Jurnal Riset Statistika (JRS)
e-ISSN 2798-6578 | p-ISSN 2808-3148
https://journals.unisba.ac.id/index.php/JRS

Tersedia secara online di
Unisba Press
https://publikasi.unisba.ac.id/

Metode Regresi Probit Biner untuk Pemodelan Faktor-Faktor yang Mempengaruhi Diagnosis Penyakit Jantung

Hasna, Anneke Iswani Achmad*

Prodi Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Islam Bandung, Indonesia.

A R T I C L E I N F O

Article history :
Received : 2/4/2022
Revised : 30/6/2022
Published : 8/7/2022

Volume : 2
No. : 1
Halaman : 28-34
Terbitan : **Juli 2022**

A B S T R A K

Analisis regresi merupakan suatu cara yang digunakan untuk menjelaskan hubungan fungsional antara variabel respon (Y) dengan variabel prediktor (X). Namun dalam kenyataan tidak jarang menggunakan data kualitatif yang berbentuk biner. Regresi probit biner merupakan suatu model regresi yang digunakan untuk menjelaskan hubungan antara variabel respon dengan satu atau lebih variabel prediktor, dimana pada variabel respon bersifat kualitatif sedangkan variabel prediktor bisa bersifat kuantitatif dan atau kualitatif. Model probit biner merupakan salah satu bentuk dari model *Generalized Linear Model* (GLM) yang digunakan untuk menganalisis hubungan antara satu variabel respon dengan beberapa variabel prediktor, dimana variabel responnya berupa data kuantitatif biner yang bernilai 0 dan 1. Estimasi parameter menggunakan metode *Maximum Likelihood*, dan diselesaikan dengan metode *Newton Raphson*. Tujuan penelitian ini adalah untuk pemodelan faktor-faktor yang mempengaruhi diagnosis penyakit jantung. Pada skripsi ini, data yang digunakan merupakan data sekunder mengenai analisa & dataset prediksi serangan jantung. Berdasarkan penelitian dapat diketahui faktor-faktor yang mempengauhi yaitu jenis kelamin, kolesterol, detak jantung maksimum, angina, dan penurunan segmen ST. Ketepatan klasifikasi sebesar 79,21% dengan nilai kesalahan klasifikasi 20,79%.

Kata Kunci : Regresi Probit Biner; Maxmimun likelihood; Penyakit Jantung.

A B S T R A C T

Regression analysis is a method used to explain the functional relationship between the response variable (Y) and the predictor variable (X). But in reality, it is not uncommon to use qualitative data in the form of binary. Binary probit regression is a regression model that is used to explain the relationship between the response variable and one or more predictor variables, where the response variable is qualitative while the predictor variable can be quantitative and/or qualitative. The binary probit model is a form of the Generalized Linear Model (GLM) model that is used to analyze the relationship between one response variable and several predictor variables, where the response variable is binary quantitative data with values of 0 and 1. Parameter estimation using the Maximum Likelihood method, and solved by Newton Raphson's method. The purpose of this study is to model the factors that influence the diagnosis of heart disease. In this thesis, the data used is secondary data regarding the analysis & prediction of heart attack datasets. Based on the research, it can be seen that the influencing factors are gender, cholesterol, maximum heart rate, angina, and a decrease in the ST segment. Classification accuracy is 79.21% with a misclassification value of 20.79%.

Keywords : Binary Probit Regression; Maximum likelihood; Heart Disease.

Corresponding Author : Email : annekeiswani11@gmail.com
Indexed : Garuda, Crossref, Google Scholar
DOI : https://doi.org/10.29313/jrs.vi.721

A. Pendahuluan

Statistika merupakan ilmu yang berperan sebagai sarana analisis dan interpretasi data memperoleh suatu kesimpulan. Dalam kehidupan di era globalisasi statistika sangat berperan dalam membantu kehidupan manusia. Statistika berperan untuk menjelaskan hubungan antara variabel-variabel yang berpengaruh terhadap suatu objek, salah satu metode yang digunakan adalah analisis regresi. Analisis regresi merupakan suatu cara yang digunakan untuk menjelaskan hubungan fungsional antara variabel respon (*Y*) dengan variabel prediktor (*X*). Hubungan fungsional antara variabel-variabel tersebut akan menjadi suatu persamaan regresi [1]. Pada analisis regresi biasanya variabel respon menggunakan data kuantitatif. Namun dalam kenyataan tidak jarang menggunakan data kualitatif yang berbentuk biner. Pada variabel respon bersifat kualitatif sedangkan variabel prediktor bisa bersifat kuantitatif dan atau kualitatif [2].

Analisis regresi yang digunakan untuk menjelaskan hubungan antara variabel respon dan prediktor, dimana variabel respon berupa data kualitatif yaitu Linear Probability Model, model logit dan model probit [3]. Linear Probability Model merupakan metode biner yang paling sederhana, namun memiliki kelemahan yaitu nilai peluang yang dihitung berada diluar 0 dan 1. Regresi logistik merupakan metode yang digunakan untuk menggambarkan hubungan antara variabel respon dengan satu atau lebih variabel prediktor. Regresi probit merupakan metode yang digunakan untuk menganalisis hubungan antara satu variabel respon dengan beberapa variabel prediktor, dimana variabel responnya berupa data kualitatif biner yang diasumsikan bernilai 0 yang menyatakan ketidakberadaan suatu karakteristik dan nilai 1 yang menyatakan keberadaan suatu karakteristik dan variabel prediktor bertipe kontinu dan atau diskrit berskala nominal dan atau biner. Berdasarkan penelitian terdahulu oleh Masitoh [4] yang menyatakan bahwa perbedaan antara regresi logistik dan probit yaitu berdasarkan *link function*. *Link function* pada regresi logistik biner adalah $\ln\left(\frac{\pi_i}{1-\pi_i}\right)$ dan pada regresi probit biner adalah $\Phi^{-1}(\pi_i)$. Regresi logistik menggunakan fungsi distribusi kumulatif dari distribusi logistik (*cumulative logistic function*), sedangkan regresi probit menggunakan fungsi distribusi kumulatif dari distribusi normal.

Regresi probit biner dapat mempermudah untuk manusia dalam berbagai macam bidang, salah satunya ialah membantu mengetahui faktor-faktor yang mempengaruhi diagnosis penyakit jantung. Penyakit jantung atau istilah medis penyakit jantung koroner harus diperhatikan, karena akan berakibat fatal dan mengakibatkan kematian. Bagi manusia jantung merupakan alat vital yang penting. Menurut *World Health Organization* (WHO) penyakit jantung tetap menjadi penyebab kematian pertama di tingkat global selama 20 tahun terakhir [5]. Penyakit jantung di Indonesia juga masih menjadi penyebab pertama dari seluruh kematian, angka kematiannya sebesar 26,4%. Penderita penyakit Jantung di Indonesia sekitar 2.754.064 individu [6].

Perlu diketahui faktor-faktor yang mempengaruhi diagnosis penyakit jantung untuk pencegahan, pengobatan serta memantau agar penyakit jantung tidak meningkat dan memburuk. Untuk lebih jelas dilakukan pengaplikasian ilmu statistika melalui pemodelan faktor-faktor yang mempengaruhi diagnosis penyakit jantung. Variabel respon pada penelitian ini adalah diagnosis penyakit jantung kategori lebih kecil kemungkinan terkena serangan jantung diberi nilai 0, dan kategori lebih besar kemungkinan terkena serangan jantung diberi nilai 1. Variabel prediktor yang digunakan adalah usia, kenis kelamin, tekanan darah, kolesterol, gula darah puasa, hasil elektrokardiogram (EKG), detak jantung (*heart rate*) maksimum, Angina, dan penurunan segmen ST.

Berdasarkan latar belakang yang telah diuraikan, maka perumusan masalah dalam penelitian ini sebagai berikut: "bagaimana pemodelan regresi probit biner berdasarkan faktor-faktor yang mempengaruhi diagnosis penyakit jantung?" dan "Bagaimana ketepatan klasifikasi faktor-faktor yang mempengaruhi diagnosis penyakit jantung?". Selanjutnya, tujuan dalam penelitian ini diuraikan dalam pokok-pokok yaitu: (1) Untuk memperoleh model regresi probit biner berdasarkan faktor-faktor yang mempengaruhi diagnosis penyakit jantung; (2) Menentukan ketepatan klasifikasi faktor-faktor yang mempengaruhi diagnosis penyakit jantung.

B. Metode Penelitian

Peneliti menggunakan metode regresi probit biner. Data yang digunakan merupakan data sekunder yang didapatkan dari *website* kaggle.com mengenai *Heart Attack Analysis & Prediction Dataset.*

Variabel yang digunakan terdapat sepuluh variabel hasil reduksi, yang terdiri dari satu variabel respon dan Sembilan variabel prediktor. Berikut ini variabel-variabel yang digunakan dalam penelitian ini:

Tabel 1. Variabel Penelitian

Variabel	Nama Variabel	Tipe	Keterangan
Y	Diagnosis penyakit jantung	Kategori	0 = Lebih kecil kemungkinan terkena serangan jantung 1 = Lebih besar kemungkinan terkena serangan jantung
x_1	Usia	Kontinu	-
x_2	Jenis Kelamin	Kategori	0 = Perempuan 1 = Laki-laki
x_3	Tekanan darah (mm/Hg)	Kontinu	-
x_4	Kolesterol (mg/dL)	Kontinu	-
x_5	Gula darah puasa	Kategori	0 = di bawah 119 mg/dL 1 = di atas sama dengan 120 mg/dL
x_6	Hasil elektrokardiogram (EKG)	Kategori	0 = Normal 1 = Tidak normal
x_7	Detak jantung (*heart rate*) maksimum	Kontinu	-
x_8	Angina	Kategori	0 = Akibat olahraga 1 = Bukan akibat olahraga
x_9	Penurunan segmen ST	Kontinu	-

Langkah analisis yang digunakan dalam penelitian ini menggunakan metode regresi probit biner sebagai berikut ini: (1) Mendeteksi multikolinearitas pada variabel prediktor dengan menghitung nilai VIF; (2) Melakukan analisis regresi probit biner sebagai berikut ini: (1) Membuat model regresi probit biner untuk mengidentifikasi variabel prediktor terhadap variabel respon, berdasarkan pengujian parameter secara serentak dan parsial; (2) Menghitung peluang regresi probit biner; (3) Interpretasi model regresi probit biner dengan efek marginal; (4) Pengujian kesesuaian model probit biner dengan rumus; (5) Mengukur kebaikan model melalui ketepatan klasifikasi dengan rumus; (6) Membuat kesimpulan.

C. Hasil dan Pembahasan

Uji Multikolinieritas

Dilakukan uji multikolinieritas untuk mengetahui apakah terdapat hubungan yang linear antara variabel-variabel prediktor. Hasil uji multikolinearitas menggunakan *software* RStudio sebagai berikut:

Tabel 2. Hasil Uji Multikolinearitas

Variabel	Nilai VIF	Keputusan
X_1	1,3896	Tidak Terdapat Multikolinearitas
X_2	1,0913	Tidak Terdapat Multikolinearitas
X_3	1,1652	Tidak Terdapat Multikolinearitas
X_4	1,1298	Tidak Terdapat Multikolinearitas
X_5	1,0493	Tidak Terdapat Multikolinearitas
X_6	1,0633	Tidak Terdapat Multikolinearitas

Lanjutan Tabel 2. Hasil Uji Multikolinearitas

Variabel	Nilai VIF	Keputusan
X_7	1,4868	Tidak Terdapat Multikolinearitas
X_8	1,2485	Tidak Terdapat Multikolinearitas
X_9	1,2298	Tidak Terdapat Multikolinearitas

Berdasarkan Tabel 2 nilai VIF kurang dari 10, maka tidak terdapat multikolinearitas pada masing-masing variabel prediktor. Artinya tidak terdapat hubungan yang linear antara variabel-variabel prediktor, maka dapat dilakukan pemodelan regresi probit biner menggunakan seluruh variabel.

Analisis Regresi Probit Biner

Sebelum melakukan pemodelan regresi probit biner dilakukan pemeriksaan uji signifikan secara serentak. Pengujian ini dilakukan untuk melihat keberartian koefisien β dengan variabel-variabel prediktor secara keseluruhan. Pada metode ini uji serentak dilakukan dengan uji *likelihood ratio test* dengan α = 5%, dengan hipotesis dari uji signifikan parameter β secara serentak adalah sebagai berikut $H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = \beta_9 = 0$; H_1: minimai ada satu $\beta_j \neq 0$; dimana j=1,2,…,p

Hasil pengujian parameter serentak dari s*oftware* RStudio sebagai berikut:

Tabel 3. Hasil Pengujian Parameter Secara Serentak

G^2	$\chi^2_{(9;0,05)}$	Keputusan
145.5099	16,919	Tolak H_0

Karena $G^2 > \chi^2_{(5;0,05)} = 16{,}919$ maka H_0 ditolak, artinya minimal ada satu variabel prediktor yang signifikan memberikan pengaruh terhadap diagnosis penyakit jantung.

Setelah pengujian parameter serentak dilakukan pengujian parameter β parsial untuk mengetahui variabel prediktor mana yang mempengaruhi terhadap variabel respon. Dilakukan uji Wald dengan hipotesis sebagai berikut: (1) $H_0: \beta_j = 0$; dimana j=1,2,…,p; (2) $H_1: \beta_j \neq 0$; dimana j=1,2,…,p

Hasil pengujian parameter parsial dari *software* sebagai berikut:

Tabel 4. Hasil Pengujian Parameter Secara Parsial

Variabel	Koeifisien	SE Koefisien	*Wald*	P-*Value*	Keputusan
X_2	-0,9406	0,199966	-4,704	0,00000255	Tolak H_0
X_4	-0,0042	0,001727	-2,408	0,016	Tolak H_0
X_7	0,0183	0,004354	4,203	0,0000263	Tolak H_0
X_8	-0,8241	0,193938	-4,249	0,0000214	Tolak H_0
X_9	-0,3843	0,083484	-4,603	0,00000417	Tolak H_0

Dengan menggunakan taraf signifikan α = 5%, maka berdasarkan nilai mutlak uji Wald lebih kecil dari nilai $|Z_{0,05/2}| = 1{,}96$ atau nilai p-*value* kurang dari α = 5%, maka terima H_0. Berdasarkan Tabel 4.4 didapatkan lima variabel prediktor yang signifikan yaitu X_2, X_4, X_7, X_8, dan X_9. Variabel prediktor yang signifikan terhadap diagnosis penyakit serangan jantung adalah jenis kelamin, kolesterol, detak jantung maksimum, angina, dan penurunan segmen ST. Variabel prediktor yang signifikan yang akan digunakan untuk membentuk model regresi probit biner.

Pemodelan Regresi Probit Biner

Setelah menguji parameter secara serentak dan parsial, dilakukan pemodelan terhadap variabel-variabel yang signifikan. Berikut ini model regresi dengan variabel jenis kelamin, kolesterol, detak jantung maksimum, angina, dan penurunan segmen ST:

$$Y^* = -0{,}2967 - 0{,}94063X_2 - 0{,}0042X_4 + 0{,}0183X_7 - 0{,}8241X_8 - 0{,}38434X_9$$

Persamaan terbentuk akan membentuk model regresi probit biner, model regresi probit biner dengan persamaan probabilitas pasien dalam kategori lebih kecil kemungkinan terkena serangan jantung adalah sebagai berikut:

$$P(Y = 0) = \Phi(\gamma - \beta^T X) = \Phi(-0{,}2967 - 0{,}9406X_2 - 0{,}0042X_4 + 0{,}0183X_7 - 0{,}8241X_8 - 0{,}3843X_9)$$

Persamaan model regresi probit biner dengan persamaan probabilitas pasien dalam kategori lebih besar kemungkinan terkena serangan jantung adalah sebagai berikut:

$$P(Y = 1) = 1 - \Phi(\gamma - \beta^T X) = 1 - \Phi(-0{,}2967 - 0{,}9406X_2 - 0{,}0042X_4 + 0{,}0183X_7 - 0{,}8241X_8 - 0{,}3843X_9)$$

Persamaan efek marginal pada variabel kolesterol terhadap masing-masing kategori diagnosis penyakit jantung sebagai berikut:

$$\frac{\partial P(Y = 0)}{\partial X_4} = -0{,}0042\phi(-0{,}2967 - 0{,}9406x_2 - 0{,}0042x_4 + 0{,}0183x_7 - 0{,}8241x_8 - 0{,}3843x_9)$$

$$\frac{\partial P(Y = 1)}{\partial X_4} = 0{,}0042\phi(-0{,}2967 - 0{,}9406x_2 - 0{,}0042x_4 + 0{,}0183x_7 - 0{,}8241x_8 - 0{,}3843x_9)$$

Dengan persamaan diatas, dapat dihitung besar pengaruh kolesterol terhadap diagnosis penderita penyakit jantung. Pada salah satu pasieun laki-laki dengan kadar kolesterol 233mg/dL, detak jantung maksimum sebesar 150, angina akibat olahraga, dan terjadi penurunan segmen ST sebesar 2,3 dengan efek marginal nya sebagai berikut:

$$\frac{\partial P(Y = 0)}{\partial x_4} = -0{,}0042\phi(-0{,}2967 - 0{,}9406(1) - 0{,}0042(233) + 0{,}0183(150) - 0{,}8241((0{,}3843(2{,}3))$$

$$= -0{,}0042\left[\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{0{,}3548^2}{2}\right)\right]$$

$$= -0{,}0016$$

$$\frac{\partial P(Y = 1)}{\partial x_4} = 0{,}0042\phi(-0{,}2967 - 0{,}9406(1) - 0{,}0042(233) + 0{,}0183(150) - 0{,}8241(0) - 0{,}3843(2{,}3))$$

$$= 0{,}0042\left[\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{0{,}3548^2}{2}\right)\right]$$

$$= 0{,}0016$$

Untuk mengetahui model yang digunakan uji kesesuaian dengan hipotesis sebagai berikut: H_0 : tidak terdapat perbedaan antara hasil prediksi dengan hasil observasi dan H_1 : terdapat perbedaan antara hasil prediksi dengan hasil observasi

Dengan perhitungan *software* RStudio maka diperoleh nilai D=227,9. Digunakan taraf signifikan α = 5% dengan nilai db = 297 maka nilai $\chi^2_{(297,0,05)} = 338{,}193$. Karena $D < \chi^2_{(db,\alpha)}$ maka terima H_0, dapat

disimpulkan bahwa model sesuai artinya tidak terdapat perbedaan yang signifikan antara hasil prediksi model dengan hasil observasi.

Alat ukur untuk mendapatkan model terbaik yaitu pengukuran dengan ketepatan klasifikasi yang mampu memprdiksi secara akurat. Ketepatan klasifikasi dihitung melalui perhitungan APER. Berikut ini hasil pengelompokan data aktual dengan data prediksi:

Tabel 5. Tabulasi Silang Klasifikasi Aktual dan Hasil Prediksi Model

Kelompok Aktual	Kelompok Prediksi		Total
	Y = 0	Y = 1	
Y = 0	96	42	138
Y = 1	24	141	165

Berdasarkan tabel tersebut dapat diketahui bahwa 96 pasien yang didiagnosis lebih kecil terkena terkena serangan jantung diprediksi benar oleh model dan terdapat 141 pasien yang didiagnosis lebih besar terkena kemungkinan terkena serangan jantung diprediksi benar oleh model. Tingkat ketepatan klasifikasi dapat dihitung berdasarkan rumus (2.22) sebagai berikut:

$$\text{APER} = \frac{42+24}{138+165} \times 100\% = \frac{66}{303} \times 100\% = 0{,}2178$$

Ketepatan klasifikasi = 1- 0,2178 = 78,22%

Bedasarkan perhitungan tersebut, diperoleh nilai ketepatan klasifikasi diagnosis penyakit jantung dari model regresi probit biner sebesar 79,21% dengan tingkat kesalahan klasifikasi sebesar 20,79%. Ketepatan klasifikasi model probit biner.

D. Kesimpulan

Berdasarkan analisis dan pembahasan mengenai diagnosis penyakit jantung menggunakan metode regresi probit biner, diperoleh kesimpulan seperti, pemodelan regresi probit biner menghasilkan lima variabel prediktor yang signifikan, yaitu variabel jenis kelamin, variabel kolesterol, variabel hasil elektrokardiogram (EKG), variabel detak jantung maksimum, variabel angina, dan variabel penurunan segmen ST. Model regresi probit biner sebagai berikut:

$$P(Y = 0) = \Phi(-0{,}2967 - 0{,}9406X_2 - 0{,}0042X_4 + 0{,}0183X_7 - 0{,}8241X_8 - 0{,}3843X_9)$$

$$P(Y = 1) = 1 - \Phi(-0{,}2967 - 0{,}9406X_2 - 0{,}0042X_4 + 0{,}0183X_7 - 0{,}8241X_8 - 0{,}3843X_9)$$

Nilai ketepatan klasifikasi adalah sebesar 78,22% dengan nilai kesalahan klasifikasi 21,78%. Artinya bahwa akurasi model probit biner terbentuk untuk mendiagnosis penyakit jantung yang tepat sebesar 78,22%

Daftar Pustaka

[1] Sudjana, *Metode statistika*. Bandung: Tarsito, 2002.

[2] M. Damayanti CR and T. S. Yanti, "Regresi Poisson Invers Gaussian (PIG) untuk Pemodelan Jumlah Kasus Pneumonia pada Balita di Provinsi Jawa Tengah Tahun 2019," *J. Ris. Stat.*, vol. 1, no. 2, pp. 143–151, Feb. 2022, doi: 10.29313/jrs.v1i2.523.

[3] A. Agresti, *Categorical Data Analysis, 2nd Edition.* New York, 2003.

[4] F. Masitoh and V. Ratnasari, "Pemodelan Status Ketahanan Pangan di Provinsi Jawa Timur dengan Pendekatan Metode Regresi Probit Biner," vol. 5, 2016, doi: 10.12962/j23373520.v5i2.16549.

[5] WHO, "WHO reveals leading causes of death and disability worldwide: 2000-2019," 2020. https://www.who.int/news/item/09-12-2020-who-reveals-leading-causes-of-death-and-disability-worldwide-2000-2019#:~:text=Heart disease has remained the,nearly 9 million in 2019.

[6] P2PTM Kemenkes RI, "Yuk, kenali apa itu penyakit jantung koroner (PJK)?," 2021. http://p2ptm.kemkes.go.id/infographic-p2ptm/hipertensi-penyakit-jantung-dan-pembuluh-darah/yuk-kenali-apa-itu-penyakit-jantung-koroner-pjk#:~:text=22 Desember 2021-,Yuk%2C kenali apa itu penyakit jantung koroner (PJK)%3F,dinding pembuluh darah (Ateroskler (accessed Apr. 03, 2020).