UNP JOURNAL OF STATISTICS AND DATA SCIENCE
Volume 1 Nomor 5 Tahun 2023 Halaman 464-472.
ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

Comparison of Error Rate Prediction in CART for Imbalanced Data

Lifia Zullani, Dodi Vionanda*, Syafriandi, dan Dina Fitria
Departemen Statistika, Universitas Negeri Padang, Padang, Indonesia
*Corresponding author: dodi_vionanda@fmipa.unp.ac.id

Submitted : 08 Oktober 2023
Revised : 24 Oktober 2023
Accepted : 26 Oktober 2023

ABSTRACT

CART is one of the tree based classification algorithms. CART is a tree consisting of root nodes, internal nodes, and terminal nodes. The accuracy of the model in CART can be calculated by measuring prediction errors in the model. One common method used to predict error rates is cross-validation. There are three cross-validation algorithms, namely leave one out, hold out, 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. Every algorithm has its shortcomings; hold out cannot guarantee that the training set represents the entire dataset, leave one out is very time-consuming and requires significant computation because it has to train the model as many times as there are data points, and k-fold provides longer computation time because the training algorithm must be run k times. In reality, the data often encountered is imbalanced. Imbalanced data refers to data with a different number of observations in each class. In CART, imbalanced data affects the prediction results. The CART model tends to produce decision trees that are more inclined to predict the majority class because there are more instances of the majority class. This can reduce the prediction accuracy for the minority class. Therefore, the selection of an appropriate cross-validation method is crucial to ensure that cross-validation maintains the correct class proportions during model evaluation. This research focuses on comparing error rate prediction methods in the CART model with imbalanced data. The study uses three types of data: univariate, bivariate, and multivariate, obtained from differences in population means and correlations between independent variables. The results obtained indicate that the k-fold algorithm is the most suitable error rate prediction algorithm applied to CART with imbalanced data.

Keywords: *CART, cross validation, error rate predicton, imbalanced data.*

I. PENDAHULUAN

Classification and Regression Tree (CART) dikembangkan oleh Leo Breiman, Jerome H. Friedman, Richard A. Olshen dan Charles J. Stone sekitar tahun 1980-an (Breiman et al., 1993). CART merupakan metodologi statistik nonparametrik yang dikembangkan untuk topik analisis klasifikasi, baik untuk variabel terikat kategorik maupun kontinu. CART menghasilkan suatu pohon klasifikasi jika variabel terikatnyanya kategorik, dan menghasilkan pohon regresi jika variabel terikatnyanya kontinu

Akurasi model yang telah dibuat dengan metode CART dapat dihitung dengan menggunakan metode trediksi galat. Untuk menghitung prediksi laju galat terdapat dua metode yang dapat digunakan yaitu *training error rate* dan *test error rate*. Pada *training error rate* data set digunakan untuk membangun pohon dan kemudian digunakan kembali untuk melakukan prediksi. Penggunaan metode ini menyebabkan adanya bias pada estimasi dan *overfitting*. Untuk menghindari terjadinya *overfitting* dapat digunakan metode *test error rate*, dimana sebagian data digunakan untuk membangun pohon dan sisanya digunakan untuk memprediksi. Data yang digunakan untuk prediksi dipakai untuk menghitung prediksi laju galat. Dibandingkan metode *training error rate*, metode *test error rate* memberikan hasil yang lebih akurat sehingga Breiman dkk (1984) menamai metode ini sebagai prediksi laju galat klasifikasi yang sebenarnya.

Metode pembagian data pada *test error rate* dilakukan dengan menggunakan *cross validation*. *Cross validatian* merupakan metode pembagian data menjadi data latih dan data uji. Dalam *cross validation* terdapat berbagai algoritma yang dapat digunakan dimana masing-masing algoritma memiliki karakteristik yang berbeda, yaitu *hold out*, *leave one*

DOI: https://doi.org/10.24036/ujsds/vol1-iss5/117

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

out dan *k-fold cross validation*. Perbedaan dari ketiga metode ini terletak pada pembagian data untuk data latih dan data uji. Algoritma *leave one out* bekerja dengan setiap pengamatan berperan sebagai data latih dan data uji, *hold out* membagi data secara acak menjadi dua kelompok sebagai data data uji dan data latih, sementara itu *k-fold cross validation* mengelompokkan data secara acak menjadi *k* kelompok terlebih dahulu kemudian 1 kelompok dijadikan sebagai data uji dan kelompok lainnya menjadi data latih. Perbedaan kinerja ketiga metode prediksi laju galat ini menyebabkan adanya kekurangan dan kelebihan pada masing-masing metode dalam memprediksi laju galat pada model sehingga dilakukan perbandingan untuk menentukan metode prediksi laju galat terbaik pada metode CART.

Sebagian besar data riil yang dihasilkan dalam penelitian merupakan data yang tidak seimbang. Data tidak seimbang adalah data yang memiliki jumlah kelas amatan yang berbeda. Ketidakseimbangan data berdampak pada hasil prediksi yang tidak stabil. Model CART memiliki kecenderungan untuk menghasilkan pohon keputusan yang lebih cenderung memprediksi kelas mayoritas karena ada lebih banyak contoh kelas mayoritas (Sari, 2019). Hal ini dapat mengurangi akurasi prediksi untuk kelas minoritas. sehingga pemilihan metode prediksi laju galat yang sesuai sangat penting untuk memastikan bahwa metode prediksi laju galat dapat mempertahankan proporsi kelas yang benar selama evaluasi model. Permasalahan yang diajukan dalam penelitian ini adalah kinerja metode prediksi LOOCV, *hold out* dan *k-fold cross validation* untuk data seimbang dengan rasio kelas data yang berbeda dan algoritma prediksi laju galat mana yang cocok pada metode CART untuk data tidak seimbang. Tujuan dari penelitian ini adalah untuk mengetahui kinerja masing-masing algoritma prediksi laju galat pada data yang tidak seimbang dengan rasio kelas data yang berbeda, membandingkan kinerja algoritma prediksi laju galat untuk memberikan algoritma prediksi laju galat terbaik untuk diterapkan pada model CART dengan data yang tidak seimbang.

II. METODE PENELITIAN

A. *Classification and Regression Tree* (CART)

CART merupakan salah satu metode atau algoritma dari salah satu teknik eksplorasi data yaitu teknik pohon keputusan. CART terbilang sederhana namun merupakan metode yang kuat. CART bertujuan untuk mendapatkan suatu kelompok data yang akurat sebagai penciri dari suatu pengklasifikasian, selain itu CART digunakan untuk menggambarkan hubungan antara variable terikat dengan satu atau lebih variabel bebas (Breiman et al., 1993).
Langkah-langkah penerapan Algoritma CART adalah sebagai berikut :

1. Pembentukan Pohon Klasifikasi
   Proses pembentukan pohon klasifikasi terdiri atas 3 tahapan, yaitu:
   a. Pemilihan Pemilah
      Pada langkah ini, menggunakan data pelatihan yang kemudian dibagi berdasarkan kriteria pemisahan *goodness of split*. Bagian-bagian yang dihasilkan dari proses pemisahan ini harus memiliki tingkat keseragaman yang lebih tinggi dibandingkan dengan pemisahan sebelumnya. Untuk mengukur tingkat keragaman, menggunakan *indeks Gini*, yang akan menghasilkan pohon keputusan yang memiliki dua cabang. Fungsi indeks Gini adalah sebagai berikut:

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

      Dengan asumsi bahwa P(j|t) merupakan proporsi kelas *j* pada simpul *t*, dimana j=1,2,3,...,n dan P_L, P_R adalah probabilitas dari simpul kanan dan kiri. Atribut yang terpilih akan membentuk sebuah himpunan kelas yang disebut simpul atau *node*. Langkah selanjutnya adalah menentukan kriteria *goodness of split* yang merupakan suatu evaluasi pemilihan oleh pemilah *s* pada *t* yang disebut juga sebagai penurunan keheterogenan dengan rumus sebagai berikut:

$$\phi(s,t) = i(t) - P_L i(t_L) - P_R i(t_R) \quad (2)$$

      Pemilah yang menghasilkan nilai lebih tinggi merupakan pemilah yang lebih baik karena hal ini memungkinkan untuk mereduksi keheterogenan secara lebih signifikan.

   b. Penentuan Simpul Terminal
      Suatu simpul *t* akan menjadi simpul terminal atau tidak, akan dipilah kembali bila pada simpul *t* tidak terdapat penurunan keheterogenan dengan adanya batasan minimum *n* seperti hanya terdapat satu pengamatan pada tiap simpul anak.

   c. Penandaan Label Kelas

DOI: https://doi.org/10.24036/ujsds/vol1-iss5/117

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

Penandaan kelas dilakukan pada simpul terminal, simpul nonterminal, dan akar simpul, akan tetapi, penandaan label paling dibutuhkan pada simpul terminal karena simpul ini penting digunakan untuk memprediksi suatu objek pada kelas tertentu yang berada pada simpul terminal ini. Penandaan label kelas pada simpul terminal dilakukan berdasarkan aturan jumlah terbanyak yaitu jika:

$$P(j_0|t) = max\ _j P(j|t) = max_j \frac{N_j(t)}{N(t)} \quad (3)$$

Label kelas simpul terminal t adalah j_0 yang memberi nilai dugaan kesalahan pengklasifikasian simpul t terbesar, proses pembentukan pohon klasifikasi berhenti apabila hanya ada satu pengamatan dalam tiap simpul anak atau adanya batas minimum n, semua pengamatan dalam simpul anak identik, dan adanya batasan jumlah level atau kedalaman pohon maksimal.

2. Pemangkasan Pohon Klasifikasi

Pemangkasan dilakukan dengan jalan memangkas bagian pohon yang kurang penting sehingga didapatkan pohon optimal. Ukuran pemangkasan yang digunakan untuk memperoleh ukuran pohon yang layak adalah *cost complexity minimum*. Sub pohon dari pohon terbesar T_{Max} $(T < T_{Max})$ yang merupakan ukuran *cost complexity* yaitu :

$$R_\alpha(t) = R(T) + \alpha|\tilde{T}| \quad (4)$$

Cost complexity prunning menentukan suatu pohon bagian $T(\alpha)$ yang meminimumkan $R_\alpha(t)$ pada seluruh pohon bagian, atau untuk setiap nilai α , dicari pohon bagian $T(\alpha) < T_{Max}$ yang meminimumkan $R_\alpha(t)$. Jika $R(T)$ digunakan sebagai kriteria penentuan pohon optimal maka akan cenderung pohon terbesar adalah T_1, sebab semakin besar pohon, maka semakin kecil nilai $R(T)$.

Dalam pohon keputusan CART, terdapat nilai aktual yang disimbolkan dengan y, dan nilai prediksi, yang disimbolkan dengan $\hat{y}$. Nilai aktual ini merupakan nilai sebenarnya dari variabel terikat yang ingin diprediksi oleh model. Sedangkan nilai prediksi mengacu pada nilai yang diprediksi oleh model berdasarkan kondisi atau atribut pada simpul pohon keputusan. Untuk melakukan prediksi, data baru atau pengamatan yang akan diklasifikasi (y) dimasukkan ke dalam pohon keputusan. Data ini harus memiliki atribut yang sesuai dengan atribut yang digunakan selama pelatihan model. Model akan memulai prediksi dari akar pohon, dan mengikuti jalur pohon keputusan. Pada setiap simpul dalam pohon, terdapat aturan keputusan yang memeriksa atribut tertentu pada data baru. Pada setiap simpul dalam pohon keputusan, model akan memilih salah satu cabang berdasarkan hasil aturan keputusan. Proses akan terus berlanjut sampai mencapai simpul daun di pohon. Setelah model mengikuti langkah-langkah di pohon keputusan, ia akan sampai pada simpul daun yang berisi hasil prediksi akhir ($\hat{y}$) dan digunakan sebagai prediksi dari model untuk data baru yang dimasukkan.

B. Prediksi laju galat

Prediksi laju galat terhadap model adalah mengukur kinerja model dengan menghitung segala bentuk tingkat kesalahan prediksi pada model. Prediksi laju galat pada pohon klasifikasi dapat dihitung menggunakan *misclassification rate* dengan rumus sebagai berikut:

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

Dimana, y_i merupakan data aktual pengamatan *ke=i,* dan $\hat{y}_i$ merupakan hasil prediksi pada amatan *ke-i*, dengan indikator variabel $I(y_i \neq \hat{y}_i)$ bernilai 1 jika $(y_i \neq \hat{y}_i)$ dan bernilai 0 jika $(y_i = \hat{y}_i)$. Jika $I(y_i \neq \hat{y}_i)$ bernilai 0 maka pengamatan *ke-1* diklasifikasikan dengan benar menggunakan metode klasifikasi, jika bernilai 1 maka terjadi *misclassified.*

Metode paling sederhana dan paling banyak digunakan untuk memperkirakan prediksi laju galat adalah *cross validation* (Hastie, 2008: 241). Menurut James dkk (2013), *Cross validation* merupakan suatu teknik pengujian keefektifan suatu model, yang dibentuk dengan mengambil data dan membaginya menjadi dua bagian, yaitu data latih dan data uji. Data latih digunakan untuk melatih model agar model dapat memahami pola yang ada pada datanya, sedangkan validasi model menggunakan data uji sebagai alat pengujiannya. Ada beberapa algoritma pembelajaran di c*ross validation* :

1. *Hold Out*

Hold out adalah strategi validasi yang secara acak membagi set pengamatan menjadi dua bagian, data latih dan data uji. Model dibentuk menggunakan data latih, dan model yang telah dibentuk digunakan untuk memprediksi respons

DOI: https://doi.org/10.24036/ujsds/vol1-iss5/117

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

pengamatan pada data uji(James, 2013: 176). Prediksi laju galat dengan metode *hold out* dapat dihitung dengan menggunakan rumus sebagai berikut:

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

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

LOOCV melakukan pemisahan set pengamatan menjadi dua bagian. Satu pengamatan digunakan untuk data uji, dan pengamatan yang tersisa digunakan sebagai data latih (James, 2013: 178). Amatan pertama sebagai data uji dan amatan lainnya sebagai data latih yang menghasilkan *error rate* pertama. Prosedur ini di ulang hingga semua amatan menjadi data uji, sehingga menghasilkan *error rate* sebanyak *n* amatan. Prediksi laju galat dengan metode LOOCV dapat dihitung dengan menggunakan rumus sebagai berikut:

$$\hat{\mathrm{E}}^{\mathrm{LOOCV}} = \frac{1}{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{I}\left(\mathrm{y}_{(\mathrm{i})} \neq \hat{\mathrm{y}}_{(\mathrm{i})}\right) \quad (6)$$

3. *K-Fold Cross Validation*

Pendekatan *k-fold cross validation* melibatkan secara acak yang membagi himpunan pengamatan ke dalam *k* grup, atau lipatan, dengan ukuran yang kira-kira sama (Suyanto, 2013) Lipatan pertama digunakan sebagai data uji, dan $k - 1$ lipatan yang tersisa menjadi data latih,yang kemudian digunakan sebagai himpunan data pertama,. Demikian seterusnya hingga didapatkan *k* himpunan data, sehingga setiap lipatan pernah menjadi data uji sebanyak satu kali. Menurut James dkk (2013) Prediksi laju galat dengan metode *k-fold cross validation* dapat dihitung dengan menggunakan rumus sebagai berikut:

$$\hat{E}^{CV} = \frac{1}{k} \sum_{i=1}^{k} \frac{1}{n_k} \sum_{i=1}^{n_{kk}} I(y_1 \neq \hat{y}_1) \quad (7)$$

C. *Imbalanced Data*

Data Imbalance atau data tidak seimbang merupakan kondisi dimana suatu kelompok kelas memiliki jumlah data yang jauh berbeda dibandingkan dengan kelas lainnya (Sari, 2019). Kelas yang memiliki jumlah data lebih banyak disebut dengan *majority class* dan kelas yang mempunyai jumlah data lebih sedikit disebut dengan *minority class.* Karakteristik dari *data imbalance* dapat mempengaruhi hasil prediksi yang dilakukan oleh algoritma. Menurut Vluymans (2019) untuk mengetahui seberapa besar tingkat ketidak seimbangan data yang ada dapat dihitung menggunakan IR *(Imbalanced Ratio)* dengan perbandingan sebagai berikut.

$$Imbalanced\ Ratio = \frac{n_{majority}}{n_{minority}}$$

Perbandingan diatas menunjukkan besarnya tingkat ketidakseimbangan data berdasarkan perbandingan kelas major dan kelas minor. Ketika IR = 1, dataset seimbang sempurna. Nilai yang lebih besar menunjukkan perbedaan yang lebih besar dalam ukuran kelas. Setiap kumpulan data dengan rasio ketidakseimbangan melebihi 1,5 dianggap tidak seimbang, sedangkan IR = 9 sering digunakan sebagai ambang batas di mana kumpulan data dianggap sangat tidak seimbang (Vluymans, 2019:83)

D. Jenis dan Sumber Data

Jenis data yang digunakan pada penelitin ini adalah data primer. Data primer yang digunakan merupakan 3 jenis data acak yaitu univariat, bivariat dan multivariat yang dibangkitkan menggunakan *software R studio* sebanyak 100 amatan. Data acak untuk variabel prediktor berupa data numerik yang berdistribusi normal$\mathrm{N}(\mu, \Sigma)$. Sedangkan untuk variabel respon berupa data kategorik dengan dua kelas yaitu 0 dan 1. Data dibangkitkan dengan berbagai proporsi, rataan dan korelasi antar variabel. Perbedaan nilai rataan yang digunakan yaitu $\mu^{(1)}$ yang merupakan rataan untuk kelas data yang bernilai 0 dan $\mu^{(2)}$ untuk kelas data yang bernilai 1

Pembangkitan data acak berdistribusi normal pada kasus variabel prediktor univariat diterapkan perbedaan nilai rataan populasi dengan nilai variansi yang sama dengan uraian pada Tabel 1.

DOI: https://doi.org/10.24036/ujsds/vol1-iss5/117

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

Tabel 1. Ketentuan untuk Data Univariat.

	$\mu^{(1)}$	$\mu^{(2)}$
Pengaturan 1	0	1
Pengaturan 2	0	2

Pembangkitan data acak berdistribusi normal pada kasus variabel prediktor bivariat dan multivariat diterapkann pengaturan beda rataan populasi yang akan diiringi dengan penerapan struktur korelasi. Berikut adalah uraian mengenai pengaturan beda rataan populasi untuk variabel prediktor bivariat dan multivariat pada Tabel 2.

Tabel 2. Ketentuan Struktur Rataan untuk Data Bivariat

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	$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{pmatrix}2\\0\end{pmatrix}$	-	-

Berdasarkan uraian Tabel 2 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. Untuk pengaturan struktur korelasi pada kasus data bivariat dan multivariat akan dijelaskan pada Tabel 3 berikut.

Tabel 3. Ketentuan Struktur Korelasi Kasus Bivariat

Pengaturan	Struktur korelasi	
	Bivariat	Multivariat
1	$\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}$	$\begin{bmatrix}1 & 0 & 0\\0 & 1 & 0\\0 & 0 & 1\end{bmatrix}$
2	$\begin{bmatrix}1 & 0.5\\0.5 & 1\end{bmatrix}$	$\begin{bmatrix}1 & 0.5 & 0\\0.5 & 1 & 0\\0 & 0 & 1\end{bmatrix}$
3	$\begin{bmatrix}1 & 0.9\\0.9 & 1\end{bmatrix}$	$\begin{bmatrix}1 & 0.9 & 0\\0.9 & 1 & 0\\0 & 0 & 1\end{bmatrix}$
4		$\begin{bmatrix}1 & 0.5 & 0.5\\0.5 & 1 & 0.5\\0.5 & 0.5 & 1\end{bmatrix}$
5		$\begin{bmatrix}1 & 0.9 & 0.9\\0.9 & 1 & 0.9\\0.9 & 0.9 & 1\end{bmatrix}$

DOI: https://doi.org/10.24036/ujsds/vol1-iss5/117

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

Tabel 3 merupakan pengaturan korelasi yang digunakan dimana pengaturan 1 menunjukkan kasus tanpa korelasi. Untuk pengaturan 2 menunjukkan kondisi kasus dengan korelasi sedang sedangkan pengaturan 3 menunjukkan kasus dengan korelasi tinggi. Penelitian ini mengkaji 3 kondisi yaitu tanpa korelasi, korelasi sedang, dan korelasi tinggi untuk kondisi antara dua variabel relevan dan antar variabel relevan dengan variabel irelevan dengan mengkombinasikan penerapan pengaturan beda rataan populasi dan struktur korelasi antar variabel.

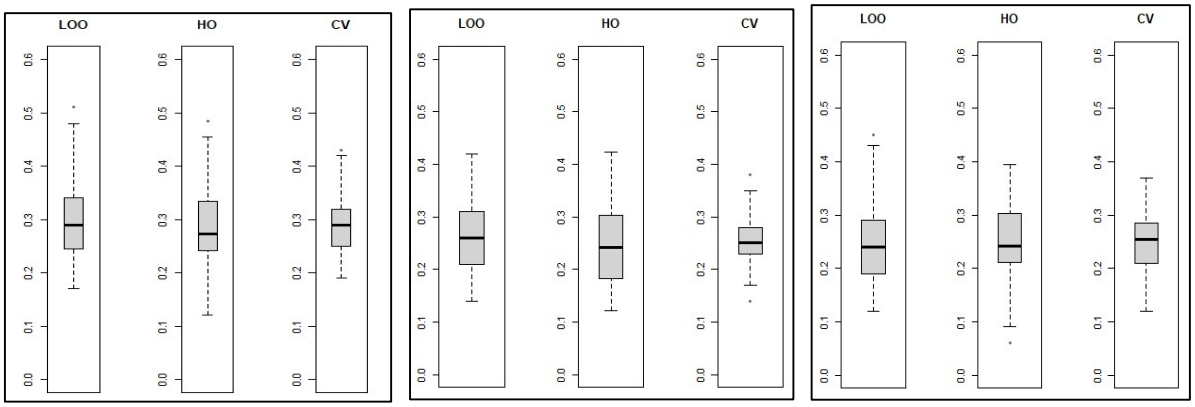
Penelitian ini mengkaji kasus data tidak seimbang maka data akan dibangkitkan dengan berbagai macam proporsi. Berikut merupakan proporsi yang digunakan yang diuraikan pada Tabel 5.

Tabel 4. Rasio Jumlah Sampel dan Kelompok

Proporsi	n_1	n_2
1	50	50
2	60	40
3	70	30
4	80	20
5	90	10

III. HASIL DAN PEMBAHASAN

Tujuan dari penelitian ini adalah untuk menemukan algoritma prediksi laju galat yang cocok digunakan untuk metode klasifikasi CART pada kasus data tidak seimbang. Terdapat tiga algoritma yang dibandingkan yaitu LOOCV, *hold out*, dan *k-fold*. Hasil yang diperoleh dari penelitian ini disajikan dalam bentuk *boxplot*. Algoritma prediksi laju galat yang terbaik dapat dilihat dari algoritma yang memiliki nilai *Inter Quartil Range* (IQR) terendah. Gambar 1 menunjukkan bahwa data multivariat memiliki nilai *error rate* yang paling rendah, disusul oleh data bivariat dan kemudian data univariat, yang artinya semakin banyak variabel independen maka nilai *error rate* semakin kecil.



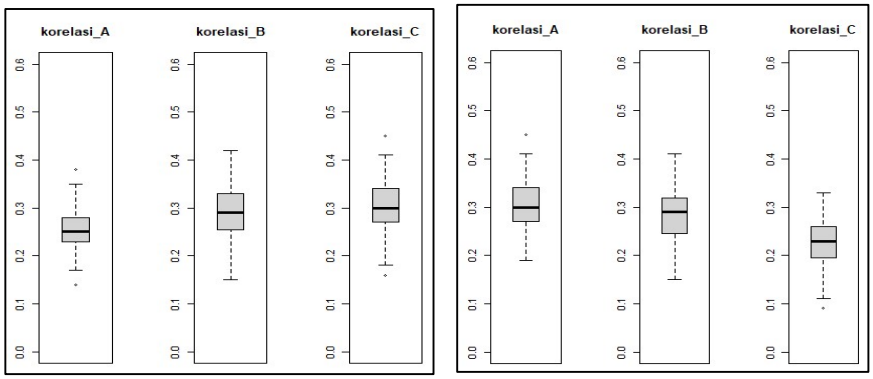
(a) **(b)** **(c)**

Gambar 1. Prediksi laju galat untuk kasus 1 (a) univariat, (b) bivariat, dan (c) multivariat.

Berdasarkan Gambar 1 dapat dilihat bahwa algoritma LOOCV, *hold out* dan *k-fold cross validation* memiliki variasi *error rate* yang sangat berbeda. Variasi *error rate* yang paling besar dihasilkan oleh algoritma *hold out*, sedangkan variasi *error rate* yang paling kecil dihasilkan oleh algoritma *k-fold cross validation*. Dengan demikian, bisa disimpulkan bahwa algoritma yang tepat untuk memproyeksikan tingkat kesalahan dalam pemodelan CART dengan data yang tidak seimbang adalah algoritma *k-fold cross validation*. Hasil yang sama dapat diperoleh dalam pengaturan berikutnya yang menunjukkan bahwa algoritma yang sesuai untuk memprediksi tingkat kesalahan adalah *k-fold cross validation*.

DOI: https://doi.org/10.24036/ujsds/vol1-iss5/117

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

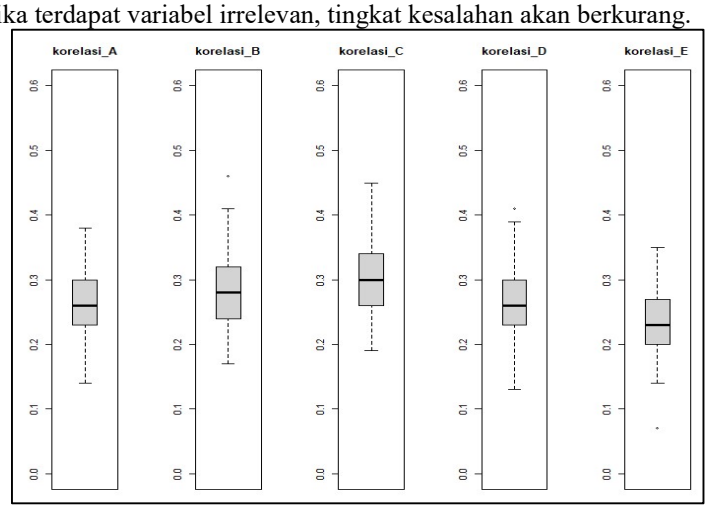


(a) **(b)**

Gambar 2. Prediksi laju galat metode *k-fold cross validation* dengan korelasi berbeda untuk (a) kasus 1 variabel relevan dan (b) kasus 2 variabel irelevan untuk data bivariat.

Data bivariat dan multivariat memiliki perbedaan dalam tingkat prediksi laju galat yang dapat diamati dalam berbagai pengaturan korelasi. Dalam situasi di mana data dengan semua variabelnya adalah variabel relevan saling berkorelasi, yang berarti variabelnya memiliki hubungan yang signifikan satu sama lain, *error rate* akan meningkat seiring dengan peningkatan korelasi, seperti yang terlihat pada Gambar 2(a). Sebaliknya, dalam Gambar 2(b), terdapat dua variabel, dimana yang pertama relevan dan yang kedua tidak relevan. Ketika dua variabel tersebut memiliki korelasi, tingkat kesalahan akan mengalami penurunan seiring dengan peningkatan korelasi antara keduanya. Temuan serupa juga berlaku dalam konteks lainnya.

Data multivariat melibatkan dua variabel bebas yang memiliki korelasi yaitu variabel pertama dengan variabel kedua, dan variabel pertama dengan variabel ketiga. Gambar 3 menunjukkan hasil *error rate* pada variabel yang memiliki korelasi dengan menggunakan algoritma *k-fold cross validation*. Dalam kasus ini, variabel pertama dan kedua merupakan variabel relevan, sementara variabel ketiga merupakan variabel irrelevan. Dari Gambar 3, terlihat bahwa ketika semua variabel yang relevan dalam data memiliki korelasi, *error rate* yang dihasilkan akan meningkat. Namun, jika terdapat variabel yang tidak relevan dalam data, *error rate* yang dihasilkan justru akan berkurang. Hal yang serupa berlaku pada data dengan korelasi antara variabel lainnya. Ketika semua variabel dalam data relevan, tingkat kesalahan akan meningkat, namun jika terdapat variabel irrelevan, tingkat kesalahan akan berkurang.

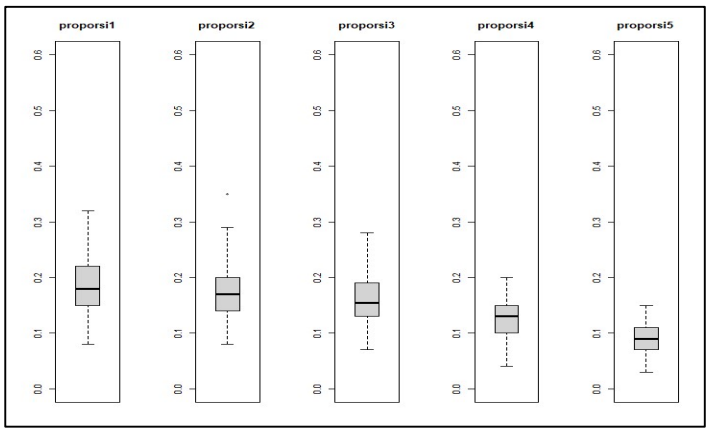


Gambar 3. Hasil *error rate* algoritma *k-fold cross validation* dengan korelasi antara variabel relevan dengan variabel irrelevan pada data multivariat kasus 2

Data yang dibangkitkan dengan proporsi kelas data yang berbeda menghasilkan nilai prediksi yang berbeda seperti dapat dilihat pada Gambar 4.

DOI: https://doi.org/10.24036/ujsds/vol1-iss5/117

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X



Gambar 4. Perbandingan hasil *error rate* dengan jumlah kelas amatan yang berbeda pada algoritma *k-fold cross validation* data univariat kasus 1

Berdasarkan Gambar 4 dapat dilihat bahwa semakin tidak seimbang data maka menghasilkan nilai prediksi yang semakin kecil. Namun hal ini bukan berarti semakin tidak seimbang suatu data maka dikatakan baik tetapi ini merupakan sebuah kesalahan. Pada data tidak seimbang, kesalahan prediksi atau *error* sering terjadi pada kelas minoritas sehingga *error rate* yang dihasilkan cenderung lebih kecil. Seperti yang terdapat pada Gambar 4 ketika perbandingan proporsi kelas data seimbang dengan proporsi1 dimana n = 50:50, hasil *error rate* menyebar di antara selang 0.08 sampai 0.32, tetapi pada data dengan proporsi kelas yang sangat tidak seimbang dengan perbandingan kelas pada proporsi4 dengan n = 80:20 hasil *error rate* berada diantara selang angka 0.04 sampai 0.2 begitu juga dengan perbandingan kelas pada proporsi5 dengan n = 90:10 hasil *error rate* berada diantara selang angka 0.03 sampai 0.13.

IV. KESIMPULAN

Algoritma LOO dan *k-fold cross validation* menunjukkan performa yang hampir sama dalam memprediksi *error rate*, tetapi algoritma *k-fold cross validation* memiliki variasi error rate yang paling stabil. Oleh karena itu, algoritma *k-fold cross validation* lebih disarankan untuk digunakan dalam memprediksi *error rate* pada pemodelan CART dalam konteks data yang tidak seimbang. Korekasi antar variabel dalam data bivariat dan multivariat mempengaruhi hasil prediksi *error rate*. Korelasi antara variabel yang relevan dalam data menghasilkan hasil yang berbeda dibandingkan dengan korelasi antara variabel yang relevan dan yang irrelevan. Saat variabel yang relevan saling berkorelasi, tingkat kesalahan prediksi meningkat seiring dengan peningkatan korelasi, sementara ketika terjadi korelasi antara variabel yang relevan dan irrelevan, *error rate* prediksi cenderung menurun seiring dengan peningkatan korelasi. Ketidakseimbangan dalam data juga mempengaruhi *error rate* yang dihasilkan oleh ketiga algoritma. Semakin tidak seimbang data, *error rate* cenderung semakin rendah. Hal ini terjadi karena pada data yang tidak seimbang, *error rate* akan cenderung mendekati proporsi kelas minoritas tanpa perlunya penanganan khusus. Oleh karena itu, untuk penelitian selanjutnya, disarankan untuk mengembangkan penelitian yang mempertimbangkan metode penanganan ketidakseimbangan data agar dapat menghasilkan *error rate* yang lebih optimal.

DAFTAR PUSTAKA

Arlot, Sylvain. 2010. A Survey of Cross Validation Procedures for Model Selection. Statistic Surveys. Volume 4, pp. 40-79. DOI: 10.1214/09-SS054.

Berrar, Daniel. 2018. Cross-validation. Encyclopedia of Bioinformatics and Computational Biology. Volume 1, Elsevier, pp. 542–545. https://doi.org/10.1016/B978-0-12-809633-8.20349-X.

Breiman, L., Friedman, J.H., Olshen R.A & Stone, C.J. 1984. Classification And Regression Tree. New York: Chapman and Hall.

Chawla, N. V. 2003. C4.5 and Imbalanced Data Sets : Investigating the Effect of Sampling Method, Probabilistic Estimate, and Decision Tree Structure. ICML Whorkshop Learning from Imbalanced Data Sets II. Washington D.C.

DOI: https://doi.org/10.24036/ujsds/vol1-iss5/117

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

Hastie, T., Tibshirani, R. & Friedman, J. 2008. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. California: Springer.

James, G., Witten, D., Hastie, T., & Tibshirani, R., 2013. An Introduction to Statistical Learning. New York: Spinger.

Sari, D. (2019). Klasifikasi Kadae Glukosa Darah Menggunakan Metode Regresi Logistik Ordinal Pada Data Tidak Seimbang [TESIS]. Bogor: Instut Pertanian Bogor.

Vluymans, Sarah. 2019. Dealing with Imbalanced and Weakly Labelled Data in Machine Learning using Fuzzy and Rough Set Methods. Belgium: Springer.

DOI: https://doi.org/10.24036/ujsds/vol1-iss5/117