

Analisis Perbandingan Model Regresi Logistik Dan Probit Dengan K-Fold Cross Validation Dalam Mengidentifikasi Faktor Signifikan Pada Penyakit Diabetes Melitus

Aqilla Khairunnisa ^{a,1,*}

^{*a} Universitas Ahmad Dahlan, Yogyakarta, Indonesia;

¹ aqilla2000015033@webmail.uad.ac.id

^{*}Correspondent Author

Received:

Revised:

Accepted:

KATAKUNCI

Analisis Regresi Logistik
Analisis Regresi Probit
Algoritma BHHH
k-fold cross validation

ABSTRAK

Statistika adalah cabang matematika yang berkaitan dengan pengumpulan, analisis, interpretasi, presentasi, dan organisasi data. Statistika digunakan dalam berbagai disiplin ilmu untuk membuat keputusan berdasarkan data. Terdapat dua jenis utama statistika, yaitu statistika deskriptif dan statistika inferensial. Salah satu metode statistika inferensial yang biasa digunakan adalah Analisis Regresi Logistik dan Regresi Probit. Regresi Logistik dan Regresi Probit merupakan teknik dalam statistika inferensial yang digunakan untuk menemukan hubungan di antara variabel prediktor dan variabel respons yang bersifat dikotomis (memiliki dua kategori) atau polikotomis (memiliki lebih dari dua kategori). Regresi logistik menggunakan fungsi distribusi kumulatif dari distribusi logistik sedangkan Regresi Probit menggunakan fungsi distribusi kumulatif dari distribusi normal. Tujuan dari penelitian ini adalah mengetahui model terbaik antara model logit dan model probit berdasarkan validasi model menggunakan k-fold cross validation. Data yang digunakan adalah data sekunder tentang prediksi diabetes yang tersedia dalam Kaggle. Berdasarkan hasil yang diterapkan pada data tersebut didapatkan faktor faktor yang berpengaruh signifikan terhadap penyakit diabetes adalah jenis kelamin, usia, riwayat hipertensi, riwayat penyakit jantung, riwayat merokok, BMI, kadar HbA1c, dan kadar gula darah. Hasil perbandingan model didapatkan dari rata-rata akurasi yang sama menggunakan k-fold cross validation untuk model logit dan probit yaitu sebesar 93.7%. Perbandingan ini diperkuat dengan empat kriteria dalam pemilihan model terbaik yaitu AIC, Pseudo-R², AUC, Logloss keakuratan klasifikasi, dan uji kesesuaian model (Goodness of fit). Secara keseluruhan dapat disimpulkan bahwa model probit lebih baik daripada model logit dalam kasus data tersebut.

KEYWORDS

Logistic Regression Analysis
Probit Regression Analysis
BHHH algorithm
k-fold cross validation

ABSTRACT

Statistics is a discipline of mathematics concerned with the collection, analysis, interpretation, presentation, and organization of data. Statistics is used in various disciplines to make decisions based on data. There are two general types of statistics, descriptive statistics and inferential statistics. One of the commonly used inferential statistical methods is Logistic Regression Analysis and Probit Regression. Logistic

regression and Probit regression are techniques in inferential statistics used to find the relationship between predictor variables and response variables that are dichotomous (have two categories) or polycotomous (have more than two categories). Logistic regression uses the cumulative distribution function of the logistic distribution while Probit regression uses the cumulative distribution function of the normal distribution. The purpose of this study is to determine the best model between the Logistic Regression model and the Probit Regression model based on model validation using k-fold cross validation. The data used is secondary data on diabetes prediction available in Kaggle. Based on the results applied to the data, the factors that have a significant effect on diabetes are gender, age, history of hypertension, history of heart disease, smoking history, BMI, HbA1c levels, and blood sugar levels. The results of the model comparison showed the same average accuracy using k-fold cross validation for logistic regression and probit regression models, which was 93.7%. This comparison is supported by four criteria in selecting the best model, namely AIC, Pseudo-R², AUC, Logloss, classification accuracy, and goodness of fit test. Overall, it can be concluded that the Probit Regression model is better than the Logistic Regression model in the case of these data.

This is an open-access article under the [CC-BY-SA](#) license.



Pendahuluan

Statistika merupakan cabang ilmu matematika yang berfokus pada pengumpulan, analisis, interpretasi, dan penyajian data. Dalam aplikasinya, statistika terbagi menjadi dua, yaitu statistika deskriptif dan inferensial. Salah satu teknik dalam statistika inferensial yang paling umum digunakan adalah analisis regresi, yang bertujuan untuk memahami hubungan antara variabel dependen (respon) dengan satu atau lebih variabel independen (prediktor) [1]. Model regresi tidak hanya digunakan untuk mengukur kekuatan hubungan antarvariabel, tetapi juga sebagai alat prediksi dalam berbagai bidang, termasuk kesehatan [2].

Di antara jenis-jenis regresi, regresi logistik dan regresi probit banyak digunakan saat variabel respon bersifat biner atau ordinal. Regresi logistik menggunakan fungsi distribusi logistik kumulatif (CDF), sedangkan regresi probit menggunakan fungsi distribusi normal kumulatif [2]. Kedua model ini dapat digunakan untuk mengidentifikasi faktor-faktor signifikan yang berkontribusi terhadap suatu kejadian atau kondisi, seperti penyakit.

Penyakit Tidak Menular (PTM), khususnya diabetes melitus (DM), menjadi masalah kesehatan global yang terus meningkat. Data WHO menunjukkan bahwa lebih dari 40 juta kematian setiap tahunnya disebabkan oleh PTM, seperti penyakit kardiovaskular dan diabetes [3]. Di Indonesia, prevalensi diabetes mencapai angka 10 juta jiwa dan menempatkan

Indonesia sebagai negara dengan peringkat ketujuh tertinggi di dunia [4]. Diabetes melitus merupakan gangguan kronis yang memengaruhi metabolisme glukosa akibat gangguan produksi atau fungsi insulin [5]. Oleh karena itu, identifikasi faktor-faktor signifikan penyebab diabetes sangat penting dilakukan sebagai upaya pencegahan dan penanganan dini.

Beberapa penelitian terdahulu menunjukkan keberhasilan penerapan regresi logistik dan probit dalam mengidentifikasi faktor risiko terhadap suatu kondisi kesehatan. Penelitian di Puskesmas Pijorkoling (2020) menggunakan regresi logistik biner dan menemukan bahwa konsumsi makanan asin, pola tidur, dan kebiasaan olahraga berhubungan signifikan dengan kejadian hipertensi pada lansia [6]. Di Kota Jambi, regresi logistik biner digunakan untuk mengkaji risiko kejadian diare pada anak balita, dengan hasil bahwa status imunisasi, sanitasi makanan, dan ketersediaan jamban merupakan faktor signifikan [7].

Pada kasus penyakit jantung, regresi probit digunakan untuk memodelkan pengaruh variabel seperti jenis kelamin, angina, kadar kolesterol, detak jantung maksimum, dan perubahan segmen ST, yang semuanya terbukti signifikan [8]. Penelitian lain di Jawa Timur menggunakan regresi probit biner untuk mengidentifikasi faktor yang memengaruhi status stunting pada balita, dengan hasil bahwa pemberian ASI yang tidak eksklusif serta pendidikan ibu merupakan faktor utama [9]. Di Jawa Barat, regresi logistik dan probit biner diterapkan untuk mengevaluasi pencapaian target unmet need pada pasangan usia subur, dan ditemukan bahwa tingkat pendidikan kepala keluarga, jumlah anak, serta metode kontrasepsi berhubungan signifikan [10].

Berdasarkan studi-studi sebelumnya, regresi logistik biner dan regresi probit biner terbukti efektif dalam mengidentifikasi faktor signifikan terhadap kondisi kesehatan. Mengingat pentingnya diabetes melitus sebagai penyakit yang serius dan meningkat, penelitian ini bertujuan untuk membandingkan performa model regresi logistik dan probit dalam mengidentifikasi faktor signifikan penyebab diabetes melitus dengan menggunakan pendekatan K-Fold Cross Validation untuk validasi model.

Metode

Langkah awal penelitian ini yaitu melakukan proses preprocessing data yang mencakup analisis statistik deskriptif, penyeimbangan data menggunakan metode SMOTE-ENN, serta uji normalitas dan asumsi non-multikolinearitas. Selanjutnya diterapkan regresi logistik biner dengan menghitung estimasi parameter, melakukan uji parameter secara serentak (Likelihood Ratio Test) dan parsial (Wald Test), serta menguji kecocokan model menggunakan Hosmer-Lemeshow Test. Model logit yang terbentuk kemudian divalidasi menggunakan k-fold cross validation. Tahapan yang sama juga dilakukan pada regresi probit biner, mulai dari estimasi

parameter, pengujian parameter, pengujian kecocokan model, hingga validasi model, sehingga dapat diperoleh perbandingan performa antara kedua model dalam memprediksi kejadian diabetes melitus.

Hasil dan Pembahasan

1. Analisis Deskriptif Statistik

Penelitian ini menggunakan data prediksi diabetes dari situs Kaggle. Terdapat 100.000 observasi dalam kasus ini. Variabel yang digunakan meliputi variabel dependen/respon dan variabel independen/prediktor. Variabel dependen adalah prediksi penyakit diabetes melitus, sedangkan variabel independen adalah faktor-faktor yang memengaruhi risiko penyakit diabetes. Berikut adalah tabel deskripsi statistik dari data diabetes:

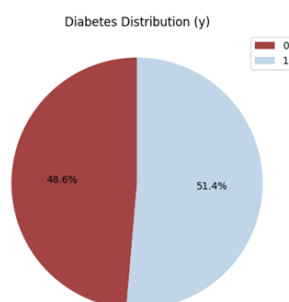
Tabel 1. Deskriptif Statistik

<i>Variabel</i>	<i>Min</i>	<i>Max</i>	<i>Mean</i>	<i>Median</i>	<i>Variance</i>	<i>Std</i>
x_2	0,08	80	41,80	43	504,60	22,46
x_6	10,01	95,69	27,32	27,32	45,80	6,77
x_7	3,5	9	5,53	5,8	1,15	1,07
x_8	80	300	138,22	140	1673,72	40,91

Berdasarkan Tabel 1 memperlihatkan ciri-ciri masing-masing variabel prediktor yang merupakan tipe rasio. Variabel prediktor usia (x_2) mempunyai nilai terkecil sebesar 0,08 dan nilai terbesar yaitu 80. Hal ini mengindikasikan rentang usia individu bervariasi dari 1 bulan hingga 80 tahun. Selain itu, variansi usia (x_2) sebesar 507,01, yang menunjukkan variasi usia yang signifikan berdasarkan nilai varian yang tinggi.

2. Hybridsampling SMOTE-ENN

Hasil statistik deskriptif pada Tabel 1, diketahui bahwa orang yang terkena penyakit diabetes (kategori 0) pada variabel respon y lebih sedikit dibandingkan dengan orang yang tidak terkena penyakit diabetes (kategori 1). Penerapan metode hybridsampling yaitu SMOTE-ENN bekerja dengan membuat sampel sintesis dari kelas minoritas dan menghapus sampel yang ditandai noise. Hasil dari penerapan dari SMOTE-ENN ditunjukkan dengan gambar 1.



Gambar 1. Data setelah dilakukan SMOTE-ENN

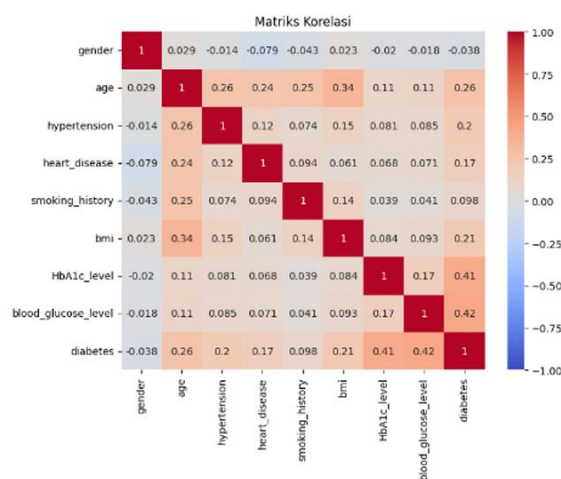
Berdasarkan Gambar 1 diketahui persentase kategori antara terkena diabetes dan tidak terkena diabetes masing masing yaitu 51,4% dan 48,6% dengan jumlah observasi 154.822 data. Jumlah data yang terkena diabetes setelah dilakukan metode SMOTE-ENN menjadi 79.528 data. Sedangkan jumlah data tidak terkena diabetes setelah dilakukan SMOTE-ENN menjadi 75.294 data.

3. Uji Normalitas

Berdasarkan uji Kolmogorov-Smirnov dapat diketahui bahwa data prediksi penyakit diabetes tidak mengikuti distribusi normal sehingga regresi linear tidak cocok dalam data ini. Hal ini disebabkan oleh $p\text{-value} < 0,05$. Maka dari itu, data tersebut dapat diuji dengan metode regresi logistik biner dan regresi probit biner. Pada studi kasus, terdapat 154.822 sampel, dengan 75.924 di antaranya termasuk dalam kategori 0 dan 79.528 di antaranya termasuk dalam kategori 1.

4. Uji Multikolinearitas

Metode yang dapat digunakan untuk mengetahui adanya multikolinearitas, antara lain dengan memperoleh nilai koefisien korelasi Pearson (r_{ij}) dan nilai Variance Inflation Factor (VIF). Nilai korelasi antar variabel predictor disajikan pada Gambar 2. Sedangkan untuk memeriksa Variance Inflation Factor (VIF) ditampilkan pada Tabel 2.



Gambar 2. Matriks Korelasi antar Variabel Prediktor

Berdasarkan Gambar 2 menggambarkan bahwa nilai korelasi antara semua variabel prediktor tidak melebihi 0,85, menunjukkan bahwa korelasi tersebut tidak signifikan. Maka dari itu, tidak ada kolinearitas antara variabel prediktor.

Tabel 2. Nilai VIF Variabel Prediktor

<i>Variabel Prediktor</i>	<i>VIF</i>
x_1	1,022644
x_2	1,396945
x_3	1,092557
x_4	1,092190
x_5	1,073114
x_6	1,153447
x_7	1,233966
x_8	1,192298

Dalam Tabel 2 menunjukkan tidak ada multikolinearitas antara variabel variabel prediktor. Hal ini dibuktikan dengan masing-masing variabel prediktor mempunyai nilai VIF < 10. Dengan demikian, semua variabel prediktor dapat dimasukkan ke dalam analisis regresi logistik biner dan regresi probit biner.

5. Hasil Analisis Regresi Logistik Biner

Prosedur klasifikasi menggunakan regresi logistik meliputi estimasi parameter model logistik, uji signifikansi parameter, dan validasi model.

a. Estimasi Parameter Model Logit

Pada kasus ini, variabel respon berskala biner untuk model regresi menggunakan kategori variabel diabetes (y) yang dikategorikan $y = 0$ “tidak terkena diabetes” dan $y = 1$ “terkena diabetes”. Salah satu variabel respon digunakan sebagai pembanding, sehingga dari kedua kategori tersebut akan diperoleh satu fungsi logit. Kategori yang digunakan sebagai pembanding dalam kasus ini adalah kategori untuk $y = 1$. Sehingga estimasi model regresi logit biner yang diharapkan berdasarkan persamaan berikut:

$$\pi = P(y_i = 0|x_i) = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8)} \quad (1)$$

Dan $1 - \pi_i = p(y_i = 1|x_i)$

Nilai estimasi parameter diperoleh dengan menggunakan BHHH (Berndt, Hall, Hall, and Hausman). Berikut diberikan nilai estimasi koefisien β pada setiap variabel prediktor untuk model logit dalam Tabel 3. Berdasarkan parameter yang terdapat pada tabel 3, dilakukan substitusi kedalam persamaan 1. Hasil perhitungan dapat disimpulkan bahwa sampel pertama memberikan nilai peluang untuk model logit 42 43 sebesar 0,085 untuk kategori 0 dan 0,915 untuk kategori 1. Artinya, sampel pertama berpeluang besar terkena penyakit diabetes sebesar 91,5% sedangkan peluang tidak terkena penyakit diabetes hanya sebesar 8,5%.

Tabel 3. Estimasi Parameter Model Logit

<i>Parameter</i>	<i>Nilai Parameter</i>	<i>Nilai Statistik Wald</i>	<i>p-value</i>
$\widehat{\beta}_0$	-18,529	17.041,296	0,000
$\widehat{\beta}_1$	-0,381	194,761	0,000
$\widehat{\beta}_2$	6,172	7.579,088	0,000
$\widehat{\beta}_3$	1,051	668,540	0,000
$\widehat{\beta}_4$	0,682	149,244	0,000
$\widehat{\beta}_5$	0,352	128,368	0,000
$\widehat{\beta}_6$	13,632	4.592,006	0,000
$\widehat{\beta}_7$	17,052	10.412,213	0,000
$\widehat{\beta}_8$	9,309	9.105,188	0,000

b. Uji Signifikansi Parameter

1) Uji Simultan

Berdasarkan uji simultan diatas, dapat dilihat nilai G sebesar 134.860,98, dimana nilai $G \geq \chi^2_{(0,05;8)}$ yaitu $134.860,98 > 15,51$, dapat diambil keputusan H_0 ditolak. Dengan demikian, dapat diambil kesimpulan bahwa semua variabel prediktor memiliki pengaruh signifikan terhadap penyakit diabetes secara keseluruhan.

2) Uji Parsial

Hasil analisis untuk variabel prediktor lainnya dapat dilihat pada Tabel 2. Dari Tabel 2, terlihat bahwa setiap variabel prediktor secara statistik signifikan memengaruhi variabel respon. Hal ini ditunjukkan oleh p-value masing-masing variabel prediktor yang lebih kecil dari $\alpha = 0,05$ atau membandingkan nilai Chi-Square tabel yaitu $\chi^2_{(0,05;1)} = 3,84$.

c. Validasi Model

Validasi model digunakan untuk mengetahui performa machine learning dalam membentuk klasifikasi. Metode yang akan digunakan untuk validasi model logit adalah K-Fold Cross Validation. Nilai k yang dipilih adalah 10 dan dilakukan pengulangan pengujian sebanyak 10 kali yang ditunjukkan pada tabel 3.

Tabel 3. Hasil K-Fold Cross Validation

<i>Fold</i>	<i>Accuracy</i>	<i>Sensitivity</i>	<i>Specificity</i>
1	0,9375	0,9409	0,9376
2	0,9360	0,9368	0,9385
3	0,9384	0,9418	0,9386
4	0,9376	0,9418	0,9371
5	0,9373	0,9354	0,9421
6	0,9342	0,9366	0,9355
7	0,9389	0,9413	0,9399
8	0,9347	0,9339	0,9387

9	0,9403	0,9418	0,9420
10	0,9361	0,9389	0,9369
Rata-Rata	0,9371	0,9389	0,9387

Berdasarkan Tabel 3, Proses klasifikasi menggunakan model logit dengan 10 kali pengulangan diperoleh bahwa nilai akurasi berkisar 93%-94%, dengan nilai sensitivitas 94%, dan nilai spesitifitas yang sangat tinggi yaitu 93%-94%. Hal ini menunjukkan bahwa model logit mampu mengklasifikan penentuan prediksi penyakit diabetes dengan sangat baik di setiap bagian data.

6. Hasil Analisis Regresi Probit Biner

Prosedur klasifikasi menggunakan regresi probit meliputi estimasi parameter model logistik, uji signifikansi parameter, dan validasi model.

a. Estimasi Parameter Model Probit

Dalam kasus ini, variabel respon berskala biner untuk model regresi menggunakan kategori variabel diabetes (y) yang dikategorikan $y = 0$ "tidak terkena diabetes" dan $y = 1$ "terkena diabetes". Salah satu variabel respon digunakan sebagai pembanding, sehingga dari kedua kategori tersebut akan diperoleh satu fungsi probit. Kategori yang digunakan sebagai pembanding dalam kasus ini adalah kategori untuk $y = 1$. Sehingga estimasi model probit yang diharapkan berdasarkan persamaan berikut.

$$\pi_i = \phi(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8) \quad (2)$$

$$\text{Dan } 1 - \pi_i = p(y_i = 1 | x_i)$$

Berdasarkan parameter yang terdapat pada tabel 2, dilakukan substitusi kedalam persamaan 2. Hasil perhitungan dapat disimpulkan sampel pertama memberikan nilai peluang untuk model logit sebesar 0,085 untuk kategori 0 dan 0,915 untuk kategori 1. Artinya, sampel pertama berpeluang besar terkena penyakit diabetes sebesar 91,5% sedangkan peluang tidak terkena penyakit diabetes hanya sebesar 8,5%.

b. Uji Signifikansi Parameter

1) Uji Simultan

Berdasarkan uji simultan diatas, dapat dilihat nilai G sebesar 134.873,76, dimana nilai $G \geq \chi^2_{(0,05;8)}$ yaitu $134.873,76 > 15,51$, dapat diambil keputusan H_0 ditolak. Dengan demikian, dapat diambil kesimpulan bahwa semua variabel prediktor memiliki pengaruh signifikan terhadap penyakit diabetes secara keseluruhan.

2) Uji Parsial

Hasil analisis untuk variabel prediktor lainnya dapat dilihat pada Tabel 4.7. Dari Tabel 4.7, terlihat bahwa setiap variabel prediktor secara statistik signifikan memengaruhi variabel respon. Hal ini ditunjukkan oleh p-value masing-masing variabel prediktor yang lebih kecil

dari $\alpha=0,05$ atau membandingkan nilai Chi-Square tabel yaitu $\chi^2 (0,05;1)=3,84$, nilai ini dapat dibandingkan dengan masing masing nilai pada Tabel 4.7.

c. Validasi Model

Validasi model bertujuan mengevaluasi performa machine learning dalam membentuk klasifikasi. Metode yang akan digunakan untuk validasi model probit adalah K-Fold Cross Validation. Nilai k yang dipilih adalah 10 dan dilakukan pengulangan pengujian sebanyak 10 kali.

Tabel 3. Hasil K-Fold Cross Validation

<i>Fold</i>	<i>Accuracy</i>	<i>Sensitivity</i>	<i>Specificity</i>
1	0,9374	0,9405	0,9377
2	0,9358	0,9365	0,9384
3	0,9386	0,9417	0,9389
4	0,9373	0,9413	0,9369
5	0,9370	0,9350	0,9420
6	0,9341	0,9361	0,9357
7	0,9383	0,9408	0,9391
8	0,9344	0,9337	0,9383
9	0,9398	0,9408	0,9420
10	0,9359	0,9389	0,9364
Rata-Rata	0,9369	0,9385	0,9385

Proses klasifikasi menggunakan model probit dengan 10 kali pengulangan diperoleh bahwa nilai akurasi berkisar antara 93%-94%, dengan nilai sensitivitas 94%, dan nilai spesitifitas yang sangat tinggi yaitu antara 93%-94%. Hal ini menunjukkan bahwa model probit mampu mengklasifikasi penentuan prediksi penyakit diabetes dengan sangat baik di setiap bagian data.

7. Pemilihan Model Terbaik

Model terbaik dipilih berdasarkan kriteria seperti nilai AIC yang paling rendah, Pseudo- R^2 tinggi, AUC tinggi, Logloss rendah, keakuratan klasifikasi tinggi, dan uji Goodness of fit. Dari analisis yang dilakukan, hasil pemilihan model terbaik disimpulkan berdasarkan keempat kriteria tersebut seperti yang ditunjukkan dalam Tabel 4.11 berikut.

Table 1. Hasil Kriteria Pemilihan Model Terbaik

<i>Model</i>	<i>AIC</i>	<i>Pseudo-R^2</i>	<i>AUC</i>	<i>Logloss</i>	<i>Keakuratan Klasifikasi</i>	<i>Uji Goodness of Fit</i>
Model Logit	36.770,33	78,58%	98,66%	0,1482	93,69%	Model Sesuai
Model Probit	36.757,55	78,59%	98,66%	0,1481	93,69%	Model Sesuai

Berdasarkan Tabel 4.11, diperoleh hasil pemilihan model terbaik berdasarkan keakuratan klasifikasi bahwa model logit dan probit memberikan akurasi yang sama sehingga kedua

model baik. Begitu juga hasil uji goodness of fit diperoleh bahwa model logit dan model probit sesuai dengan kasus ini. Hasil pemilihan model terbaik berdasarkan AUC menunjukkan bahwa kedua model logit dan model probit memberikan hasil yang sama sehingga kedua model baik. Akan tetapi, berdasarkan Pseudo-R² model logit dan probit memiliki perbedaan yang tidak signifikan yaitu hanya 0,01%. Begitu juga dengan melihat Logloss terkecil, model logit dan model probit memiliki perbedaan tidak signifikan yaitu hanya 0,0001. Berdasarkan nilai AIC terkecil, model logit lebih unggul dibandingkan dengan model probit, yaitu masing-masing sebesar 36.770,33 < 36.757,55. Dari enam kriteria diatas, model probit lebih banyak memenuhi syarat pemilihan model terbaik dibandingkan dengan model logit sehingga dapat disimpulkan bahwa model terbaik yang dipilih adalah model probit.

Simpulan

Berdasarkan hasil analisis yang dilakukan, estimasi parameter model logit dan model probit yang diperoleh melalui algoritma BHHH menunjukkan bahwa seluruh variabel prediktor yang digunakan, yaitu jenis kelamin, usia, riwayat hipertensi, riwayat penyakit jantung, riwayat merokok, indeks massa tubuh (BMI), kadar HbA1c, dan kadar gula darah, memiliki pengaruh signifikan terhadap kemungkinan terjadinya penyakit diabetes. Evaluasi model dilakukan menggunakan *k-fold cross validation* dan menunjukkan bahwa model logit dan model probit sama-sama memiliki rata-rata akurasi sebesar 93,7%, dengan perbedaan nilai sensitivitas dan spesifisitas yang tidak signifikan. Untuk memperkuat penilaian terhadap performa kedua model, dilakukan perbandingan menggunakan beberapa kriteria statistik, seperti Akaike Information Criterion (AIC), Pseudo-R², Area Under Curve (AUC), Logloss, keakuratan klasifikasi, dan uji *Goodness of fit* dengan Hosmer-Lemeshow test. Berdasarkan keseluruhan hasil evaluasi tersebut, model regresi probit dinilai lebih unggul dibandingkan model regresi logit dalam memprediksi penyakit diabetes pada data yang digunakan dalam penelitian ini.

Daftar Pustaka

- [1] J. Neter, *Applied linear regression models*. 1983.
- [2] K. Kumari and S. Yadav, "Linear regression analysis study," *J. Pract. Cardiovasc. Sci.*, vol. 4, no. 1, p. 33, 2018, doi: 10.4103/jpcs.jpcs_8_18.
- [3] D. Rika Widianita, "PREVALENSI KEJADIAN PENYAKIT TIDAK MENULAR (PTM)," *AT-TAWASSUTH J. Ekon. Islam*, vol. VIII, no. I, pp. 1–19, 2023.
- [4] A. S. Marbun, "Pelaksanaan Empat Pilar pada Penderita Diabetes Melitus," *J. Abdimas Mutiara*, vol. 3, no. 1, pp. 366–371, 2022.
- [5] Direktorat P2PTM, "Penyakit diabetes melitus," Kemenkes. [Online]. Available: <https://p2ptm.kemkes.go.id/informasi-p2ptm/penyakit-diabetes-melitus>

-
- [6] S. Hafni, R. Begum Suroyo, J. T. Sibero, Z. Nasution, and M. Wulan, "Faktor-Faktor yang Mempengaruhi Kejadian Hipertensi Pada Lansia di Puskesmas Pijorkoling Kecamatan Padangsidempuan Tenggara Kota Padangsidempuan," *J. Healthc. Technol. Med.*, vol. 7, no. 2, pp. 2615–109, 2021.
 - [7] H. D. Sitanggang, I. T. Ramadhanti, and R. Halim, "Faktor risiko kejadian diare pada anak balita (12-59 bulan) di Puskesmas 'X' Kota Jambi," *Ris. Inf. Kesehat.*, vol. 11, no. 1, p. 54, 2022, doi: 10.30644/rik.v11i1.624.
 - [8] Hasna and A. I. Achmad, "Metode Regresi Probit Biner untuk Pemodelan Faktor-Faktor yang Mempengaruhi Diagnosis Penyakit Jantung," *J. Ris. Stat.*, pp. 28–34, 2022, doi: 10.29313/jrs.vi.721.
 - [9] C. Dewanti, V. Ratnasari, and A. T. Rumiati, "Pemodelan Faktor-Faktor yang Memengaruhi Status Balita Stunting di Provinsi Jawa Timur Menggunakan Regresi Probit Biner," *J. Sains dan Seni ITS*, vol. 8, no. 2, 2020, doi: 10.12962/j23373520.v8i2.48519.
 - [10] I. N. I. Sari and V. Ratnasari, "Pemodelan Regresi Logistik dan Probit Biner Target Unmet Need di Provinsi Jawa Barat," *Sains dan Seni*, vol. 9, no. 2, 2020.
 - [11] H. Bai, "Preparing Teacher Education Students to Integrate Mobile Learning into Elementary Education," *TechTrends*, vol. 63, no. 6, pp. 723–733, Nov. 2019, doi: 10.1007/s11528-019-00424-z.
 - [12] F. Giannakas, A. Papasalouros, G. Kambourakis, and S. Gritzalis, "A comprehensive cybersecurity learning platform for elementary education," *Inf. Secur. J. A Glob. Perspect.*, vol. 28, no. 3, pp. 81–106, May 2019, doi: 10.1080/19393555.2019.1657527.
 - [13] R. M. Vink *et al.*, "Self-reported adverse childhood experiences and quality of life among children in the two last grades of Dutch elementary education," *Child Abuse Negl.*, vol. 95, p. 104051, Sep. 2019, doi: 10.1016/j.chiabu.2019.104051.