UNP JOURNAL OF STATISTICS AND DATA SCIENCE
Volume I Nomor 4 Tahun 2023 Halaman 256 -263.
ISSN(*Online*): 2985-475X

Comparison of Error Rate Prediction Methods in Binary Logistic Regression Model for Balanced Data

Shavira Asysyifa S, Dodi Vionanda*, Nonong Amalita, dan Dina Fitria
Departemen Statistika, Universitas Negeri Padang, Padang, Indonesia
*Corresponding author: dodi_vionanda@fmipa.unp.ac.id

Submitted : 18 Juli 2023
Revised : 05 Agustus 2023
Accepted : 08 Agustus 2023

ABSTRACT

Binary Logistic Regression is one of the statistical methods that can be used to see the relations between dependent variable with some independent variables, where the dependent variable split into two categories, namely the category declaring a successful event and the category declaring a failed event. The performance of binary logistic regression can be seen from the accurary of the model. Accuracy can be measured by predicting the error rate. One method that can be used to predict error rate is cross validation. The cross validation method works by dividing the data into two parts, namely testing data and training data. Cross validation has several learning methods that are commonly used, namely Leave One Out (LOO), Hold out, and K-fold cross validation. LOO has unbiased estimation of accuracy but take a long time, hold out can avoid overfitting and works faster because no iterations, and k-fold cross validation has smaller error rate prediction. Meanwhile, data cases with different correlation are useful to find out the different correlations effect performance of error rate prediction method. In this study uses artificially generated data with a normal distribution, including univariate, bivariate, and multivariate datasets with various combination of mean differences and correlation. Considering these factors, this study focuses on comparing the three cross validation methods for predicting error rate prediction in binary logistic regression. This study finds out that k-fold cross validation method is the most suitable method to predict errors in binary logistic regression modeling for balanced data.

Keywords: *Binary Logistic Regression, Hold Out, K-fold Cross Validation, Leave One Out*

I. PENDAHULUAN

Analisis regresi logistik biner merupakan metode yang digunakan untuk mengetahui hubungan antara variabel terikat dan variabel bebas, dimana variabel terikat bersifat biner terdiri ata dua kategori $Y = 1$ menyatakan kejadian sukses dan kategori $Y = 0$ menyatakan kejadian gagal (Agresti, 2007). Hasil analisis regresi logistik akan membentuk sebuah model. Model yang telah dibentuk perlu dinilai akurasinya. Salah satu cara yang digunakan yaitu dengan melihat laju galat yang diperoleh. Metode dalam memprediksi laju galat adalah metode *cross validation. Cross validation* digunakan untuk mengestimasi prediksi laju galat dalam meningkatkan ketepatan dari pemilihan model dan mengevaluasi kinerja suatu model.

Cross validation bekerja dengan cara membagi data menjadi dua bagian yakni data latih (*training*) dan data uji (*testing*). Data *training* digunakan untuk melatih model sehingga dapat dipahami pola datanya, sedangkan data *testing* digunakan untuk pengujian dan memprediksi model. *Cross validation* mengasumsikan bahwa data *training* dan data *testing* bersifat *independen*. Menurut Rafaeilzadeh dkk (2016), *Cross validation* memiliki beberapa metode diantaranya: *Leave One Out* (LOO), *Hold out,* dan *K-fold cross validation*. Perbedaan ketiga metode terletak dalam pemisahan data *training* dan data *testing*. LOO berkerja dengan cara membagi data menjadi n-1 pengamatan untuk data *training* dan 1 pengamatan tersisa untuk data *testing*. *Hold out* membagi data menjadi 2/3 data *training* dan 1/3 lainnya sebagai data *testing*. *K-fold cross validation* pada data *training* menggunakan k-1 *fold* pengamatan dan sisa 1 *fold* pengamatan sebagai data *testing*. Perbedaan dari kinerja ketiga metode prediksi laju galat menyebabkan adanya kekurangan dan kelebihan dari masing-masing metode dalam memprediksi laju galat. Oleh karena itu akan dilakukan perbandingan kinerja metode prediksi laju galat dalam pemodelan regresi logistik biner.

Perbandingan prediksi laju galat pada pemodelan regresi logistik biner dipengaruhi oleh beberapa faktor seperti jumlah variabel, variasi dari rataan populasi sampel, dan hubungan korelasi. Kurniawan dan Yuniarto (2016)

DOI: https://doi.org/10.24036/ujsds/vol1-iss4/90

menyebutkan bahwa korelasi merupakan hubungan antara variabel, dimana hubungan korelasi yang digunakan terbagi tiga macam yaitu tanpa korelasi, korelasi sedang dan korelasi tinggi. Korelasi akan memberikan dampak terhadap hasil prediksi, dimana pada regresi logistik ketika menggunakan korelasi tinggi akan menyebabkan multikolinearitas dan *overfitting.* Penelitian menggunakan data seimbang dalam melihat pengaruh metode *cross validation* terhadap regresi logistik. Berdasarkan penelitian yang dilakukan oleh Shelke dkk (2017) menyimpulkan bahwa sebagian besar algoritma pembelajaran memiliki kinerja lebih baik ketika kumpulan data hampir seimbang karena dalam data seimbang jumlah pengamatan pada tiap kelas populasi memiliki pengamatan yang sama. Tujuan dari penelitian ini adalah mengidentifikasi dan membandingkan kinerja dari masing-masing metode prediksi laju galat yakni *LOO*, *hold out*, dan *k-fold cross validation* sehingga diperoleh metode yang paling cocok diterapkan dalam pemodelan regresi logistik biner untuk data seimbang.

II. METODE PENELITIAN

A. Regresi Logistik Biner

Hosmer dan Lemeshow (2013), menyatakan pada regresi logistik biner variabel terikat bersifat dikotomi dan variabel bebas dapat berupa numerik maupun kategorik. Variabel dikotomi atau biner terdiri atas dua kategori saja seperti sukses dan gagal. Kategori sukses bernilai satu (1) dan kategori gagal bernilai nol (0). Pada regresi logistik tidak mengasumsikan hubungan linear antara variabel terikat dan variabel bebas. Abonazel dan Ibrahim (2018), menyatakan bahwa regresi logistik hampir mirip dengan regresi linier, perbedaan antara regresi logistik dan regresi linier terletak pada distribusi dan nilai prediksi. Pada regresi linier distribusi yang digunakan adalah distribusi normal dengan nilai prediksi berbentuk kontinu, sedangkan regresi logistik menggunakan distribusi Bernouli dengan nilai prediksi berbentuk probabilitas. Sehingga dikatakan bahwa regresi logistik merupakan kasus khusus dari model linier umum. Model dari regresi logistik:

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n)} \quad (1)$$

Dimana:
$\pi(x)$: probabilitas dari kejadian dengan nilai $0 \leq \pi(x) \leq 1$
n : banyak variabel bebas

Nilai probabilitas $\pi(x)$ berkisar antara nol dan satu. Untuk mempermudah penaksiran parameter dan mendapatkan fungsi linier maka dilakukan transformasi menggunakan transformasi logit. Menezes dkk (2017) memaparkan Transformasi logit harus dilakukan untuk menjamin nilai $\pi(x)$ akan selalu berada pada selang [0,1]. Hasil dari transformasi logit:

$$g(x) = \ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n \quad (2)$$

Nilai prediksi dilambangkan dengan $\hat{y}$ dan nilai asli dilambangkan dengan y. Nilai probabilitas $\pi(x)$ merupakan nilai prediksi dari regresi logistik, namun pada regresi logistik biner nilai prediksinya adalah $g(x)$.

Korelasi pada regresi logistik dapat berpengaruh terhadap kinerja estimasi, terutama ketika korelasi yang digunakan merupakan korelasi tinggi. Korelasi yang tinggi pada regresi logistik dapat menyebabkan terjadinya multikolinearitas dan *overfitting*. Korelasi tinggi pada variabel bebas dapat menyulitkan dalam estimasi koefisien model sehingga estimasi koefisien tidak stabil dan sulit untuk diinterpretasikan. Courvoisier (2011) menjelaskan bahwa saat memasangkan model regresi logistik dengan korelasi tinggi dapat menyebabkan bias terlalu tinggi terhadap variabel terikat, korelasi tinggi juga dapat mengurangi ketepatan dalam estimasi model.

B. Prediksi Laju Galat

Galat atau *error* merupakan selisih antara data aktual dengan data prediksi. Galat didefenisikan sebagai ukuran yang digunakan untuk menilai keakuratan sebuah model yang telah dibangun menggunakan dataset dalam memprediksi data baru. Prediksi laju galat yang dilakukan terhadap sebuah model berarti mengukur kinerja dari suatu model dengan menghitung tingkat kesalahan prediksi model tersebut. Prediksi laju galat digunakan untuk pemilihan model karena masuk akal untuk memilih model yang memiliki akurasi prediksi terendah diantara analisis lainnya. Metode untuk mengestimasi prediksi laju galat dalam meningkatkan ketepatan dari pemilihan model adalah *cross validation.*

Menurut Yadav dan Shukla (2016), tujuan *cross validation* adalah memprediksi kinerja model yang dipelajari dari data yang tersedia menggunakan suatu metode, kemudian membandingkan kinerja dari beberapa metode sehingga

DOI: https://doi.org/10.24036/ujsds/vol1-iss4/90

ISSN(*Online*): 2985-475X

diperoleh metode yang cocok untuk data tersebut. Dalam mengukur galat atau *error rate* dapat menggunakan indikator variabel (James, 2013) dengan rumus:

$$err_i = I\,(y_i \neq \hat{y}_i) \tag{3}$$

Dimana:

y_i : data aktual amatan ke-i, dimana $i = 1,2,\dots,n$

$\hat{y}_i$: data hasil prediksi ke-i, dimana $i = 1,2,\dots,n$

Dengan indikator variabel bernilai 1 jika $I\,(y_i \neq \hat{y}_i)$ dan bernilai 0 jika $I\,(y_i = \hat{y}_i)$.

C. *Leave One Out* (LOO)

Menurut James dkk (2013), LOO merupakan metode dalam prediksi laju galat yang membagi data pengamatan menjadi data *training* dan data *testing* berdasarkan jumlah pengamatan. Proses pengulangan perhitungan dapat dilakukan sebanyak n pengamatan. Setiap pengulangan akan memiliki semua data kecuali satu pengamatan yang akan digunakan untuk data *testing*. Sehingga pada LOO tidak perlu dilakukan pengacakan data. Model yang diperoleh dari data *training* akan diuji menggunakan data *testing*. Rafaeilzadeh (2016) dalam penelitiannya mengemukakan kelebihan dan kekurangan dari LOO. Kelebihan LOO adalah estimasi kinerja yang dihasilkan hampir tidak bias. Namun kekurangan dari LOO memakan waktu yang cukup lama, dan menghasilkan variansi yang lebih tinggi. Rumus dalam menghitung nilai *error rate* prediksi laju galat menggunakan metode LOO:

$$\hat{E}^{LOO} = \frac{1}{n}\sum_{i=1}^{n} I\,(y_i \neq \hat{y}_i) \tag{4}$$

D. *Hold out*

Menurut James dkk (2013), *Hold out* merupakan metode prediksi laju galat yang membagi data secara acak. Pembagian data dilakukan secara acak namun dengan ketentuan jumlah data *training* harus lebih banyak dari pada data *testing*. Rafaeilzadeh (2016), memaparkan metode *hold out* digunakan untuk menghilangkan masalah *overfitting* pada data. *Hold out* menghindari tumpang tindih antara data *training* dengan data *testing*, sehingga dapat menghasilkan perkiraan tingkat kinerja lebih akurat. Namun kekurangan *hold out* adalah hasil akhir sangat bergantung kepada pemisahan data *training* dan data *testing.* Diantara metode prediksi galat lainnya, *hold out* merupakan metode yang paling sederhana dimana metode ini mengambil data secara acak dalam data pengamatan kemudian menjadikannya data *training*, dan sisa pengamatan menjadi data *testing*. Rumus menghitung nilai *error rate* prediksi galat metode *hold out*.

$$\hat{E}^{HO} = \frac{1}{n_{test}}\sum_{i=1}^{n_{test}} I\,(y_i \neq \hat{y}_i) \tag{5}$$

E. *K-Fold Cross Validation*

James dkk (2013) menyatakan, *K-fold cross validation* merupakan metode prediksi laju galat dengan cara membagi data ke dalam k grup dengan sampel pengamatan yang sama. Pada data *training* akan digunakan sebanyak $k-1$ grup dan sisanya 1 grup akan menjadi data *testing*. Pada setiap data *training* akan dihitung rata-rata *error* grup. Pengulangan akan dilakukan sebanyak *k-fold* yang telah ditentukan. Pembagian umum yang dilakukan pada penelitian lainnya seperti 5 *fold,* 7 *fold,* dan 10 *fold*. Dalam penelitian Rafaeilzadeh (2016), kelebihan dari metode *k-fold cross validation* menghasilkan estimasi kinerja yang lebih baik dan variansi yang lebih kecil dengan hasil yang lebih akurat. Namun kekurangan metode *k-fold cross validation* kurang cocok digunakan pada data dengan sampel kecil. Menurut Wong (2014), terdapat empat faktor yang mempengaruhi perkiraan akurasi yang diperoleh dengan menggunakan metode *k-fold cross validation* yaitu jumlah *fold,* jumlah pengamatan dalam satu *fold*, tingkat rata-rata, dan pengulangan dari *k-fold*. Menurut James dkk (2013) rumus dalam menghitung nilai *error rate* prediksi galat pada metode *k-fold cross validation* dengan 5 kelompok dapat dilihat pada persamaan 6.

$$\hat{E}^{CV} = \frac{1}{5}\sum_{k=1}^{5}\frac{1}{n_k}\sum_{i=1}^{n_k} I\,(y_i \neq \hat{y}_i) \tag{6}$$

F. Jenis Penelitian dan Sumber Data

Jenis penelitian merupakan penelitian eksperimen. Payadnya dan Jayantika (2018) menyebutkan penelitian eksperimen ditujukan untuk meneliti hubungan sebab akibat dengan memanipulasi satu atau lebih variabel pada sebuah kelompok eksperimental kemudian membandingkan hasilnya dengan kelompok yang tidak mengalami manipulasi.

DOI: https://doi.org/10.24036/ujsds/vol1-iss4/90

Jenis data yang digunakan pada penelitian ini merupakan data simulasi. Data simulasi diperoleh dari data bangkitan pada *software R-studio version 4.1.1*. Simulasi data pada penelitian ini bertujuan untuk membandingkan kinerja dari metode prediksi laju galat dalam pemodelan regresi logistik biner untuk kasus data seimbang. Metode prediksi laju galat yang digunakan LOO*, Hold out* dan *K-fold cross validation*. Dari nilai laju galat yang diperoleh akan dibandingkan kinerja dari metode prediksi laju galat dengan memperhatikan nilai median dan variansi dari *error rate*.

Pemodelan regresi logistik biner menggunakan variabel terikat berbentuk biner yang bernilai satu (1) dan nol (0). Sedangkan untuk variabel bebas dibangkitkan dengan beberapa ketentuan, diantaranya yaitu jumlah variabel bebas, perbedaan rataan populasi dari sampel berasal, dan korelasi pada variabel bivariat dan multivariat.Variabel bebas yang digunakan terdiri atas satu variabel bebas (univariat), dua variabel bebas (bivariat), dan tiga variabel bebas (multivariat). Untuk kasus data bivariat dan multivariat melakukan simulasi dengan melibatkan struktur korelasi dan struktur rataan yang berbeda untuk rataan populasi. Menurut Sukertiyarno dan Agoestanto (2017) kasus data univariat dapat dibangkitkan dari distribusi normal $N\ (\mu, \sigma)$ dengan perbedaan rataan populasi seperti pada Tabel 1.

Tabel 1. Ketentuan Nilai Rataan Populasi Data Univariat

Jumlah variabel	Kasus	Rataan	
		$\boldsymbol{\mu^{(1)}}$	$\boldsymbol{\mu^{(2)}}$
Univariat	1	0	1
	2	0	2

Menurut Tong (1990) perbedaan rataan populasi pada kasus data bivariat dapat dibangkitkan dengan ketentuan pada Tabel 2.

Tabel 2. Ketentuan Nilai Rataan Populasi Data Bivariat

Jumlah variabel	Kasus	Struktur rataan populasi	
		$\boldsymbol{\mu^{(1)}}$	$\boldsymbol{\mu^{(2)}}$
Bivariat	1	$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{pmatrix}1\\1\end{pmatrix}$
	2	$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{pmatrix}1\\0\end{pmatrix}$
	3	$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{pmatrix}2\\2\end{pmatrix}$
	4	$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{pmatrix}2\\0\end{pmatrix}$

Berdasarkan Tabel 2 untuk data bivariat dengan kasus 1 dan kasus 3 kedua variabel bebas dinamakan variabel relevan karena variabel memuat informasi tentang perbedaan kelas populasi. Sedangkan, kasus 2 dan kasus 4 hanya variabel pertama yang merupakan variabel relevan dan variabel kedua dinamakan variabel irrelevan. Dinamakan variabel irrelevan karena variabel yang kedua tidak memuat informasi tentang perbedaan kelas populasi. Korelasi merupakan angka yang menunjukkan kuatnya hubungan variabel yang diteliti. Kurniawan dan Yuniarto (2016) menafrsirkan bahwa korelasi bernilai 0 berarti hubungan variabel sangat kecil dan dianggap tidak ada korelasi, korelasi bernilai 0.5 berarti memiliki hubungan yang sedang, dan korelasi bernilai 0.9 berarti memiliki korelasi yang erat. Pengaturan hubungan korelasi pada kasus data bivariat dipaparkan pada Tabel 3.

Tabel 3. Struktur Korelasi pada Data Bivariat

korelasi	Struktur Korelasi
A	$\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}$
B	$\begin{bmatrix}1 & 0.5\\0.5 & 1\end{bmatrix}$
C	$\begin{bmatrix}1 & 0.9\\0.9 & 1\end{bmatrix}$

Pada Tabel 3 korelasi A menunjukkan hubungan tanpa korelasi, korelasi B menunjukkan hubungan korelasi sedang, dan korelasi C menunjukkan hubungan korelasi tinggi. Kemudian untuk pembangkitan data, pengaturan struktur korelasi digabung dengan pengaturan struktur rataan populasi sehingga dapat dikaji kondisi data dengan variabel relevan yang tidak memiliki korelasi, memiliki korelasi sedang atau korelasi tinggi, dan kondisi data dimana terdapat variabel relevan dan variabel irrelevan yang tidak memiliki korelasi, memiliki korelasi sedang atau korelasi tinggi. Menurut

DOI: https://doi.org/10.24036/ujsds/vol1-iss4/90

Tong (1990) perbedaan rataan populasi pada kasus data multivariat dapat dibangkitkan dengan ketentuan perbedaan pada Tabel 4.

Tabel 4. Ketentuan Nilai Rataan Data Multivariat

Jumlah variabel	Kasus	Struktur rataan populasi	
		$\mu^{(1)}$	$\mu^{(2)}$
Multivariat	1	$\begin{pmatrix}0\\0\\0\end{pmatrix}$	$\begin{pmatrix}1\\1\\1\end{pmatrix}$
	2	$\begin{pmatrix}0\\0\\0\end{pmatrix}$	$\begin{pmatrix}1\\1\\0\end{pmatrix}$
	3	$\begin{pmatrix}0\\0\\0\end{pmatrix}$	$\begin{pmatrix}1\\0\\1\end{pmatrix}$
	4	$\begin{pmatrix}0\\0\\0\end{pmatrix}$	$\begin{pmatrix}1\\0\\0\end{pmatrix}$
	5	$\begin{pmatrix}0\\0\\0\end{pmatrix}$	$\begin{pmatrix}0\\0\\1\end{pmatrix}$
	6	$\begin{pmatrix}0\\0\\0\end{pmatrix}$	$\begin{pmatrix}2\\2\\2\end{pmatrix}$
	7	$\begin{pmatrix}0\\0\\0\end{pmatrix}$	$\begin{pmatrix}2\\2\\0\end{pmatrix}$
	8	$\begin{pmatrix}0\\0\\0\end{pmatrix}$	$\begin{pmatrix}2\\0\\2\end{pmatrix}$
	9	$\begin{pmatrix}0\\0\\0\end{pmatrix}$	$\begin{pmatrix}2\\0\\0\end{pmatrix}$
	10	$\begin{pmatrix}0\\0\\0\end{pmatrix}$	$\begin{pmatrix}0\\0\\2\end{pmatrix}$

Pada data multivariat kasus 1 dan kasus 6 ketiga variabel merupakan variabel relevan. Sementara kasus 2, kasus 3, kasus 7, dan kasus 8 terdiri atas 2 variabel relevan dan 1 variabel irrelevan, dan sisanya kasus 4, kasus 5, kasus 9 dan kasus 10 terdiri atas 1 variabel relevan dan 2 variabel irrelevan. Sama seperti data bivariat, hubungan korelasi yang digunakan juga terdiri atas 0, 0.5 dan 0.9, namun pada data multivariat juga memungkinkan melihat hubungan korelasi yang terjadi antar 2 variabel bebas. Menurut Kurniawan dan Yuniarto (2016) pengaturan hubungan korelasi yang digunakan pada kasus data multivariat dipaparkan pada Tabel 5.

Tabel 5. Struktur korelasi pada data multivariat

Korelasi	Struktur Korelasi
A	$\begin{bmatrix}1 & 0 & 0\\0 & 1 & 0\\0 & 0 & 1\end{bmatrix}$
B	$\begin{bmatrix}1 & 0.5 & 0\\0.5 & 1 & 0\\0 & 0 & 1\end{bmatrix}$
C	$\begin{bmatrix}1 & 0.9 & 0\\0.9 & 1 & 0\\0 & 0 & 1\end{bmatrix}$
D	$\begin{bmatrix}1 & 0.5 & 0.5\\0.5 & 1 & 0.5\\0.5 & 0.5 & 1\end{bmatrix}$
E	$\begin{bmatrix}1 & 0.9 & 0.9\\0.9 & 1 & 0.9\\0.9 & 0.9 & 1\end{bmatrix}$

DOI: https://doi.org/10.24036/ujsds/vol1-iss4/90

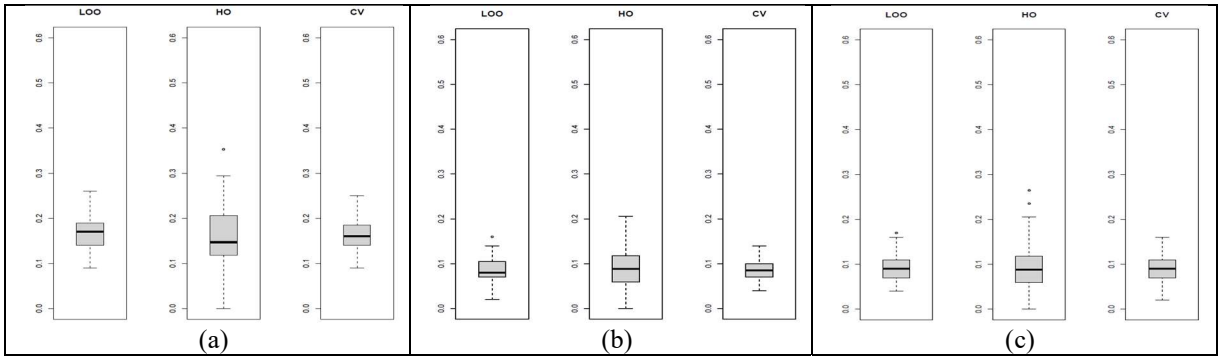
ISSN(*Online*): 2985-475X

Pada Tabel 5 korelasi A menunjukkan hubungan tanpa korelasi, korelasi B menunjukkan hubungan korelasi sedang yang terjadi antara 2 variabel x_1x_2, korelasi C menunjukkan hubungan korelasi tinggi antara variabel x_1x_2, korelasi D menunjukkan hubungan korelasi sedang, dan korelasi E menunjukkan hubungan korelasi tinggi. Semua kasus data univariat, bivariat, maupun multivariat dibangkitkan dari distribusi normal dengan variansi satu $(\sigma = 1)$.

III. HASIL DAN PEMBAHASAN

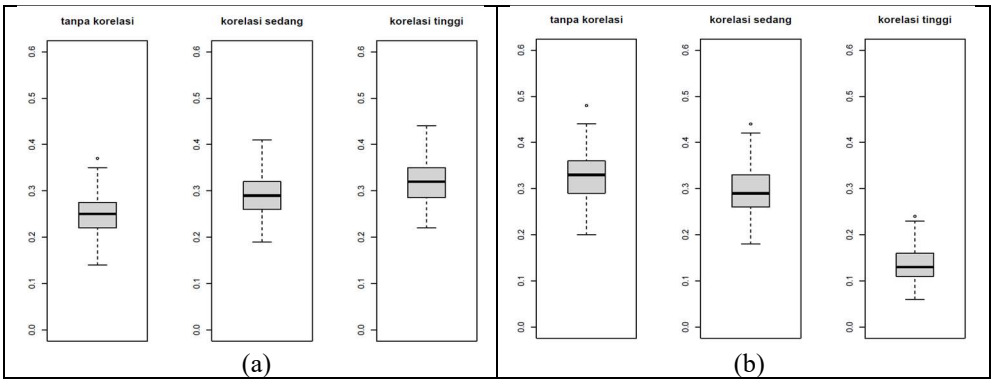
A. Hasil Analisis

Tujuan dari penelitian ini adalah melihat metode prediksi laju galat yang paling cocok diterapkan pada regresi logistik dengan data seimbang. Metode prediksi laju galat dapat dilihat dari variasi *error rate* yang kecil. Variasi *error rate* dapat dilihat berdasarkan *boxplot* yang dihasilkan. Selain melihat variasi terkecil, dapat melihat pengaruh antara variasi data dan hubungan korelasi terhadap prediksi laju galat. Pada Gambar 1 dapat dilihat perbedaan hasil *error rate* pada setiap jenis data.



Gambar 1. Boxplot hasil prediksi galat pada data (a) Univariat (b) Bivariat (c) Multivariat dengan kasus 1

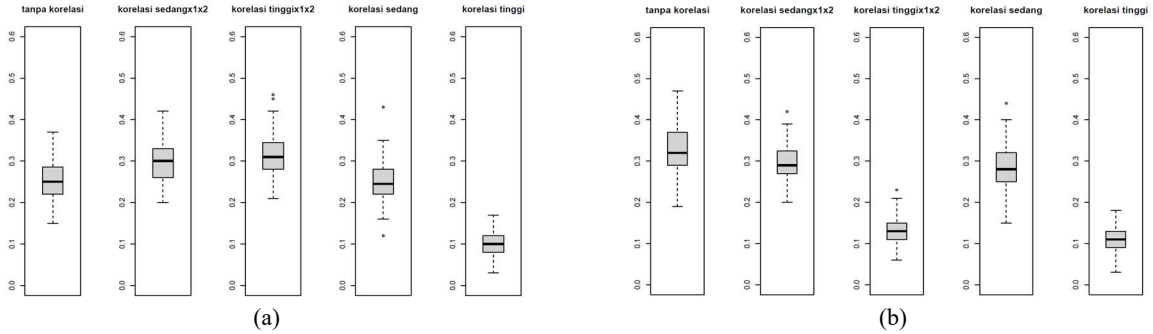
Percobaan pada Gambar 1 dilakukan untuk melihat struktur prediksi laju galat pada masing-masing data dengan mengabaikan nilai korelasinya. Gambar 1 juga memperlihatkan bahwa metode *hold out* menghasilkan variasi yang lebih besar dibandingkan LOO dan *k-fold cross validation*. Dari variasi *error rate* yang diperoleh menunjukkan bahwa metode *k-fold cross validation* memiliki kinerja yang lebih baik dibandingkan LOO dan *hold out* dalam memprediksi laju galat. Meskipun variasi *error rate* yang dihasilkan oleh LOO dan *k-fold cross validation* memiliki ukuran variasi yang hampir sama. Namun, jika dilihat dari seluruh percobaan baik pada data univariat, bivariat, maupun data multivariat, *k-fold cross validation* lebih dominan menunjukkan variasi yang lebih kecil dibandingkan metode prediksi galat lainnya. Sehingga dapat dikatakan bahwa *k-fold cross validation* lebih cocok digunakan pada pemodelan regresi logistik pada data seimbang.



Gambar 2. Hasil *error rate* data bivariat dengan (a) sesama variabel relevan (b) variabel relevan dan variabel irrelevan dari metode *k-fold cross validation*

DOI: https://doi.org/10.24036/ujsds/vol1-iss4/90

Pada Gambar 2 dapat dilihat Pengaruh variabel relevan dan variabel irrelevan pada data bivariat. Pada data bivariat dan multivariat pengaruh korelasi antar hubungan variabel merupakan hal yang harus diperhatikan. Perbedaan nilai korelasi juga mempengaruhi hasil prediksi laju galat. Percobaan yang dilakukan pada sesama variabel relevan ketika berkorelasi menghasilkan nilai *error rate* yang semakin besar ketika korelasinya juga meningkat. Hal ini dapat dilihat pada Gambar 2 (a) yang memuat percobaan setting 1 dan setting 3. Pada Gambar 2 (b) merupakan percobaan hubungan korelasi yang terjadi pada setting 2 dan setting 4 terjadi antara variabel relevan dengan variabel irrelevan menyebabkan semakin tinggi nilai korelasi menghasilkan nilai median *error rate* semakin kecil.



(a) (b)

Gambar 3. Hasil *error rate* data multivariat antara (a) 2 variabel relevan dengan 1 variabel irrelevan dan (b) 1 variabel relevan dengan 2 variabel irrelevan dari metode *k-fold cross validation*

Pada Gambar 3 merupakan pengaruh hubungan antara variabel relevan dengan variabel irrelevan data multivariat. Gambar 3 dapat dilihat ketika hubungan korelasi terjadi antara data yang memiliki 1 variabel relevan dan 2 variabel irrelevan menghasilkan nilai *error rate* yang lebih besar dibandingkan dengan data multivariat dengan 2 variabel relevan dan 1 variabel irrelevan. Pada data multivariat ketika ketiga variabel yang berkorelasi merupakan variabel relevan maka semakin tinggi nilai korelasi yang digunakan akan menghasilkan nilai median *error rate* yang semakin besar, namun jika terdapat salah satu variabel merupakan variabel irrelevan maka *error rate* yang dihasilkan akan menurun. Data yang memuat variabel relevan dan variabel irrelevan ketika berkorelasi semakin besar korelasi maka nilai *error rate* akan semakin kecil.

B. Pembahasan

Dari hasil analisis yang telah dilakukan, kinerja dari masing-masing metode prediksi galat memiliki hasil yang berbeda-beda. Kinerja *hold out* menghasilkan variasi yang besar dibandingkan *k-fold cross validation* dan LOO. Namun median *error rate* yang dihasilkan *hold out* lebih kecil dibandingkan dengan metode lain. *Hold out* juga banyak menghasilkan *outlier* pada setiap percobaan, karena *hold out* cenderung lebih tidak stabil dalam memprediksi laju galat. Hal ini dapat disebabkan karena pemisahan data *training* dan data *testing* yang hanya dilakukan sekali tanpa perulangan. Sehingga *hold out* memuat model dengan jumlah data *training* yang lebih sedikit, dengan kata lain model yang dibangun menggunakan lebih sedikit data dari pada LOO maupun *k-fold cross validation.*

Kinerja yang dihasilkan oleh *k-fold cross validation* dan LOO memiliki kinerja yang hampir sama dalam memprediksi galat. Pada data univariat dan bivariat, metode *k-fold cross validation* menunjukkan hasil yang lebih baik. Pada data multivariat terdapat beberapa percobaan yang menunjukkan LOO lebih baik, namun secara keseluruhan dilihat dari nilai median *error rate* dan variasi yang dihasilkan*, k-fold cross validation* menghasilkan variasi yang lebih kecil dibandingkan LOO. Sehingga dikatakan *k-fold cross validation* adalah metode yang paling cocok digunakan dalam pemodelan regresi logistik biner. Metode *k-fold cross validation* menghasilkan variasi yang lebih kecil karena data yang dimiliki dibagi kedalam kelompok, kemudian dalam kelompok tersebut dilakukan perulangan, sehingga menghasilkan hasil yang lebih akurat. Hal ini juga sesuai dengan kelebihan dari *k-fold cross validation* yaitu menghasilkan nilai variasi yang lebih kecil dengan hasil lebih akurat.

Penelitian ini menggunakan data univariat, bivariat dan multivariat. Masing-masing data dibangkitkan dengan struktur rataan populasi yang berbeda. Pada data bivariat dan multivariat data juga dibangkitkan dengan struktur korelasi yang berbeda pula. Secara keseluruhan ketika nilai rataan kelasnya dibedakan akan mempengaruhi hasil *error rate*. Pada data bivariat dan multivariat, perbedaan nilai korelasi mempengaruhi hasil prediksi laju galat. Selain itu, akan ada perbedaan nilai *error rate* ketika korelasi dilakukan antar sesama variabel relevan dengan percobaan ketika korelasi

DOI: https://doi.org/10.24036/ujsds/vol1-iss4/90

dilakukan antar variabel relevan dan variabel irrelevan. Ketika terjadi peningkatan korelasi akan disertai dengan peningkatan nilai *error rate*. Hubungan korelasi tinggi akan menghasilkan nilai *error rate* yang lebih besar dibandingkan nilai *error rate* yang dilakukan tanpa korelasi. Sedangkan ketika percobaan dilakukan antara variabel relevan dengan variabel irrelevan menyebabkan terjadi peningkatan nilai korelasi akan disertai penurunan nilai *error rate*.

IV. KESIMPULAN

Kinerja LOO dan *k-fold cross validation* pada data seimbang memiliki kinerja yang hampir sama dalam memprediksi laju galat. Namun metode *k-fold cross validation* memiliki variasi *error rate* yang lebih kecil dibandingkan metode LOO. Sehingga metode yang paling cocok digunakan untuk memprediksi laju galat dalam pemodelan regresi logistik biner dengan data seimbang adalah metode *k-fold cross validation*. Pada data bivariat dan multivariat ketika korelasi dilakukan antar sesama variabel relevan, saat terjadinya peningkatan korelasi akan disertai dengan peningkatan nilai *error rate*. Sebaliknya, ketika korelasi dilakukan antar variabel relevan dengan variabel irrelevan, saat terjadinya peningkatan korelasi akan disertai dengan penurunan nilai *error rate*. Pada penelitian selanjutnya dapat melakukan perbandingan kinerja metode *k-fold cross validation fold=5* dengan metode metode *k-fold cross validation fold =10*, ataupun penggunaan metode ini dapat diimplementasikan pada kasus data lainnya.

DAFTAR PUSTAKA

Abonazel, M., dan Ibrahim, M. (2018). On Estimation Methos for Binary Logistic Regression Model with Missing Values. *International Journal of Mathematics and Computational Science.* Vol. 4(3):79-85.

Agresti, A. (2007). *An Introduction to Categorical Data Analysis*. New York: John Wiley and Sons.

Courvoisier, D. S., Combescure, C., Agoritsas, T., Gayet-Ageron, A., & Perneger, T. V. (2011). Performance of logistic regression modeling: beyond the number of events per variable, the role of data structure. Journal of clinical epidemiology, 64(9), 993-1000.

Hosmer, D.W., dan S. Lemeshow. (2013). *Applied Logistic Regression*. Edisi ke-3, New Jersey: Canada.

James, G., Witten, D., Hastie, T., Tibshirani R. (2013). *An Introduction to Statistical Learning,* New York: Springer.

Kurniawan, R., dan Yuniarto, B. (2016). *Analisis Regresi Dasar dan Penerapannya dengan R*. Jakarta: Kencana.

Menezes, F. S., Liska, G. R., Cirillo, M. A., Vivanco, M. J. (2017). Data Classification with Binary Response through the Boosting Algorithm and Logistic Regression. *Expert Systems with Application*, vol 69: 63-65.

Payadnya, I., Jayantika, I. (2018). *Panduan penelitian eksperimen beserta analisis statistic dengan SPSS.* Yogyakarta: Deepublish.

Refaeilzadeh, P., Tang, L., Liu, H. (2016). Cross Validation. *Encyclopedia of Database Systems.* DOI:10.1007/978-1-4899-799-3_565-2.

Shelke, M., Deshmukh, P., Shandilya, V. (2017). A Review on Imbalanced Data Handling Using Undersampling and Oversampling Technique. *International Journal of Recent Trends in Engineereing & Research*. Vol. 3. ISSN: 2455-1457. DOI: 10.23883/IJRTER.2017.3168.0UWXM.

Sukestiyarno, Y. L., dan Agoestanto, A. (2017). Batasan prasyarat uji normalitas dan uji homogenitas pada model regresi linier. *Unnes Journal of Mathematics*, 6(2),168-177.

Tong, Yuang L., dan Tong, Y. L. (1990). *Fundamental properties and sampling distributions of the multivariate normal distribution*. Pp:23-61. Springer New York.

Wong., T., T. (2015). Performance Evaluation of Classification Algorithms by K-Fold and Leave-One-Out Cross Validation. *Journal of Elsevier*, pp:0031-3203.

Yadav, S., Shukla, S. (2016). Analysis of K-fold Cross Validation Over Hold-Out Validation on Colossal datasets for Quality Classsification. *IEEE 6th International Advanced Computing,* pp: 788-813.

DOI: https://doi.org/10.24036/ujsds/vol1-iss4/90