UNP JOURNAL OF STATISTICS AND DATA SCIENCE
Volume I Nomor 4 Tahun 2023 Halaman 271 -279.
ISSN(*Online*): 2985-475X

Comparison of Error Rate Prediction Methods in Classification Modeling with Classification and Regression Tree (CART) Methods for Balanced Data

Fitria Panca Ramadhani, Dodi Vionanda*, Syafriandi, Admi Salma
Departemen Statistika, Universitas Negeri Padang, Padang, Indonesia
*Corresponding author: dodi_vionanda@fmipa.unp.ac.id

Submitted : 12 Juni 2023
Revised : 28 Juli 2023
Accepted : 08 Agustus 2023

ABSTRACT

CART (Classification and Regression Tree) is one of the classification algorithms in the decision tree method. The model formed in CART is a tree consisting of root nodes, internal nodes, and terminal nodes. After the model is formed, it is necessary to calculate its accuracy. The aim is to see the performance of the model. The accuracy of this model can be determined by calculating the predicted error rate in the model. The error rate prediction method works by dividing the data into training data and testing data. There are three methods in the error rate prediction method: Leave One Out Cross Validation (LOOCV), Hold Out (HO), and K-Fold Cross Validation. These methods have different performance in dividing data into training data and testing data, so there are advantages and disadvantages to each method. Therefore, a comparison was made between the three error rate prediction methods with the aim of determining the appropriate method for the CART algorithm. This comparison was made by considering several factors, for instance, variations in the mean, the number of variables, and correlations in normally distributed random data. The results of the comparison will be observed using a boxplot by looking at the median error rate and the lowest variance. The results of this study indicate that the K-Fold Cross Validation method has the lowest median error rate and the lowest variance, so the most suitable error prediction method for the CART method is the K-Fold Cross Validation method.

Keywords: *Classification and Regression Tree, Hold Out, K-Fold Cross Validation, Leave One Out Cross Validation*

I. PENDAHULUAN

Classification and Regression Tree (CART) merupakan algoritma pohon keputusan dengan pelabelan berdasarkan kelas pada variabel dependennya. Model yang dihasilkan oleh CART berupa pohon yang berbentuk seperti diagram alur. Akurasi model pada CART dapat diukur dengan menghitung prediksi laju galat pada model. Kinerja model ini dapat dihitung menggunakan metode *test error rate.* Metode *test error rate* bekerja dengan membagi data menjadi dua bagian yang masing-masing berfungsi untuk membentuk model dan menguji akurasi model.

Metode *test error rate* disebut juga dengan metode *cross validation*. Menurut Hastie dkk (2008: 241), *cross validation* merupakan salah satu metode yang paling sederhana dan paling banyak digunakan untuk memperkirakan prediksi galat pada model. Metode ini terbagi kedalam tiga macam yaitu *Leave One Out Cross Validation (LOOCV),* metode *Hold Out* (HO), dan metode *K-Fold Cross Validation.* Perbedaan dari ketiga metode ini terletak pada pembagian data untuk data *training* dan data *testing*. LOOCV bekerja dengan setiap pengamatan berperan sebagai data *training* dan data *testing*, HO membagi data menjadi 2/3 sebagai data *training* dan 1/3 menjadi data *testing*, sementara itu *K-Fold* mengelompokkan data secara acak terlebih dahulu kemudian membaginya menjadi dua bagian. Perbedaan kinerja ketiga metode prediksi galat ini menyebabkan adanya kekurangan dan kelebihan pada masing-masing metode dalam memprediksi galat pada model sehingga dilakukan perbandingan untuk menentukan metode prediksi galat terbaik pada metode CART.

Perbandingan prediksi laju galat pada CART dipengaruhi oleh beberapa faktor seperti korelasi, variasi rataan, dan jumlah variabel. Kurniawan dan Yuniarto (2016) menyebutkan bahwa nilai prediksi akan semakin besar jika nilai koefisien korelasi semakin kecil. Dalam hal ini, rentang koefisien korelasi terbagi menjadi tiga macam yaitu tidak adanya korelasi, korelasi sedang, dan korelasi tinggi. Korelasi ini akan dikombinasikan dengan variasi rataan pada

DOI: https://doi.org/10.24036/ujsds/vol1-iss4/73

ISSN(*Online*): 2985-475X

jumlah variabel bivariat dan multivariat sehingga dapat diamati pengaruhnya dengan membandingkan masing-masing metode prediksi laju galat.

Pada penelitian terdahulu seperti penelitian yang dilakukan oleh Kohavi (1995) yang membandingkan HO, LOOCV, *K-Fold,* dan *bootstrap* didapatkan kesimpulan bahwan *ten-fold cross validation* merupakan metode terbaik dalam seleksi model. Kesimpulan yang sama juga disebutkan oleh Payam dkk (2016) pada penelitiannya yang meneliti tentang perbandingan metode *resubstitution validation,* HO, LOOCV, dan *K-Fold* dalam menyeleksi model klasifikasi. Berdasarkan pemaparan di atas, maka dalam artikel ini akan dibahas tentang perbandingan metode prediksi galat dalam pemodelan klasifikasi dengan metode CART untuk data seimbang.

II. METODE PENELITIAN

Metode pada penelitian ini dimulai dari pembangkitan data menggunakan R Studio kemudian dilanjutkan dengan pembagian data menjadi data *training* dan data *testing* yang disesuaikan dengan proses masing-masing metode prediksi laju galat, setelahnya dilakukan pembentukan model pada metode CART menggunakan data *training*. Tahap akhir penelitian ini yaitu validasi pada model menggunakan data *testing*. Berikut merupakan penjelasan dari tahapan penelitian pada tulisan ini yaitu sebagai berikut.

A. Pembangkitan Data

Jenis data pada penelitian ini berupa data simulasi menggunakan data acak berdistribusi normal $N(\mu, \sigma)$ yang dibangkitkan melalui *software R Studio* sebanyak 100 amatan. Simulasi data pada penelitian ini bertujuan untuk membandingkan kinerja metode prediksi laju galat pada pemodelan CART untuk kasus data seimbang. Metode prediksi laju galat yang dibahas pada penelitian ini yaitu metode LOOCV, HO, dan metode *K-Fold Cross Validation.* Dari nilai laju galat yang diperoleh, akan diperhatikan nilai median dan variansi dari masing-masing metode prediksi laju galat sebagai perbandingan. Sementara itu, pemodelan CART pada penelitian ini merupakan pemodelan CART biner sehingga variabel dependen yang dibangkitkan bernilai 1 dan 2. Sedangkan variabel independen menggunakan data simulasi dengan memperhatikan beberapa aspek yaitu jumlah variabel independen yang digunakan, perbedaan rataan populasi berdasarkan pengambilan sampel, dan perbedaan korelasi antara variabel untuk kasus bivariat dan multivariat. Jumlah variabel independen terdiri dari satu variabel (univariat), dua variabel (bivariat), dan tiga variabel (multivariat). Pada kasus bivariat dan multivariat diaplikasikan struktur korelasi dan struktur rataan yang berbeda. Sementara itu perbedaan rataan populasi univariat dibangkitkan dari distribusi normal yang dijabarkan pada Tabel 1 berikut.

Tabel 1. Ketentuan Data Bangkitan untuk Data Univariat

Jumlah Variabel	Pengaturan	Rataan	
		$\mu^{(1)}$	$\mu^{(2)}$
Univariat	1	0	1
	2	0	2

Tabel 1 merupakan ketentuan aturan data bangkitan untuk jenis data univariat dengan variansi 1. Pada Tabel 1 terdapat nilai rataan pada dua kelas data berbeda dengan nilai simpangan baku yang sama. Sedangkan pengaturan untuk perbedaan rataan populasi untuk kasus bivariat dan multivariat disajikan pada Tabel 2 berikut.

Tabel 2. Ketentuan Data Bangkitan untuk Data Bivariat dan Data Multivariat

Pengaturan	Bivariat		Multivariat	
	$\mu^{(1)}$	$\mu^{(2)}$	$\mu^{(1)}$	$\mu^{(2)}$
1	$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{pmatrix}1\\1\end{pmatrix}$	$\begin{pmatrix}0\\0\\0\end{pmatrix}$	$\begin{pmatrix}1\\1\\1\end{pmatrix}$
2	$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{pmatrix}1\\0\end{pmatrix}$	$\begin{pmatrix}0\\0\\0\end{pmatrix}$	$\begin{pmatrix}1\\1\\0\end{pmatrix}$
3	$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{pmatrix}2\\2\end{pmatrix}$	$\begin{pmatrix}0\\0\\0\end{pmatrix}$	$\begin{pmatrix}1\\0\\1\end{pmatrix}$
4	$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{pmatrix}2\\0\end{pmatrix}$	$\begin{pmatrix}0\\0\\0\end{pmatrix}$	$\begin{pmatrix}1\\0\\0\end{pmatrix}$

DOI: https://doi.org/10.24036/ujsds/vol1-iss4/73

ISSN(*Online*): 2985-475X

Pengaturan	Bivariat		Multivariat	
	$\mu^{(1)}$	$\mu^{(2)}$	$\mu^{(1)}$	$\mu^{(2)}$
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	-	-	$\begin{pmatrix}0\\0\\0\end{pmatrix}$	$\begin{pmatrix}1\\1\\1\end{pmatrix}$

Pada Tabel 2 di atas untuk kasus bivariat pengaturan 1, kedua variabel independennya memuat informasi tentang perbedaan kelas sehingga variabel independen pada pengaturan 1 dinamakan sebagai variabel relevan. Sementara itu, pada pengaturan 2 kasus bivariat terdapat dua kondisi variabel yang berbeda yaitu variabel pertama merupakan variabel yang memuat informasi tentang perbedaan kelas sehingga variabel pertama ini dinamakan variabel relevan sedangkan variabel kedua merupakan variabel dengan kondisi tidak memuat informasi tentang perbedaan kelas sehingga variabel ini disebut sebagai variabel irrelevan. Dalam hal ini, kedua variabel pada pengaturan 2 masing-masing berperan sebagai variabel relevan dan variabel irrelevan. Hal ini juga berlaku untuk kasus data multivariat yang juga memuat variabel relevan dan variabel irrelevan dengan 10 pengaturan perbedaan rataan populasi.

Selain itu, perlakuan korelasi juga diterapkan pada jenis data bivariat dan multivariat yang ditampilkan pada Tabel 3 berikut.

Tabel 3. Pengaturan Korelasi untuk Data Bivariat

Pengaturan	Stuktur Korelasi
1	$\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}$
2	$\begin{bmatrix}1 & 0{,}5\\0{,}5 & 1\end{bmatrix}$
3	$\begin{bmatrix}1 & 0{,}9\\0{,}9 & 1\end{bmatrix}$

Tabel 3 merupakan pengaturan korelasi yang diaplikasikan pada kasus bivariat. Pengaturan 1 merupakan kasus tanpa korelasi, pengaturan 2 merupakan kasus dengan korelasi sedang, dan pengaturan 3 merupakan kasus dengan korelasi tinggi. Pada penelitian ini diteliti kasus tanpa korelasi, korelasi sedang, dan korelasi tinggi dengan kondisi korelasi antara sesama variabel relevan dan antara variabel relevan dengan variabel irrelevan yang menggabungkan pengaturan pada struktur rataan dengan pengaturan pada struktur korelasi. Simulasi ini juga diterapkan pada kasus data multivariat dengan pengaturan korelasi yaitu sebagai berikut.

Tabel 4. Pengaturan Korelasi untuk Kasus Data Multivariat

Pengaturan	Stuktur Korelasi
1	$\begin{bmatrix}1 & 0 & 0\\0 & 1 & 0\\0 & 0 & 1\end{bmatrix}$
2	$\begin{bmatrix}1 & 0{,}5 & 0\\0{,}5 & 1 & 0\\0 & 0 & 1\end{bmatrix}$

DOI: https://doi.org/10.24036/ujsds/vol1-iss4/73

ISSN(*Online*): 2985-475X

Pengaturan	Stuktur Korelasi
3	$\begin{bmatrix} 1 & 0,9 & 0 \\ 0,9 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$
4	$\begin{bmatrix} 1 & 0,5 & 0,5 \\ 0,5 & 1 & 0,5 \\ 0,5 & 0,5 & 1 \end{bmatrix}$
5	$\begin{bmatrix} 1 & 0,9 & 0,9 \\ 0,9 & 1 & 0,9 \\ 0,9 & 0,9 & 1 \end{bmatrix}$

Untuk kasus data multivariat terdapat lima pengaturan struktur korelasi yang disajikan pada Tabel 4. Pengaturan 1 merupakan struktur tanpa korelasi, pengaturan 2 merupakan korelasi sedang untuk dua variabel, pengaturan 3 merupakan korelasi tinggi untuk dua variabel, pengaturan 4 merupakan korelasi sedang untuk tiga variabel, dan pengaturan 5 merupakan korelasi tinggi untuk tiga variabel.

B. Prediksi Laju Galat

Setelah dilakukan pembangkitan data maka tahap selanjutnya yaitu membagi data menjadi data *training* dan data *testing*. Metode ini juga dikenal dengan metode prediksi laju galat yang terbagi menjadi tiga macam yaitu metode LOOCV, HO, dan *K-Fold Cross Validation*. Metode prediksi laju galat merupakan salah satu cara yang dilakukan untuk melakukan validasi pada model yang telah terbentuk nantinya. Metode ini bekerja dengan memprediksi galat yang terdapat pada model. Galat merupakan selisih data sebenarnya dengan data prediksi sedangkan prediksi laju galat yaitu kesalahan rata-rata yang dihasilkan dari penggunaan metode CART untuk mengetahui pengaruh respon pada pengamatan baru. Prediksi galat pada klasifikasi dengan variabel dependen bersifat kategorik maka digunakan rumus sebagai berikut.

$$Err_i = I(y_i \neq \hat{y}_i)$$

Dimana,

$$I = \begin{cases} 1 & y_i \neq \hat{y}_i \\ 0 & y_i = \hat{y}_i \end{cases}$$

y_i merupakan data aktual amatan *ke-i*, dimana *i=1,2,...,n* sedangkan $\hat{y}_i$ merupakan data hasil prediksi *ke-i*, dimana *i=1,2,...,n*. Hasil prediksi $\hat{y}_i$ pada pengamatan ke-*I,* dengan ketentuan indikator variabelnya yaitu $I(y_i \neq \hat{y}_i)$ bernilai 1 dan $(y_i = \hat{y}_i)$ bernilai 0. Jika $I(y_i \neq \hat{y}_i)$ bernilai 0 maka pemangatan ke-i diklasifikasikan dengan benar yang artinya tidak terdapat galat dalam klasifikasi, jika sebaliknya maka terjadi misklasifikasi (James dkk, 2013:184). Nilai prediksi $\hat{y}_i$ didapatkan dari pengujian data *testing* yang dicobakan pada model yang dibentuk menggunakan data *training.*

C. *Leave One Out Cross Validation* (LOOCV)

LOOCV bekerja dengan membagi data menjadi dua bagian yaitu data *testing* dan *data training*. Namun, pada metode LOOCV setiap pengamatan berperan sebagai *data training* dan data *testing*. Menurut James dkk (2013: 178), pengamatan (x_1, y_1) digunakan sebagai data *testing,* sementara itu sisa pengamatannya (x_2, y_2), …, (x_n, y_n) digunakan sebagai *data training.* Pengamatan x_1 pada data *testing* digunakan untuk mencari prediksi y *(ŷ)* untuk mendapatkan nilai *error rate.* Metode ini dilakukan secara berulang kali hingga pengamatan (x_n, y_n) menjadi data *testing*. Berikut merupakan perhitungan menggunakan LOOCV.

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

Dengan $\hat{E}^{LOOCV}$ merupakan prediksi galat LOOCV, *n* merupakan banyak pengamatan dan $I(y_i \neq \hat{y}_i)$ sebagai indeks variabel. Adapun kelebihan dari LOOCV menurut Wong (2015) yaitu sistem beraturan pada LOOCV dan tidak adanya pengacakan untuk setiap pengambilan data testing dan training menyebabkan hasil akurasi rataan yang selalu konstan. Kelebihan lainnya yaitu LOOCV sering kali menghasilkan perkiraan akurasi yang tidak bias untuk akurasi model klasifikasi. Sedangkan kelemahan dari LOOCV menurut Efron (1983) yaitu metode ini memakan terlalu banyak waktu dan biaya jika digunakan pada data berukuran besar dan memiliki hasil perkiraan dengan nilai varians yang sangat besar.

DOI: https://doi.org/10.24036/ujsds/vol1-iss4/73

ISSN(*Online*): 2985-475X

D. *Hold Out*

Estimasi *Hold out* bekerja dengan membagi data secara acak menjadi *data training* dan *validation set*. Rokach dan Maimon (2014: 33) menyebutkan bahwa dua pertiga dari data menjadi data *training* sementara sepertiga sisanya menjadi data *testing.* Model yang disesuaikan pada *data training* kemudian digunakan untuk memprediksi respon pada data pengamatan menggunakan data *testing*. Dalam variabel respon yang bersifat kuantitatif, data *testing* dihitung menggunakan *MSE*. Hasil dari perhitungan *MSE* akan memberikan perkiraan pada *test error*. Metode HO dapat dihitung menggunakan rumus berikut.

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

Dengan $\hat{E}^{HO}$ sebagai prediksi galat *hold out,* $n_{testing}$ merupakan banyak pengamatan pada data *testing,* dan $I(y_i \neq \hat{y}_i)$ sebagai indeks variabel. HO merupakan metode paling sederhana diantara dua metode prediksi galat lainnya. Kelemahan dari metode ini yaitu ketidakefisienan penggunaan data karena sepertiga datanya tidak digunakan untuk membentuk model pada data training.

E. *K-Fold Cross Validation*

Estimasi *K-Fold* merupakan metode lanjutan dari metode LOOCV. *K fold estimasi* mempermudah atau mempersingkat waktu dalam mencari data *testing*. Jika pada LOOCV, setiap pengamatan berperan sebagai data *testing*, sedangkan pada *K fold*, pengamatan dikelompokkan terlebih dahulu menjadi *k* kelompok. Setelah terbentuk *k* kelompok, maka *k-1* menjadi data *training*, sedangkan sisanya menjadi data *testing* hingga *k* kelompok terakhir menjadi data *testing*. Biasanya setiap kelompok terdiri dari *k*=5 atau *k*=10. jadi jika *n* memiliki 200 pengamatan dengan *k*=10 maka dalam satu kelompok terdiri dari 20 pengamatan. Berikut merupakan rumus dari *K-Fold:*

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

Kelebihan dari *K-Fold* yaitu dapat menghemat waktu dan biaya jika digunakan pada data berukuran besar, sedangkan kelemahannya yaitu akurasi hasil rataan galat tidak konstan yang disebabkan oleh adanya pengacakan pengamatan dalam kelompok yang terbentuk pada *K-Fold* (Kohavi, 1995).

F. *Classification and Regression Tree (CART)*

Pembentukan model menggunakan CART merupakan tahap selanjutnya setelah dilakukan pembagian data pada metode prediksi laju galat. CART pertama kali ditemukan oleh Leo Breiman, Jerome Friedman, Richard Olshen, dan Charles Ston pada tahun 1984. Algoritma CART berbentuk seperti struktur pohon yang terdiri dari *root node, internal node, dan terminal node.* Tujuan dari algoritma CART yaitu untuk pengambilan keputusan pada sekumpulan data besar (Han. J dkk, 2012: 330).

Pembentukan pohon CART dimulai dari pemilahan pemilah dengan menggunakan metode binari rekursif menjadi dua cabang dengan memilih variabel prediktor *Xj* dan titik potong *s* sehingga membagi variabel prediktor menjadi sebuah simpul t_1 yang homogen. Hal ini didefiniskan sebagai berikut.

$$t_1(j,s) = \{X | X_j < s\} \; dan \; t_2(j,s) = \{X | X_j \geq s\}$$

Dalam memilih variabel prediktor *Xj* menjadi pemilah yang tepat, digunakan fungsi keheterogenan *index gini* dan nilai impurity. Nilai impurity merupakan tingkat keheterogenan di dalam suatu simpul. Semakin rendah nilai impurity suatu simpul berarti kehomogenan di dalam simpul semakin besar, sedangkan semakin tinggi nilai impurity suatu simpul berarti kehomogenan di dalam simpul semakin kecil. Berikut merupakan rumus dari *index gini* yaitu.

$$i(t) = 1 - \sum_{j=1}^{n} p^2(j|t)$$

dengan

$$p(j|t) = \frac{n_j(t)}{n(t)}$$

$i(t)$ merupakan fungsi keheterogenan indeks gini; $p(j|t)$, proporsi kelas *j* pada simpul *t,* dimana j=1,2,3, ..., n; $p(i|t)$ banyak pengamatan kelas *j* pada simpul *t;* $n(t)$ merupakan banyak pengamatan pada simpul *t.* Sedangkan nilai impuritas dirumuskan sebagai berikut.

$$\Delta i(s,t) = i(t) - P_L i(t_L) - P_R i(t_R)$$

Nilai impurity kelas ke-*s* simpul ke-*t* dilambangkan dengan $\Delta i(s,t)$*;* $i(t)$ *sebagai* fungsi keheterogenan; P_L merupakan peluang observasi pada simpul kiri; $i(t_L)$= nilai impurity simpul ke-*t* kiri; P_R= peluang observasi pada simpul kanan; dan $i(t_R)$ merupakan nilai impurity simpul ke-*t* kanan. Simpul yang memiliki nilai $\Delta i(s,t)$ maksimum akan dipilih sebagai pemilah. Proses ini akan berlanjut hingga mencapai *terminal node.*

DOI: https://doi.org/10.24036/ujsds/vol1-iss4/73

ISSN(*Online*): 2985-475X

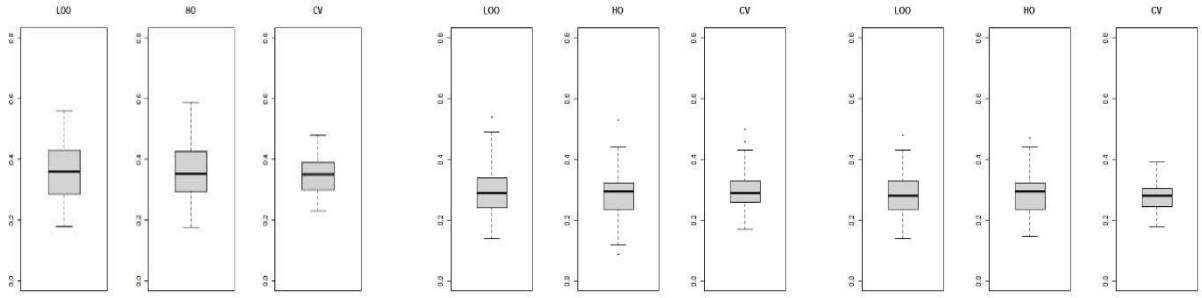
Setelah model pohon CART terbentuk, variabel independen akan diaplikasikan pada model hingga menghasilkan prediksi ke-*i*. Hasil prediksi ini dilambangkan sebagai $\hat{y}_i$. $\hat{y}_i$ ini nantinya akan digunakan untuk menghitung galat yang terdapat model yang telah terbentuk dengan tujuan menghitung akurasi kinerja pada model.

Pada CART variabel yang berkorelasi dapat menyebabkan redudansi dalam informasi yang dimuat pada pohon. Klusowski menyebutkan bahwa korelasi berpengaruh terhadap variabel independen pada setiap node. Korelasi menyebabkan proses splitting yang kurang efektif sehingga berakibat pada berkurangnya kemampuan prediksi pohon serta. Hal ini juga menyebabkan overfitting yang berakibat pada generalisasi yang buruk pada pohon.

III. HASIL DAN PEMBAHASAN

A. Hasil Analisis

Data yang digunakan pada penelitian ini menggunakan pengaturan rataan populasi dan korelasi yang berbeda-beda. Hal ini memberikan pengaruh yang berbeda pula untuk nilai prediksi laju galat yang dihasilkan. Hasil prediksi laju galat divisualisasikan menggunakan boxplot dengan kriteria prediksi galat terbaik dapat disimpulkan dari nilai median prediksi laju galat dan variansnya. Hasil ini bisa didapatkan dengan melakukan perbandingan antara tiga tipe data yang berbeda yaitu univariat, bivariat, dan multivariat. Pada perbandingan ini, perlakuan korelasi tidak diterapkan pada data bivariat dan multivariat. Berikut merupakan boxplot perbandingan untuk metode prediksi galat.



(a) **(b)** **(c)**

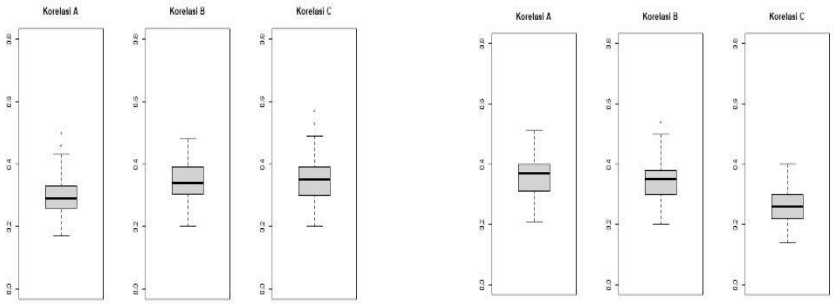
Gambar 1. Prediksi Galat untuk Data (a) Univariat, (b) Bivariat, dan (c) Multivariat

Gambar 1 merupakan hasil prediksi laju galat dari data univariat, bivariat, dan multivariat untuk pengaturan data 1. Dapat dilihat pada Gambar 1, boxplot LOO menunjukkan hasil prediksi laju galat untuk metode LOOCV, boxplot HO menunjukkan hasil untuk metode *Hold Out*, sedangkan boxplot CV menunjukkan hasil prediksi laju galat untuk metode *K-Fold Cross Validation*. Adapun Gambar 1 menunjukkan bahwa boxplot prediksi laju galat LOOCV dan HO cenderung memiliki rentang varian prediksi laju galat yang hampir sama besar meskipun pada beberapa kasus HO menunjukkan rentang varian yang lebih besar daripada LOOCV. Hal ini ditunjukkan pada Gambar 1.a untuk tipe data univariat yaitu boxplot HO memiliki varian prediksi galat yang lebih besar daripada LOOCV. Sementara itu, boxplot *K-Fold* memiliki varian prediksi laju galat yang lebih kecil dibandingkan boxplot yang lain, hal ini terlihat pada semua tipe data yang disajikan pada Gambar 1. Sedangkan untuk nilai median prediksi laju galat, ketiga metode ini memiliki nilai median yang hampir sama besar. Berdasarkan hal ini, dapat disimpulkan bahwa metode *K-Fold Cross Validation* merupakan metode prediksi laju galat yang lebih baik dibandingkan metode LOOCV dan HO untuk metode CART dengan data seimbang.

Nilai prediksi laju galat yang telah diperoleh dipengaruhi oleh beberapa kategori yaitu pengaruh pengaturan perlakuan yang berbeda seperti pengaruh pengaturan rataan dengan korelasi. Berikut ini hasil perbandingan korelasi terhadap laju galat untuk pengaturan data pada data bivariat.

DOI: https://doi.org/10.24036/ujsds/vol1-iss4/73

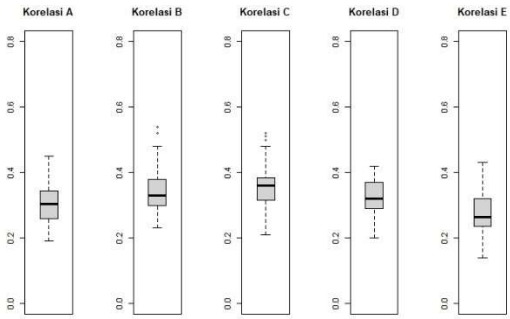
ISSN(*Online*): 2985-475X



(a) **(b)**

Gambar 2. Perbandingan Boxplot *K-Fold* berdasarkan Korelasi pada (a) Pengaturan Data 1 antara sesama variabel relevan (b) Pengaturan Data 2 antara variabel relevan dengan variabel irrelevan

Gambar 2 merupakan perbandingan korelasi pada metode *K-Fold*. Perbandingan serupa juga terdapat pada metode LOOCV dan HO namun pada artikel ini hanya ditampilkan hasil untuk metode *K-Fold* saja. Korelasi A pada Gambar 2 merupakan kondisi tanpa korelasi, korelasi B yaitu korelasi sedang, dan korelasi C merupakan korelasi tinggi. Gambar 2.a menunjukkan nilai prediksi laju galat terlihat semakin meningkat ketika nilai korelasinya semakin tinggi sedangkan pada Gambar 2.b nilai prediksi laju galat semakin menurun ketika nilai korelasi semakin tinggi. Adapun perbedaan ini terjadi karena pada Gambar 2.a variabel yang berkorelasi merupakan antara sesama variabel yang relevan sementara itu Gambar 2.b menunjukkan hasil prediksi laju galat pada variabel relevan dan variabel irrelevan yang saling berkorelasi. Hal ini menyebabkan ketika data memiliki variabel yang sama-sama relevan dan berkorelasi semakin tinggi maka nilai median prediksi laju galat akan semakin meningkat. Sebaliknya ketika data berkorelasi semakin tinggi antara variabel relevan dengan variabel irrelevan maka nilai median prediksi laju galat akan semakin kecil. Hasil yang sama juga terdapat pada data multivariat dengan nilai rataan pengaturan 1 dan pengaturan 6 pada Tabel 6 yaitu semua variabel pada pengaturan ini merupakan variabel relevan dan data multivariat dengan variasi rataan dua variabel yang relevan sedangkan pada variabelnya yang ketiga merupakan variabel irrelevan.



Gambar 3. Perbandingan Boxplot Metode Prediksi Laju Galat *K-Fold* berdasarkan Korelasi pada Data Multivariat untuk Pengaturan 2

Pada Gambar 3, dua variabel x_1 dan x_2 merupakan variabel relevan sedangkan variabel x_3 merupakan variabel irrelevan. Untuk kasus data ini diberikan lima macam perlakuan korelasi yang berbeda yaitu korelasi A menunjukkan perlakuan variabel tanpa korelasi, korelasi B merupakan korelasi sedang dua variabel, korelasi C yaitu korelasi tinggi tiga variabel, korelasi D korelasi sedang tiga variabel, dan korelasi E merupakan korelasi tinggi tiga variabel. Hasil yang didapatkan yaitu nilai prediksi galat akan semakin besar ketika dua variabel yang merupakan variabel relevan berkorelasi semakin tinggi. Hal ini ditunjukkan pada nilai rataan prediksi galat yang semakin tinggi untuk boxplot korelasi A, korelasi B, dan korelasi C. Namun ketika variabel relevan dan irrelevan berkorelasi semakin tinggi maka nilai prediksi akan semakin kecil. Hal ini ditampilkan pada Gambar 3 untuk boxplot korelasi D dan korelasi E yang nilai rataan galatnya semakin rendah.

DOI: https://doi.org/10.24036/ujsds/vol1-iss4/73

ISSN(*Online*): 2985-475X

B. Pembahasan

Hasil pada penelitian ini menunjukkan bahwa *K-Fold* memiliki kinerja yang baik dalam memvalidasi model untuk algoritma CART. Metode *K-Fold*, LOOCV, dan HO menunjukkan hasil yang cenderung hampir sama pada nilai median laju galat. Sementara itu, aspek varian menunjukkan bahwa HO memiliki varian yang paling besar. Hal ini disebabkan oleh kinerja HO yang hanya menggunakan sebagian datanya pada data training untuk model seperti yang disebutkan oleh Maimon (2014:33). Oleh karena itu hal ini berpengaruh pada nilai prediksi laju galat yang dihasilkan. Adapun, *K-Fold* memiliki varian paling kecil diantara ketiga metode prediksi laju galat tersebut, sehingga dapat disimpulkan bahwa metode prediksi laju galat yang paling baik digunakan untuk algoritma CART yaitu metode prediksi laju galat *K-Fold Cross Validation*. Hasil ini sejalan dengan hasil penelitian yang dilakukan oleh Kohavi (1995) dan Payam dkk (2016) pada penelitian terdahulu.

Penelitian ini menggunakan data simulasi dengan tujuan untuk membandingkan metode prediksi galat pada algoritma CART. Aspek yang digunakan pada data simulasi terletak pada perbedaan rataan populasi, jumlah variabel independen dan korelasi dengan proporsi kelas data seimbang. Variasi rataan terdiri dari variabel relevan dan variabel irrelevan yang terdapat pada tiga jenis data yaitu data univariat, bivariat, dan multivariat sedangkan pengaruh korelasi diaplikasikan pada data bivariat dan multivariat.

Pengaturan pada data simulasi memberikan pengaruh terhadap nilai prediksi laju galat. Nilai prediksi laju galat akan semakin besar ketika nilai korelasi juga semakin besar. Kondisi ini terjadi ketika sesama variabel relevan berkorelasi sebaliknya jika variabel relevan dan variabel irrelevan berkorelasi maka nilai prediksi laju galat akan semakin kecil ketika nilai korelasi semakin besar.

IV. KESIMPULAN

Berdasarkan pengujian tiga metode prediksi galat untuk algoritma CART dengan tiga jenis data yaitu univariat, bivariat, dan multivariat didapatkan bahwa metode prediksi laju galat yang memiliki nilai yang paling baik yaitu metode *K-Fold Cross validation*. Selain itu aturan perbedaan nilai korelasi pada data memberikan dampak terhadap hasil nilai prediksi laju galat. Ketika nilai korelasi semakin besar antara sesama variabel relevan maka nilai median prediksi laju galat akan semakin besar namun sebaliknya ketika variabel relevan dan variabel irrelevan saling berkorelasi dengan nilai korelasi yang semakin besar maka nilai median prediksi laju galat akan semakin kecil. Pada penelitian selanjutnya diharapkan dapat mengembangkan penelitian yang baru sehingga dapat menghasilkan hasil yang lebih baik.

DAFTAR PUSTAKA

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J. (1984). *Classification And Regression Trees*. New York: Chapman & Hall.

Efron, B. (1983). Estimating the error rate of a prediction rule: improvement on *cross validation*. *Journal of the American Statistical Association*, 78:382, pp:316-331, Doi: http://dx.doi.org/10.1080/01621459.1983.10477973.

Han, J., Kamber, M., & Pei, J. (2012). *Data mining Concepts and Technique. Third Edition.* Massachusetts: Elsevier, Inc.

Hastie, Trevor., Tibshirani, Robert., & Friedman, Jerome. (2008). *The Elements of Statistical Learning, 2nd Editions*. New York: Springer.

James, Gareth., Witten, Daniela., Hastie, Trevor., & Tibshirani, Robert. (2013). *An Introduction to Statistical Learning with Applications in R*. New York: Springer Science Business Media.

Klusowski, Jason M. (2020). Sparse Learning with CART. *34th Conferenceon Neural Information Processing Systems.*

Kohavi, Ron. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the 14th International Joint Conference on Articial Intelligence*, Volume 2, IJCAI'95, Morgan Kaufmann Publishers, pp. 1137-1143.

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

DOI: https://doi.org/10.24036/ujsds/vol1-iss4/73

ISSN(*Online*): 2985-475X

Refaeilzadeh, Payam., Tang, Lei., & Liu, Huan. (2016). *Cross validation*. *Encyclopedia of Database Systems*. New York: Springer Science Business Media.

Rokach, Lior & Maimon, Oded. (2014). *Data Mining with Decision Trees, Theory and Applications, 2nd edition*. Singapore: World Publishing Co. Pte. Ltd.

T-T.Wong. (2015). Performance evaluation of classification algorithms by k-fold and leave-one-out *cross validation*. Doi: http://dx.doi.org/10.1016/j.patcog.2015.03.009

DOI: https://doi.org/10.24036/ujsds/vol1-iss4/73