FP)
Negatif	False Negatif (FN)	True Negatif (TN)

Persentase banyaknya data yang diprediksi kelasnya benar diukur menggunakan akurasi, persentase banyaknya data yang prediksi kelas positif dan benar diukur menggunakan sensitivitas dan persentase banyaknya data yang prediksi kelas negatif dan benar diukur menggunakan spesifisitas. Ukuran akurasi, sensitivitas dan spesifisitas yang digunakan adalah sebagai berikut:

$$\text{Akurasi} = \left(\frac{TP+TN}{TP+FP+FN+TN}\right) \times 100\% \tag{4}$$

$$\text{Sensitivitas} = \left(\frac{TP}{TP+FN}\right) \times 100\% \quad (5)$$

$$\text{Spesifisitas} = \left(\frac{TN}{TN+FP}\right) \times 100\% \quad (6)$$

3. Hasil dan Pembahasan

Beberapa kondisi data training dan testing yang telah dicobakan menunjukkan bahwa kondisi tiga menghasilkan tingkat efektifitas yang lebih tinggi dibandingkan kondisi lainnya. Hasil akurasi, sensitivitas dan spesifisitas yang dihasilkan tiap metode pada berbagai kondisi secara rinci dapat dilihat pada Tabel 2.

Tabel 2. Efektifitas Metode Regresi Logistik Biner dan K-Nearest Neighbor pada Beberapa Kondisi Data Training dan Testing

Kondisi	Training	Testing	Metode	Akurasi	Sensitivitas	Spesifisitas
1	70% ($n = 210$)	30% ($n = 88$)	Regresi Logistik Biner	0.57	0.55	0.59
			K-Nearest Neighbor	0.60	0.48	0.72
2	75% ($n = 225$)	25% ($n = 73$)	Regresi Logistik Biner	0.67	0.69	0.66
			K-Nearest Neighbor	0.64	0.60	0.68
3	80% ($n = 239$)	20% ($n = 59$)	Regresi Logistik Biner	0.78	0.68	0.87
			K-Nearest Neighbor	0.71	0.71	0.71
4	85% ($n = 254$)	15% ($n = 44$)	Regresi Logistik Biner	0.68	0.62	0.74
			K-Nearest Neighbor	0.73	0.62	0.83

Kondisi data yang menghasilkan efektifitas lebih baik ditemukan pada persentase data training 80% dan data testing 20%. Pada kondisi ini, metode regresi logistik biner menghasilkan tingkat keakuratan 78% sedangkan metode KNN sebesar 71%. Dengan demikian besarnya tingkat kesalahan klasifikasi yang dihasilkan metode regresi logistik biner sebesar 22% sedangkan metode KNN sebenar 29%. Berdasarkan ukuran sensitivitas yang dihasilkan, metode regresi logistik biner dapat memprediksi wanita yang terkena menopause dini dengan benar sebesar 68% sedangkan metode KNN sebesar 71%. Berdasarkan ukuran spesifisitas yang dihasilkan, metode regresi logistik biner dapat memprediksi wanita yang tidak terkena menopause dini dengan benar sebesar 87% sedangkan metode KNN sebesar 71%.

4. Kesimpulan

Penelitian ini menghasilkan beberapa kesimpulan yaitu proporsi data training 80% dan testing 20% menghasilkan tingkat efektifitas yang terbaik. Secara umum akurasi model yang dihasilkan metode regresi logistik biner lebih tinggi dibandingkan metode k-nearest neighbor. Persentase yang dihasilkan metode KNN dalam ketepatan memprediksi wanita yang terkena menopause dini lebih tinggi dibandingkan metode regresi logistik biner.

Daftar Pustaka

[1] Brownlee. Jason, (2019), *Statistical Methods for Machine Learning*, Machine Learning Mastery.

[2] Dougherty. Geoff, (2013), *Pattern Recognition and Classification*, Springer, USA.

[3] Hosmer. DW, Lemeshow. JS, (2000), *Applied Logistic Regression.* John Wiley & Sons, Canada.

[4] Kusumawati. Eka Dewi, *et all*, (2021), Early Menopause: Reproductive Adaption of Javannese Women in West Papua, *The Conference International Guide Book and Abstract*, 79.

[5] Matloff. Norman, (2017), *Statistical Regression and Classification from Linear Models to Machine Learning*. Taylor & Francis Group, LLC.

[6] Nelso H. D, (2008), Menopause. *Journal Lancet*, **371**, 760 – 770.

[7] Nilson. Nils J, *et all*, (1996), *Introduction to Machine Learning*, Stanford University, CA

[8] World Health Organization, (1981), Research on the menopause. World Health Organization, Genewa.

[9] Zhu. D, Chung H. F, Dobson A. J, Pandeya. N, Giles G. G, Bruinsma. F, Brunner. E. J, Kuh. D, Hardy. R, Avis N. E, Gold E. B, Derby C. A, Matthews K. A, Cade J. E, Greenwood D. C, Demakakos. P, Brown D. E, Sievert L. L, Anderson. D, Hayashi. K, Lee J. S, Mizunuma. H, Tillin. T, Simonsen M. K, Adami H. O, Weiderpass. E, Mishra G. D, (2019), Age at natural menopause and risk of incident cardiovascular disease: a pooled analysis of individual patient data. *Journal Lanchet Public Health*, **4**, 55 –563.