Indones. J. Math. Nat. Sci. 44(2) 2021
**Indonesian Journal of Mathematics and Natural Sciences**
http://journal.unnes.ac.id/nju/index.php/JM

# Peningkatan Ketepatan Klasifikasi Model Regresi Logistik Biner dengan Metode Bagging (Bootstrap Aggregating)

**Dwi Liana Wella Putri*, Scolastika Mariani, Sunarmi**

Jurusan Matematika, FMIPA, Universitas Negeri Semarang, Indonesia
Gedung D7 Lt.1, Kampus Sekaran Gunungpati, Semarang 50229
*E-mail: dwilianawella@gmail.com

Diterima 5 Januari 2021 Disetujui 12 September 2021 Dipublikasikan 31 Oktober 2021

**Abstrak**

Tujuan penelitian ini adalah mengetahui ketepatan klasifikasi regresi logistik dan bagging (*bootstrap aggregating*) regresi logistik biner pada status peserta KB Kota Tegal tahun 2016 serta mengetahui model terbaik regresi logistik biner. Penelitian ini melakukan estimasi status peserta KB Kota Tegal dengan metode maximum likelihood estimation disertai dengan algoritma newton raphson, dilanjutkan dengan pengujian signifikansi parameter baik secara simultan dengan Uji likelihood ratio dan parsial dengan uji wald. Selanjutnya uji kesesuaian model menggunakan Hosmer dan Lemeshow, uji ketepatan klasifikasi regresi logistik biner dan bootstrap aggregating, dan pemilihan model terbaik dengan melihat nilai ketepatan klasifikasi tertinggi dengan tingkat kesalahan terkecil. Software yang digunakan adalah program R 3.4.1. Disimpulkan, ketepatan klasifikasi metode bagging (bootstrap aggregating) sebesar 75,641% dengan kesalahan klasifikasi 24,359%. Metode bagging (bootstrap aggregating) meningkatkan ketepatan klasifikasi pada model regresi logistik biner dengan nilai ketepatan klasifikasi pada model regresi logistik biner 69,74% meningkat menjadi 75,641%. Model yang terbaik adalah model regresi logistik biner dengan menggunakan bagging (*bootstrap aggregating*).

Kata kunci: *KB, Kota Tegal, Metode Bagging, Regresi Logistik*

***Abstract***

*The purpose of this research is to know the accuracy of classification of logistic regression and bagging (bootstrap aggregating) logistic regression with binary data of KB participant in Tegal City year 2016 as well as logistic regression model find out binary. This research do estimation KB participant in Tegal City by the method of maximum likelihood estimation algorithm with newton raphson accompanied, followed by testing the significance of parameters simultaneously with partial likelihood ratio test and the testwith wald. The next test of suitability of the model by using the hosmer and lemeshow test and the accuracy of the classification of binary logistic regression and bootstrapping aggregating and conducted the selection of the best model by looking at the value of the highest level of classification accuracy the smallest mistake. The software used in the program R.3.4.1. the conclusion obtained from this research method of bagging (bootstrap aggregating) is 75,651% with 24,359% classification error. The method of bagging (bootstrap aggregating) improves the accuracy of classification of binary logistic regression models with the highest precision of the logistic regression model is a classification of binary 69,74% increase to 75,64%. The best model is a binary logistic regression model using bagging (bootstrap aggregating).*

Key words: Bagging method, KB, Kota Tegal, Logistic regression

**How to cite:**
Putri, D.L.W., Mariani, S., & Sunarmi. (2021). Peningkatan Ketepatan Klasifikasi Model Regresi Logistik Biner dengan Metode Bagging (Bootstrap Aggregating). *Indonesian Journal of Mathematics and natural Sciences*, 44(2), 61-72

## PENDAHULUAN

Salah satu metode analisis statistik yang digunakan untuk mencari hubungan antar variabel respon dengan variabel prediktor adalah regresi logistik biner. Pada regresi logistik biner, variabel

respon bersifat biner dan variabel prediktor bersifat polikotomus atau kontinu. Pada regresi logistik biner, nilai variabel respon bersifat dikotomi yaitu variabel yang hanya mempunyai dua kemungkinan, misalnya "sukses" atau "gagal". Dalam memudahkan penerapan regresi logistik, diperlukan adanya notasi pada tiap variabel yaitu variabel Y sebagai variabel respon dan X sebagai variabel prediktor (Hosmer & Lemeshow, 2000).

Pada masalah klasifikasi data, terdapat banyak klasifikasi seperti *Decision Tree, Artificial Neural Network, Logistic Regression, dan Naive Bayes*. Naive Bayes bekerja sangat baik ketika distribusi suatu variabel respon pada sekumpulan data set yang seimbang (Nai-arun & Moungmai, 2015). Akan tetapi, pada kenyataannya masalah ketidakseimbangan dapat terjadi seperti pada klasifikasi dokumen (Laza *et al.*, 2011), dan prediksi kegagalan *credit-scoring* (Brown & Mues, 2012). Salah satu klasifikasi data yang masuk dalam kategori ini adalah data dengan variabel respon biner. Sekumpulan data yang seimbang sangat penting untuk membuat model prediksi yang baik. Untuk memprediksi suatu model regresi logistik biner, dibutuhkan nilai akurasi yang stabil. Namun, terkadang pendugaan parameter pada regresi logistik biner mengalami ketidakseimbangan klasifikasi (*missclasification*). Hal ini disebabkan data memiliki nilai varians yang tinggi (*noise*) sehingga menyebabkan kurangnya akurasi pada klasifikasi meskipun nilai akurasi keseluruhan tinggi (Prasetio & Pratiwi, 2015; Kim & Kang, 2012). Untuk mengatasi ketidakseimbangan model pada regresi logistik biner, dapat digunakan salah satu metode klasifikasi yaitu metode *bagging* (*bootstrap aggregating*).

Metode *bagging* merupakan salah satu metode yang mengkombinasikan *bootstrapping* dan *aggregating* (Alfaro *et al*., 2013). Metode ini digunakan untuk meningkatkan ketepatan klasifikasi pada sekumpulan data dengan cara mengurangi varians dengan tetap menjaga atau hanya sedikit meningkatkan bias. Metode bagging adalah metode *ensemble* yang sederhana namun efektif dan telah diterapkan untuk banyak aplikasi di dunia nyata (Liang & Zhang, 2011). Ide dasar dari bagging adalah menggunakan bootstrap dimana ketika dikombinasikan seharusnya hasilnya lebih baik dibandingkan dengan prediktor tunggal yang dibangun untuk menyelesaikan masalah yang sama. Menurut Hosmer dan Lemeshow (2000), pada regresi logistik biner, keluaran dari variabel respon terdiri dari dua kategori, misalnya sukses atau gagal yang dinotasikan dengan $y = 1$ berarti sukses atau $y = 0$ berarti gagal. Dengan demikian, variabel respon mengikuti distribusi Bernoulli. Fungsi probabilitas untuk setiap observasi dapat dituliskan sebagai berikut :

$$f(y) = \pi^{y}(1-\pi)^{1-y}; y = 0,1$$

Pada regresi linear variabel respon diasumsikan berdistribusi normal, namun pada regresi logistik biner variabel bersifat kategorikal. Bentuk fungsi regresi logistiknya adalah sebagai berikut:

$$f(x) = \frac{1}{1+e^{-x}}, -\infty < x < \infty$$

bentuk regresi logistiknya:

$$\pi(x) = \frac{e^{(\beta_0 + \beta_1 x_1 + \cdots + \beta_i x_i)}}{1 + e^{(\beta_0 + \beta_1 x_1 + \cdots + \beta_i x_i)}}$$

Persamaan diatas mempunyai bentuk yang tidak linier. Untuk membuat persamaan tersebut menjadi persamaan yang linier, maka persamaan diatas perlu ditransformasikan dengan menggunakan transformasi logit. Bentuk transformasinya adalah:

$$g\,(x) = ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_i x_i$$

Menurut Bakari (2016), ada beberapa metode yang baik dalam mengestimasi parameter, yaitu metode momen, maximum likelihood estimation, dan metode bayes. Namun, dari beberapa metode tersebut, metode *maximum likelihood estimation* merupakan metode yang kemungkinan menghasilkan penaksir yang baik. Dalam menentukan penaksir *maximum likelihood*, diperlukan adanya fungsi yang disebut fungsi *likelihood*. Fungsi ini menyatakan kemungkinan dari data yang diamati sebagai fungsi dari parameter yang tidak diketahui. Untuk menaksir parameter β, fungsi likelihood dimaksimumkan dengan syarat bahwa data tersebut mengikuti suatu distribusi tertentu. Pada regresi logistik biner, setiap pengamatan mengikuti distribusi Bernoulli sehingga didapat fungsi *likelihood*-nya. Untuk mendapatkan hasil yang konvergen, fungsi *likelihood* dimaksimumkan dengan algoritma newton raphson (Hosmer & Lemenshow, 2000).

Pada regresi logistik biner, uji signifikansi parameter dalam model dapat dilakukan dengan uji rasio likelihood dan uji wald. uji wald pada regresi logistik biner dapat dituliskan sebagai berikut:
Hipotesis:

$H_0 : \beta_j = 0$ (parameter $\beta$ tidak berpengaruh secara signifikan terhadap variabel respon secara individu)
$H_1 : \beta_j \neq 0 ; j = 1,2, \dots, p$ (parameter $\beta$ berpengaruh secara signifikan terhadap variabel respon secara individu)
dengan bentuk statistik uji wald:

$$W_j = \frac{(\hat{\beta}_j)}{[SE(\hat{\beta}_j)]}$$

di mana
$\beta_j$ : nilai koefisien regresi logistik untuk variabel ke- j
$\hat{\beta}_j$ : penduga dari parameter β
$SE(\hat{\beta}_j)$ : nilai standar error untuk variabel ke -$i$
Tolak $H_0$ , jika $|W_j| > Z_{(\alpha;db)}$ atau ${W_j}^2 > {\chi^2}_{(db,\alpha)}$ yang berarti parameter $\beta$ berpengaruh secara signifikan terhadap variabel respon secara individu.

Uji simultan atau serentak merupakan pengujian yang dilakukan untuk memeriksa signifikansi parameter $\beta$ terhadap variabel prediktor secara keseluruhan. Pada uji ini, statistik yang digunakan adalah statistik uji $G$ atau *Likelihood Ratio Test* (Hosmer & Lemeshow, 2000).
Hipotesis :
$H_0 : \beta_1 = \beta_2 = \cdots = \beta_j = 0$ (variabel prediktor tidak berpengaruh secara signifikan terhadap variabel respon)
$H_1$ : paling sedikit ada satu $\beta_j \neq 0$ $(i = 1,2, \dots, j)$ (variabel prediktor berpengaruh secara signifikan terhadap variabel respon)

Bentuk statistik uji *Likelihood Ratio Test*:

$$G = -2\, ln\left(\frac{L_0}{L_1}\right)$$

$L_0$ : likelihood tanpa variabel bebas
$L_1$ : likelihood dengan variabel bebas

Pada hipotesis $H_0$, bahwa $\beta_1$ sama dengan nol, statistik uji G mengikuti distribusi *Chi-Square*. Hasil uji statistik diatas bahwa $H_0$ ditolak jika $G > {\chi^2}_{(db,\alpha)}$ atau *p-value* $< \alpha$.

Uji kesesuaian model merupakan uji yang digunakan untuk mengetahui apakah terdapat perbedaan antara hasil observasi dengan kemungkinan hasil prediksi. Proses pengujian yang digunakan pada uji ini adalah uji Hosmer dan Lemeshow. Hipotesis yang digunakan adalah sebagai berikut:
Hipotesis :
$H_o$ : model sesuai (tidak ada perbedaan antara hasil observasi dengan hasil prediksi)
$H_1$ : Model tidak sesuai (ada perbedaan antara hasil observasi dengan hasil prediksi)

Bentuk statistik uji :

$$\hat{C} = \sum_{f=1}^{g} \left[ \frac{\left(O_f - n\bar{\pi}_f\right)^2}{\left(n_f \pi_f (1 - \pi_f)\right)} \right]$$

Di mana :
$g$ : jumlah grup
$O_f$ :jumlah nilai variabel respon pada grup ke -$f$
$\bar{\pi}_f$ : rata-rata taksiran peluang
$n_f$ : banyak observasi pada grup ke $-f$
Kriteria uji : $H_0$ ditolak jika $\hat{C} > {\chi^2}_{(g-2,\alpha)}$

Menurut Arleina dan Otok (2014), Press'Q adalah ukuran yang digunakan untuk mengetahui kestabilan dalam pengklasifikasian atau sejauh mana kelompok-kelompok tersebut dapat dipisahkan. Bentuk uji statistik dari press'Q dapat dituliskan sebagai berikut:

$$\frac{[N - (nK)]^2}{N\,(K-1)} \sim {\chi^2}_{(1)}$$

dimana

$N$ : banyaknya pengamatan
$n$: banyaknya individu yang tepat diklasifikasikan
$K$ : banyaknya kelompok

*APER* (*Apparent Error Rate*) merupakan suatu nilai yang digunakan untuk melihat kesalahan dalam mengklasifikasi suatu objek. Nilai *APER* didapatkan dengan perhitungan diberikan pada Tabel 1.

Tabel 1 Tabel Klasifikasi Regresi Logistik

| Hasil Observasi | Taksiran | |
|---|---|---|
| | $y_1$ | $y_2$ |
| $y_1$ | $n_{11}$ | $n_{12}$ |
| $y_2$ | $n_{21}$ | $n_{22}$ |

Keterangan:
$n_{11}$ : jumlah subjek dari $y_1$ tepat diklasifikasikan sebagai $y_1$
$n_{12}$ : jumlah subjek dari $y_1$ salah diklasifikasikan sebagai $y_2$
$n_{21}$ : jumlah subjek dari $y_2$ salah diklasifikasikan sebagai $y_1$
$n_{22}$ : jumlah subjek dari $y_2$ tepat diklasifikasikan sebagai $y_2$

$$APER(\%) = \frac{\text{jumlah prediksi salah}}{\text{jumlah total prediksi}} = \frac{n_{12} + n_{21}}{n_{11} + n_{12} + n_{21} + n_{22}} \times 100\%$$

dengan nilai akurasi:

$$AKURASI = 1 - APER = 1 - \frac{\text{jumlah prediksi salah}}{\text{jumlah total prediksi}}$$

Untuk memperoleh parameter yang stabil pda regresi logistik biner, dapat digunakan suatu pendekatan bootstrap yaitu metode *bagging* (*bootstrap aggregating*). *Bagging* merupakan metode untuk menghasilkan beberapa versi prediktor dan menggunakannya untuk mendapatkan prediktor agregat. Rata-rata agregat dari beberapa prediktor digunakan untuk memprediksi keluaran numerik. Prediktor-prediktor terbentuk dengan replikasi bootstrap pada data dan menggunakannya sebagai data baru.

Himpunan data (data set) £ terdiri dari $\{(y_n, x_n), n = 1, \dots, N\}$ dengan y dapat berupa kelas label atau numerik respon. Jika input adalah $x$, maka y diprediksi dengan $\varphi(x, £)$ dimana $\varphi(x, £)$ adalah parameter. Untuk mendapatkan prediktor yang lebih baik maka dilakukan replikasi bootstrap $\{£_k\}$ yang kemudian disebut $\{\varphi(x, £_k)\}$. Replikasi bootstrap dilakukan sebanyak B kali sehingga $\{£^{(B)}\}$ dari £ dan dibentuk prediktor $\{\varphi(x, £^{(B)})\}$ dimana $\{£^{(B)}\}$ adalah *resampling* dengan pengembalian. Jika $y$ data numerik, maka prosedur nyata untuk menggantikan $\varphi(x, £)$ dengan mengambil rata-rata dari $\{\varphi(x, £_k)\}$ untuk memprediksi kelas. Ambil bootstrap sampel dengan pengulangan $\{£^{(B)}\}$ dari £ dan membentuk $\{\varphi(x, £^{(B)})\}$ (Breiman, 1994).

Metode *aggregate classifier* $\mu_A$ secara umum dapat dituliskan sebagai berikut:

$$\mu_A(y) = E_F[\hat{\mu}(y, £_k)]$$

Langkah-langkah algoritma *bagging* dapat dituliskan sebagai berikut :

1. Mengambil sampel bootstrap sebanyak $n$ dari data set £ dengan pengulangan sebanyak $n$. Pengambilan sampel sedemikian hingga setiap variabel aggregate dalam setiap observasi.
2. Memodelkan regresi logistik hasil sampel bootstrap $£^{(B)}$.
3. Menghitung peluang kumulatif, peluang masing-masing kategori respon untuk setiap observasi dan menghitung ketepatan klasifikasi. Kesalahan klasifikasi pada langkah ini disebut $e_B$.
4. Mengulang langkah 1-3 sebanyak B kali (Replikasi bootstrap)
5. Memperoleh ketepatan klasifikasi bagging yaitu rata-rata ketepatan klasifikasi setiap pengulangan sampai B sehingga kesalahan bagging untuk replikasi B kali adalah $\bar{e}_B$.

6. Membentuk model bagging regresi logistik dari rata-rata setiap parameter pada pengulangan sampai B.

Hasil sensus penduduk yang dilakukan oleh Badan Pusat Statistik (BPS) tahun 2010 menyebutkan bahwa jumlah penduduk di Indonesia adalah 238,52 juta jiwa dan bertambah menjadi sekitar 255,46 juta jiwa pada tahun 2015. Jumlah penduduk Kota Tegal tahun 2015 berdasarkan hasil SUSENAS sendiri mencapai 276.734 jiwa. Kepadatan penduduk di Kota Tegal tercatat sebesar 6.203 jiwa setiap kilometer persegi. Untuk mengendalikan jumlah penduduk yang terus meningkat, Badan Kependudukan dan Keluarga Berencana Nasional (BKKBN) melakukan program Keluarga Berencana (KB) bagi pasangan usia subur. Penelitian ini menerapkan metode bootstrap aggregating untuk meningkatkan ketepatan klasifikasi pada model regresi logistik biner dalam kasus status peserta KB Kota Tegal pada tahun 2016. Pada penelitian ini akan dibandingkan nilai ketepatan klasifikasi antara regresi logistik biner dan bootstrap aggregating dan akan didapat model terbaik antara regresi logistik biner dengan *bagging (bootstrap aggregating*).

## METODE

Penelitian ini berfokus pada (1) Penelitian menggunakan model regresi logistik biner dan bagging (*bootstrap aggregating*), (2) Estimasi parameter yang digunakan dalam penelitian ini menggunakan metode *maximum likelihood estimation* dan algoritma newton raphson, (3) peningkatan ketepatan klasifikasi pada model regresi logistik biner menggunakan bagging (*bootstrap aggregating*) untuk meningkatkan akurasi, (4) Penelitian didukung dengan bantuan program R.3.4.1 dan SPSS v.23.

Data yang diperoleh dari BKKBN Provinsi Jawa Tengah tahun 2016 yaitu data status peserta KB ($Y$) dengan kategori: bukan peserta KB (0) dan peserta KB (1). Variabel usia istri ($X_1$)dengan kategori: $\leq$ 50 tahun, $>$ 50 tahun. Usia suami ($X_2$) dengan kategori: $\leq$ 50 tahun, $>$ 50 tahun. Pendidikan terakhir istri ($X_3$)dengan kategori: Tidak/belum sekolah, Tidak tamat SD/MI, Tamat SD/MI, Tidak tamat SLTP/MTSN, Tamat SLTP/MTSN, Tidak tamat SLTA/MA, Tamat SLTA/MA, Tidak tamat PT/akademi, Tamat PT/akademi. Pendidikan terakhir suami ($X_4$) dengan kategori: Tidak/belum sekolah, Tidak tamat SD/MI, Tamat SD/MI, Tidak tamat SLTP/MTSN, Tamat SLTP/MTSN, Tidak tamat SLTA/MA, Tamat SLTA/MA, Tidak tamat PT/akademi, Tamat PT/akademi. Pekerjaan istri ($X_5$) dengan kategori: Petani, Nelayan, Pedagang, PNS/TNI/POLRI, Pegawai swasta, Wiraswasta, Pensiunan, Pekerja lepas, Lainnya, Tidak bekerja. Pekerjaan suami ($X_6$) dengan kategori: Petani, Nelayan, Pedagang, PNS/TNI/POLRI, Pegawai swasta, Wiraswasta, Pensiunan, Pekerja lepas, Lainnya, Tidak bekerja. Variabel terakhir yaitu jumlah anak hidup ($X_7$).

Langkah-langkah yang digunakan dalam metode penelitian ini adalah :

(1) Menguji OLS (*Ordinary Least Square*).
(2) Pembentukan model regresi logistik biner dengan mengestimasi parameter menggunakan MLE.
(3) menentukan model regresi dengan pengujian secara simultan terhadap seluruh variabel prediktor dengan cara memasukkan seluruh variabel prediktor yang signifikan berpengaruh pada pengujian secara individu.
(4) Melakukan analisis regresi logistik biner dengan pengujian secara individu terhadap masing-masing variabel prediktor.
(5) Melakukan uji kesesuaian model dengan menggunakan uji Hosmer dan lemeshow lalu diperoleh variabel prediktor yang signifikan berpengaruh terhadap model regresi logistik biner.
(6) Menentukan kesalahan klasifikasi regresi logistik biner.
(7) Melakukan *bootstrap aggregating* untuk prediktor dari model regresi logistik biner sebanyak B 50 replikasi bootstrap.
(8) Menentukan ketepatan klasifikasi pada setiap pengambilan sampel B replikasi bootstrap, sehingga diperoleh kesalahan klasifikasi $e_B$.
(9) Menentukan kesalahan klasifikasi bagging $\bar{e}_B$.
(10) Membandingkan ketepatan klasifikasi model regresi logistik biner dan bagging regresi logistik biner.

## HASIL DAN PEMBAHASAN

Berdasarkan uji OLS (*Ordinary Least Square*) kriteria yang tidak memenuhi adalah kriteria normalitas, dan heterokedastisitas sehingga data ini didekati dengan pendekakatan regresi logistik biner. Uji heteroskedastisitas tidak bias memenuhi karena pada data ini berbentuk kategorik sehingga

tidak bias dilakukan penyembuhan. Selanjutnya dilakukan pembentukan model awal regresi logistik biner dengan menggunakan estimasi *maximum likelihood estimation* dan algoritma *newton raphson* seperti terlihat pada Tabel 2.

Tabel 2 Estimasi Parameter untuk Model Awal Regresi Logistik Biner

| $No$ | $(X_i)$ | $\beta_j$ |
|---|---|---|
| 1 | $(X_{1\_2})$ | 0,76527 |
| 2 | $(X_{2\_2})$ | 1,10945 |
| 3 | $(X_{3\_2})$ | -1,08088 |
| 4 | $(X_{3\_4})$ | -0,80080 |
| 5 | $(X_{3\_5})$ | -0,57856 |
| 6 | $(X_{3\_6})$ | -2,00045 |
| 7 | $(X_{3\_7})$ | -14,71783 |
| 8 | $(X_{3\_8})$ | -2,05412 |
| 9 | $(X_{3\_9})$ | 15,17975 |
| 10 | $(X_{4\_2})$ | -0,71855 |
| 11 | $(X_{4\_3})$ | 17,22801 |
| 12 | $(X_{4\_4})$ | -0,51464 |
| 13 | $(X_{4\_5})$ | 17,84836 |
| 14 | $(X_{4\_6})$ | 0,11024 |
| 15 | $(X_{4\_7})$ | 0,56627 |
| 16 | $(X_{4\_8})$ | 0,91570 |
| 17 | $(X_{4\_9})$ | -0,06683 |
| 18 | $(X_{5\_2})$ | -0,43856 |
| 19 | $(X_{5\_3})$ | 0,05113 |
| 20 | $(X_{5\_4})$ | -1,19291 |
| 21 | $(X_{5\_5})$ | 0,39449 |
| 22 | $(X_{5\_6})$ | 0,24773 |
| 23 | $(X_{5\_7})$ | 17,98805 |
| 24 | $(X_{5\_8})$ | -0,16053 |
| 25 | $(X_{5\_9})$ | 1,44986 |
| 26 | $(X_{5\_10})$ | 1,27527 |
| 27 | $(X_{6\_2})$ | -33,45318 |
| 28 | $(X_{6\_3})$ | -18,28683 |
| 29 | $(X_{6\_4})$ | -17,24723 |
| 30 | $(X_{6\_5})$ | -19,40141 |
| 31 | $(X_{6\_6})$ | -18,73082 |
| 32 | $(X_{6\_8})$ | -17,11914 |
| 33 | $(X_{6\_9})$ | -17,79022 |
| 34 | $(X_{6\_10})$ | -18,52069 |
| 35 | $(X_7)$ | 0,34417 |

dengan

$(X_{1\_2})$ : usia istri diatas 50 tahun
$(X_{2\_2})$ : usia suami diatas 50 tahun
$(X_{3\_2})$ : pendidikan terakhir istri tidak tamat SD
$(X_{3\_4})$ :pendidikan terakhir istri tidak tamat SLTP
$(X_{3\_5})$ : pendidikan terakhir istri tamat SLTP
$(X_{3\_6})$ :pendidikan terakhir istri tidak tamat SLTA
$(X_{3\_7})$ : pendidikan terakhir istri tamat SLTA
$(X_{3\_8})$ :pendidikan terakhir istri tidak tamat PT/akademi
$(X_{3\_9})$ :pendidikan terakhir istri tamat PT/akademi
$(X_{4\_2})$ : pendidikan terakhir suami tidak tamat SD
$(X_{4\_3})$ : pendidikanterakhir suami tamat SD
$(X_{4\_4})$ :pendidikan terakhir suami tidak tamat SLTP

($X_{4\_5}$) : pendidikan terakhir suami tamat SLTP
($X_{4\_6}$) : pendidikan terakhir suami tidak tamat SLTA
($X_{4\_7}$) : pendidikan terakhir suami tamat SLTA
($X_{4\_8}$) : pendidikan terakhir suami tidak tamat PT/akademi
($X_{4\_9}$) :pendidikan terakhir suami tamat PT/akademi
($X_{5\_2}$) : pekerjaan istri (nelayan)
($X_{5\_3}$) : pekerjaan istri (pedagang)
($X_{5\_4}$) : pekerjaan istri (PNS)
($X_{5\_5}$) : pekerjaan istri (pegawai swasta)
($X_{5\_6}$) : pekerjaan istri (wiraswasta)
($X_{5\_7}$) : pekerjaan istri (pensiunan)
($X_{5\_8}$) : pekerjaan istri (pekerja lepas)
($X_{5\_9}$) : pekerjaan istri (lainnya)
($X_{5\_10}$) : pekerjaan istri (tidak bekerja)
($X_{6\_2}$) : pekerjaan suami (nelayan)
($X_{6\_3}$) : pekerjaan suami (pedagang)
($X_{6\_4}$) : pekerjaan suami (PNS)
($X_{6\_5}$) : pekerjaan suami (pegawai swasta)
($X_{6\_6}$) : pekerjaan suami (wiraswasta)
($X_{6\_8}$) : pekerjaan suami (pekerja lepas)
($X_{6\_9}$) : pekerjaan suami (lainnya)
($X_{6\_10}$) : pekerjaan suami (tidak bekerja)
($X_7$) : jumlah anak hidup

dengan bentuk regresi logistik biner:

$$g(x) = 17{,}14729 + 0{,}76527\,X_{1\_2} + 1{,}10945\,X_{2\_2} - 1{,}08088\,X_{3\_3} - 0{,}80080\,X_{3\_4} - 0{,}57856\,X_{3\_5} - 2{,}00045\,X_{3\_6} - 14{,}71783\,X_{3\_7} - 2{,}05412\,X_{3\_8} + 15{,}17975\,X_{3\_9} - 0{,}71855\,X_{4\_2} + 17{,}22801\,X_{4\_3} - 0{,}51464\,X_{4\_4} + 17{,}84836\,X_{4\_5} + 0{,}11024\,X_{4\_6} + 0{,}56627\,X_{4\_7} + 0{,}9157\,X_{4\_8} - 0{,}06683\,X_{4\_9} - 0{,}43856\,X_{5\_2} + 0{,}05113\,X_{5\_3} - 1{,}19291\,X_{5\_4} + 0{,}39449\,X_{5\_5} + 0{,}24773\,X_{5\_6} + 17{,}98805\,X_{5\_7} - 0{,}16053\,X_{5\_8} + 1{,}44986\,X_{5\_9} + 1{,}27527\,X_{5\_10} - 33{,}45318\,X_{6\_2} - 18{,}28683\,X_{6\_3} - 17{,}24723\,X_{6\_4} - 19{,}40141X_{6\_5} - 18{,}73082X_{6\_6} - 17{,}11914X_{6\_8} - 17{,}79022X_{6\_9} - 18{,}52069X_{6\_10} + 0{,}34417X_7$$

di mana

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

Langkah selanjutnya dilakukan uji signifikansi parameter baik simultan dan uji parsial. Langkah pengujian simultan menggunakan uji $G$ atau uji likelihood ratio test dengan prosedur pengujian sebagai berikut:

hipotesis:

$H_0 : \beta_1 = \beta_2 = \cdots = \beta_j = 0$ (variabel prediktor tidak berpengaruh secara signifikan terhadap variabel respon)

$H_1$ : paling sedikit ada satu $\beta_j \neq 0$ $(i = 1{,}2, \ldots, j)$ (variabel prediktor berpengaruh secara signifikan terhadap variabel respon)

dengan taraf signifikansi 5% diperoleh statistik uji $G$ seperti pada Tabel 3.

Tabel 3. Likelihood Ratio Test

| | LogLik | Pr(>Chisq) |
|---|---|---|
| $L_0$ | $-198{,}25$ | |
| $L_1$ | $-245{,}38$ | 2,453-07*** |

Karena nilai $(G = 94{,}26) > (\chi^2_{(0{,}05{,}35)} = 49{,}802)$, atau dapat dilihat nilai dari *p-value* $= 2{,}453e - 07 < \alpha = 0{,}05$, dengan demikian $H_0$ ditolak. Artinya usia istri ($X_1$), usia suami ($X_2$), pendidikan terakhir istri ($X_3$), pendidikan terakhir suami ($X_4$), pekerjaan istri ($X_5$), pekerjaan suami ($X_6$) serta

jumlah anak hidup ($X_7$) secara bersama-sama berpengaruh secara signifikan terhadap status peserta KB. Selanjutnya dilakukan uji wald pada model regresi logistik biner dengan prosedur pengujian sebagai berikut:

Hipotesis:

$H_0 : \beta_j = 0$ (parameter $\beta$ tidak berpengaruh secara signifikan terhadap variabel respon secara individu)

$H_1 : \beta_j \neq 0 ; j = 1,2, \dots, p$ (parameter $\beta$ berpengaruh secara signifikan terhadap variabel respon secara individu)

Taraf signifikansi 5%. Hasil uji wald diberikan pada Tabel 4.

Tabel 4. Nilai Uji Wald untuk Setiap Parameter pada Model Awal

| No | $(X_i)$ | *p-value* | Keputusan |
|---|---|---|---|
| 1 | $(X_{1\_2})$ | 0,004 | $H_0$ ditolak |
| 2 | $(X_{2\_2})$ | 0,024 | $H_0$ ditolak |
| 3 | $(X_{3\_2})$ | 0,058 | $H_0$diterima |
| 4 | $(X_{3\_4})$ | 0,224 | $H_0$ diterima |
| 5 | $(X_{3\_5})$ | 0,690 | $H_0$diterima |
| 6 | $(X_{3\_6})$ | 0,006 | $H_0$ ditolak |
| 7 | $(X_{3\_7})$ | 0,995 | $H_0$ diterima |
| 8 | $(X_{3\_8})$ | 0,018 | $H_0$ditolak |
| 9 | $(X_{3\_9})$ | 0,995 | $H_0$ diterima |
| 10 | $(X_{4\_2})$ | 0,277 | $H_0$ diterima |
| 11 | $(X_{4\_3})$ | 0,994 | $H_0$diterima |
| 12 | $(X_{4\_4})$ | 0,473 | $H_0$ diterima |
| 13 | $(X_{4\_5})$ | 0,994 | $H_0$ diterima |
| 14 | $(X_{4\_6})$ | 0,880 | $H_0$ diterima |
| 15 | $(X_{4\_7})$ | 0,805 | $H_0$ diterima |
| 16 | $(X_{4\_8})$ | 0,320 | $H_0$diterima |
| 17 | $(X_{4\_9})$ | 0,968 | $H_0$ diterima |
| 18 | $(X_{5\_2})$ | 0,638 | $H_0$ diterima |
| 19 | $(X_{5\_3})$ | 0,960 | $H_0$ diterima |
| 20 | $(X_{5\_4})$ | 0,339 | $H_0$ diterima |
| 21 | $(X_{5\_5})$ | 0,694 | $H_0$diterima |
| 22 | $(X_{5\_6})$ | 0,802 | $H_0$diterima |
| 23 | $(X_{5\_7})$ | 0,991 | $H_0$ diterima |
| 24 | $(X_{5\_8})$ | 0,865 | $H_0$ diterima |
| 25 | $(X_{5\_9})$ | 0,203 | $H_0$diterima |
| 26 | $(X_{5\_10})$ | 0,398 | $H_0$ diterima |
| 27 | $(X_{6\_2})$ | 0,990 | $H_0$ diterima |
| 28 | $(X_{6\_3})$ | 0,994 | $H_0$diterima |
| 29 | $(X_{6\_4})$ | 0,994 | $H_0$diterima |
| 30 | $(X_{6\_5})$ | 0,994 | $H_0$ diterima |
| 31 | $(X_{6\_6})$ | 0,994 | $H_0$diterima |
| 32 | $(X_{6\_8})$ | 0,994 | $H_0$diterima |
| 33 | $(X_{6\_9})$ | 0,994 | $H_0$ diterima |
| 34 | $(X_{6\_10})$ | 0,994 | $H_0$ diterima |
| 35 | $(X_7)$ | 0,000 | $H_0$ ditolak |

$H_0$ ditolak jika $W_j^2 > \chi^2_{(0,05,1)}$ atau *p-value* $< \alpha$

Jadi, pada taraf signifikansi 5% terlihat bahwa variabel prediktor yang mempengaruhi variabel dependen adalah $X_1, X_2, X_3$ dan $X_7$, sedangkan variabel lain tidak mempengaruhi. Pada variabel pendidikan terakhir istri ($X_3$) dengan kategori tidak tamat SD/MI, tidak tamat SLTP/MTSN, tamat SLTP/MTSN, tamat SLTA/MA, dan tamat PT/Akademi dianggap signifikan karena masih satu variabel dengan pendidikan terakhir istri. Sedangkan variabel pendidikan terakhir suami ($X_4$),

pekerjaan istri $(X_5)$, pekerjaan suami $(X_6)$ dihilangkan karena tidak memberikan pengaruh yang signifikan terhadap variabel dependen. Langkah selanjutnya dibentuk model baru dengan menggunakan variabel $X_1, X_2, X_3$, dan $X_7$ dan diuji rasio likelihood serta uji wald kedua. Diperoleh bentuk model regresi logistik kedua seperti pada Tabel 5.

Tabel 5. Estimasi Parameter pada Model Regresi Logistik Kedua

| $No$ | $(X_i)$ | $\beta_j$ |
|---|---|---|
| 1 | $(X_{1\_2})$ | 0,734 |
| 2 | $(X_{2\_2})$ | 0,886 |
| 3 | $(X_{3\_2})$ | $-1,266$ |
| 4 | $(X_{3\_4})$ | $-0,536$ |
| 5 | $(X_{3\_5})$ | 0,082 |
| 6 | $(X_{3\_6})$ | $-1,311$ |
| 7 | $(X_{3\_7})$ | $-14,117$ |
| 8 | $(X_{3\_8})$ | $-1,182$ |
| 9 | $(X_{3\_9})$ | 14,742 |
| 10 | $(X_7)$ | 0,273 |

Diperoleh nilai

$$g(x) = -1,268+0,734X_{1_2}+0,886X_{2\_2} -1,266X_{3_2} -0,536X_{3_4} +0,082X_{3_5} -1,311X_{3_6} -14,117X_{3\_7} -1,182X_{3\_8} +14,742X_{3\_9} +0,273X_7$$

Setelah dibentuk model kedua dengan variabel yang signifikan yaitu $X_1, X_2, X_3$, dan $X_7$. Langkah selanjutnya yaitu uji likelihood ratio test sebagai berikut:

Hipotesis:

$H_0 : \beta_1 = \beta_2 = \cdots = \beta_j = 0$ (variabel prediktor tidak berpengaruh secara signifikan terhadap variabel respon)

$H_1$ : paling sedikit ada satu $\beta_j \neq 0$ $(i = 1,2, \dots, j)$ (variabel prediktor berpengaruh secara signifikan terhadap variabel respon)

Dengan taraf signifikansi 5% diperoleh bentuk uji *likelihood ratio test* seperti pada Tabel 6.

Tabel 6 *Likelihood Ratio Test*

| | LogLik | Pr(>Chisq) |
|---|---|---|
| $L_0$ | $-224,26$ | |
| $L_1$ | $-245,38$ | 6,844e-06*** |

Karena nilai $(G = 42,24) > (\chi^2_{(0,05,10)} = 18,307)$, atau dapat dilihat nilai dari *p-value* $= 6,844e-06 < \alpha = 0,05$, dengan demikian $H_0$ ditolak. Artinya usia istri $(X_1)$, usia suami $(X_2)$, pendidikan terakhir istri $(X_3)$, serta jumlah anak hidup $(X_7)$ secara bersama-sama berpengaruh secara signifikan terhadap status peserta KB.

Uji wald kedua digunakan untuk melihat pengaruh masing-masing variabel prediktor terhadap variable respon dengan mengunakan variabel prediktor $X_1, X_2, X_3$, dan $X_7$ diperoleh hipotesis sebagai berikut:

$H_0 : \beta_j = 0$ (parameter $\beta$ tidak berpengaruh secara signifikan terhadap variabel respon secara individu)

$H_1 : \beta_j \neq 0 ; j = 1,2, \dots, p$ (parameter $\beta$ berpengaruh secara signifikan terhadap variabel respon secara individu)

Dengan taraf signifikansi 5% diperoleh nilai uji wald seperti Tabel 7.

Jadi, pada taraf signifikansi 5% terlihat bahwa variabel prediktor yang mempengaruhi variabel dependen adalah $X_1, X_2, X_3$, dan $X_7$. Pada variabel pendidikan terakhir istri $(X_3)$ dengan kategori tidak tamat SLTP/MTSN, tamat SLTP/MTSN, tamat SLTA/MA, dan tamat PT/Akademi dianggap signifikan karena nilai tidak signifikan bukan berarti tidak mempengaruhi, akan tetapi variabel pendidikan terakhir istri dengan kategori tidak tamat SLTP/MTSN, tamat SLTP/MTSN, tamat SLTA/MA, dan tamat PT/Akademi memiliki pengaruh yang sangat sedikit sekali sehingga tidak signifikan.

Tabel 7. Nilai Uji Wald untuk Setiap Parameter pada Model Kedua

| No | $(X_i)$ | p-value | Keputusan |
|---|---|---|---|
| 1 | $(X_{1\_2})$ | 0,001 | $H_0$ ditolak |
| 2 | $(X_{2\_2})$ | 0,05 | $H_0$ ditolak |
| 3 | $(X_{3\_2})$ | 0,004 | $H_0$ ditolak |
| 4 | $(X_{3\_4})$ | 0,279 | $H_0$ diterima |
| 5 | $(X_{3\_5})$ | 0,940 | $H_0$ diterima |
| 6 | $(X_{3\_6})$ | 0,007 | $H_0$ ditolak |
| 7 | $(X_{3\_7})$ | 0,987 | $H_0$ diterima |
| 8 | $(X_{3\_8})$ | 0,030 | $H_0$ ditolak |
| 9 | $(X_{3\_9})$ | 0,987 | $H_0$ diterima |
| 10 | $(X_7)$ | 0,001 | $H_0$ ditolak |

.

Untuk melihat apakah data hasil observasi sesuai dengan hasil prediksi perlu dilakukan uji hosmer dan lemeshow dengan prosedur pengujian sebagai berikut:
Hipotesis:
$H_0$ : model sesuai (tidak ada perbedaan antara hasil observasi dengan hasil prediksi)
$H_1$ : Model tidak sesuai (ada perbedaan antara hasil observasi dengan hasil prediksi)
Dengan taraf signifikansi 5% dengan kriteria $H_0$ ditolak jika $\hat{C} > \chi^2_{(0,05,8)}$ atau *p-value* $< \alpha$ diperoleh hasil pengujian seperti pada Tabel 8.

Tabel 8. Uji Hosmer dan Lemeshow

| *Hosmer and Lemeshow goodness of fit (GOF) test* | | |
|---|---|---|
| *X-squared* | df | *p-value* |
| 8,048 | 8 | 0,4288 |

Diperoleh nilai $\left(\hat{C} = 8{,}048\right) < \left(\chi^2_{(0,05,8)} = 15{,}507\right)$ atau ($p$-*value*$= 0{,}4288) > (\alpha = 005)$. Dengan demikian, $H_0$ diterima, artinya model sesuai atau tidak ada perbedaan antara hasil observasi dengan hasil prediksi.

Berdasarkan hasil uji diatas diperoleh model akhir regresi logistik biner sebagai berikut:

$$g(x) = -1{,}268+0{,}734X_{1\_2}+0{,}886X_{2\_2} -1{,}266X_{3_2} -0{,}536X_{3_4}+0{,}082X_{3_5} -1{,}311X_{3_6} -14{,}117\ X_{3\_7} -1{,}182X_{3\_8}+14{,}742X_{3\_9}+0{,}273\ X_7$$

Untuk menghitung ketepatan klasifikasi, maka perlu dihitung probabilitas peserta KB dari data yang diolah. Diperoleh nilai ketepatan klasifikasi seperti pada Tabel 9.

Tabel 9. Ketepatan Klasifikasi

| Observasi | Prediksi | |
|---|---|---|
| | Bukan peserta KB | Peserta KB |
| Bukan peserta KB | 237 | 27 |
| Peserta KB | 91 | 35 |

$$APER(\%) = \frac{jumlah\ prediksi\ salah}{jumlah\ total\ prediksi} = \frac{n_{12}+n_{21}}{n_{11}+n_{12}+n_{21}+n_{22}} X\ 100\% = 30{,}2564\%$$

Persentase tingkat kesalahan klasifikasi (akurasi) pada model regresi logistik biner adalah 30,2564% , sehingga diperoleh nilai persentase ketepatan klasifikasi adalah $1 - APER = 1 - 0{,}302564 = 0{,}697436 = 69{,}7436\%$

Hasil dari model regresi logistik biner menyimpulkan bahwa variabel prediktor yang berpengaruh secara signifikan terhadap status peserta KB adalah usia istri $(X_1)$, usia suami $(X_2)$,

pendidikan istri ($X_3$), dan jumlah anak hidup ($X_7$). Langkah selanjutnya yaitu melakukan *resampling bootstrap* diperoleh hasil pengujian seperti pada Tabel 10.

Tabel 10 Ketepatan Klasifikasi dengan Bagging

| *Predicted Class* | *Observased Class* | |
|---|---|---|
| | 0 | 1 |
| 0 | 295 | 95 |
| 1 | 0 | 0 |

$$APER = 0{,}243590 = 24{,}359\%$$

Persentase tingkat kesalahan klasifikasi (akurasi) pada model regresi logistik biner adalah 24,359%, sehingga diperoleh nilai persentase ketepatan klasifikasi adalah $1 - APER = 1 - 0{,}24359 = 0{,}75641 = 75{,}641\%$.

Langkah selanjutnya membandingkan hasil ketepatan klasifikasi antara regresi logistik biner dan bagging (*bootstrap aggregating*) regresi logistik biner seperti pada Tabel 11.

Tabel 11. Perbandingan Ketepatan Klasifikasi

| Ketepatan Klasifikasi dengan Bagging | Ketepatan Klasifikasi regresi logistik biner | Peningkatan klasifikasi |
|---|---|---|
| 69,7436%. | 75,641%. | 5,9064% |

Berdasarkan Tabel 11 terlihat bahwa metode bagging (*bootstrap aggregating*) meningkatkan ketepatan klasifikasi dari model data set tunggal sebesar 69,7436% menjadi 75,641% sehingga peningkatan ketepatan klasifikasi sebesar 5,9064%. Dengan demikian ketepatan klasifikasi terbaik menggunakan bagging (*bootstrap aggregating*) logistik biner dengan model regresi logistik biner sebagai berikut:

$$g(x) = -1{,}8843 + 0{,}9594X_{1_2} + 1{,}1098X_{2_2} - 0{,}9515X_{3_2} - 1{,}0364X_{3_4} + 0{,}354X_{3_5} - 1{,}5016X_{3\_6} - 1{,}0836X_{3\_8} - 0{,}5354X_{3\_9} + 0{,}3501X_7$$

**SIMPULAN**

Berdasarkan analisis yang telah dilakukan dapat disimpulkan bahwa Hasil metode bagging (*bootstrap aggregating*) untuk meningkatkan ketepatan klasifikasi model regresi logistik biner pada status peserta KB Kota Tegal yaitu dengan menggunakan uji APER diperoleh kesalahan klasifikasi pada model bagging (*bootstrap aggregating*) regresi logistik biner sebesar 24,359% dengan ketepatan klasifikasi sebesar 75,641%. Ketepatan klasifikasi Ketepatan klasifikasi pada model regresi logistik biner sebesar 69,231%, dengan menggunakan metode bagging (*bootstrap aggregating*) ketepatan klasifikasi menjadi 75,641%. Dengan demikian, metode bagging meningkatkan ketepatan klasifikasi pada model regresi logistik biner sebesar 6,41%. Model terbaik regresi logistik biner menggunakan model bagging (*bootstrap aggregating*) regresi logistik biner adalah sebagai berikut:

$$g(x) = -1{,}8843 + 0{,}9594X_{1_2} + 1{,}1098X_{2_2} - 0{,}9515X_{3_2} - 1{,}0364\,X_{3_4} + 0{,}3545\,X_{3_5} - 1{,}5016\,X_{3\_6} - 1{,}0836\,X_{3\_8} - 0{,}5354X_{3\_9} + 0{,}3501X_7$$

Metode *bagging* (*bootstrap aggregating*) dapat diterapkan pada metode lain, misalnya CHAID karena hasil peningkatan ketepatan klasifikasi bekerja sangat baik pada model prediksi. Metode bagging memiliki ketepatan klasifikasi yang baik sehingga instansi dapat menggunakan metode tersebut untuk mengambil kebijakan. Variabel yang paling berpengaruh, berdasarkan hasil penelitian ini berturut-turut adalah usia suami, usia istri, pendidikan terakhir istri dan jumlah anak hidup. BKKBN dapat melakukan sosialisasi kepada masyarakat tentang pentingnya menggunakan KB dengan memperhatikan usia suami pada status peserta KB.

**DAFTAR PUSTAKA**


Alfaro, E., Gámez, M., & Garcia, N. (2013). adabag: An R package for classification with boosting and bagging. *Journal of Statistical Software*, 54(2), 1-35.

Arleina, O. D., & Otok, B. W. (2014). Bootstrap aggregating multivariate adaptive regression splines (Bagging MARS) untuk mengklasifikasikan rumah tangga miskin di kabupaten jombang. *Jurnal Sains dan Seni ITS*, 3(2), D91-D96.

Breiman, L. (1994). *Bagging Predictors. Technical report No 421*. Departement of statistics University of California.

Brown, I., & Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications*, 39(3), 3446-3453.

Hosmer, D.W. & Lemeshow, S. (2000). *Applied Logistic Regression*. USA: John Willey and Sons.

Kim, M. J., & Kang, D. K. (2012). Classifiers selection in ensembles using genetic algorithms for bankruptcy prediction. *Expert Systems with application*s, 39(10), 9308-9314.

Laza, R., Pavón, R., Reboiro-Jato, M., & Fdez-Riverola, F. (2011). Evaluating the effect of unbalanced data in biomedical document classification. *Journal of integrative bioinformatics*, 8(3), 105-117.

Liang, G. & Zhang, C. (2011). Empirical study of bagging predictors on medical data. *Journal of the centre for quantum computation & intelligent system,* 121.

Nai-arun, N. & Moungmai, R. (2015). Comparison of Classifiers for the Risk of Diabetes Prediction. *Procedia Computer Sains,* 69, 132-142.

Prasetio, R.T. & Pratiwi. (2015). Penerapan teknik bagging pada algoritma klasifikasi untuk mengatasi ketidakseimbangan kelas dataset medis. *Jurnal Informatika* **2**(2).