UNP JOURNAL OF STATISTICS AND DATA SCIENCE
Volume I Nomor 4 Tahun 2023 Halaman 240 – 247.
ISSN(*Online*): 2985-475X

Comparison of Error Rate Prediction Methods of C4.5 Algorithm for Imbalanced Data

Yunistika Ilanda, Dodi Vionanda*, Yenni Kurniawati, Dina Fitria
Departemen Statistika, Universitas Negeri Padang, Padang, Indonesia
*Corresponding author: dodi_vionanda@fmipa.unp.ac.id

Submitted : 12 Juni 2023
Revised : 18 Juli 2023
Accepted : 20 Juli 2023

ABSTRACT

Classification modeling can be formed using the C4.5 algorithm. The model formed by the C4.5 algorithm needs to be seen for its prediction accuracy using the error rate prediction method. Error rate prediction methods that distinguish training data and testing data have better performance. Three error rate prediction methods with training and testing data division that are often used are Hold Out (HO), Leave One Out Cross Validation (LOOCV), and K-Fold Cross Validation (K-Fold CV). This research focuses on the comparison of HO, LOOCV, and K-Fold CV error rate prediction methods in the C4.5 algorithm for imbalanced data cases, because this case is often encountered in real cases of classification. Imbalanced data causes an increase in the classification error of the C4.5 algorithm because the prediction results do not represent the entire data and worsen the performance of the error rate prediction method. Meanwhile, the case of data with different correlations is carried out to find out whether different correlations affect the performance of the error rate prediction method. The purpose of the research is to find out the most suitable error rate prediction method applied to the C4.5 algorithm in the case of imbalanced data and the influence of different correlations. The results show that the K-Fold CV method is the most suitable prediction method applied to the C4.5 algorithm for imbalanced data cases compared to the HO and LOOCV methods. In addition, high correlation can worsen the performance of error rate prediction methods.

Keywords: *C4.5 Algorithm, Error Rate Prediction Methods, Imbalanced Data*

I. PENDAHULUAN

Pemodelan dalam klasifikasi dapat dibentuk menggunakan algoritma C4.5. Algoritma C4.5 bekerja berdasarkan prinsip pohon keputusan yang cukup andal digunakan dan banyak diterapkan. Menurut Ling, dkk. (2003) algoritma C4.5 mampu membentuk struktur pohon keputusan yang berukuran kecil dan relatif akurat. Model pohon keputusan yang dibentuk oleh algoritma C4.5 perlu dilihat akurasi prediksinya. Akurasi merupakan kemampuan model pohon keputusan memprediksi dengan tepat (Mingers, 1989). Akurasi dapat dihitung menggunakan metode prediksi laju galat. Penggunaan metode prediksi laju galat dengan kinerja terbaik pada algoritma C4.5 menghasilkan evaluasi kinerja model yang tepat dan akurat.

Menurut Simon, dkk. (2003) kinerja metode prediksi laju galat paling baik jika dibuat dengan membedakan data *training* dan data *testing*. Metode prediksi laju galat yang bekerja berdasarkan pembagian data *training* dan *testing* serta sering dijumpai penggunaannya yaitu metode prediksi laju galat *Hold Out* (HO), *Leave One Out Cross Validation* (LOOCV), dan *K-Fold Cross Validation (K-Fold CV).* Perbedaan prinsip kerja serta besar pembagian data *training* dan *testing* ketiga metode prediksi laju galat tersebut menjadi dasar membandingkan metode prediksi laju galat yang lebih efektif dan efisien diterapkan pada algoritma C4.5. Pada metode HO, *overfitting* dapat dihindari dan waktu komputasi lebih cepat dibandingkan metode LOOCV ataupun K-Fold CV sebab bekerja tanpa iterasi. Akan tetapi, metode HO tidak menggunakan semua data yang tersedia dan hasilnya sangat bergantung pada pilihan untuk pemisahan data *training* dan *testing*. Metode LOOCV memiliki estimasi akurasi hampir tidak bias, tetapi membutuhkan waktu komputasi yang lama sebab menggunakan iterasi sebanyak n amatan. Sementara itu, metode *K-Fold CV* memanfaatkan data secara maksimal, tetapi tidak optimal digunakan pada dataset berukuran kecil (Refaeilzadeh, dkk., 2009).

Berkenaan dengan dataset yang digunakan dalam penelitian, kinerja metode prediksi laju galat dapat terganggu oleh adanya data tidak seimbang. Menurut Kotsiantis, dkk. (2006) dan Chawla (2009) data tidak seimbang merupakan dataset yang memiliki kelas kategori klasifikasi dengan jumlah amatan yang tidak sama. Data tidak seimbang

DOI: https://doi.org/10.24036/ujsds/vol1-iss4/89

ISSN(*Online*): 2985-475X

menyebabkan peningkatan kesalahan klasifikasi sebab data kategori kelas minoritas akan masuk kedalam data kategori kelas mayoritas sehingga hasil prediksi tidak merepresentasikan data secara keseluruhan.

Penelitian yang telah dilakukan oleh Kohavi (1995) dalam membandingkan metode prediksi laju galat Bootstrap, HO, LOOCV, dan *K-Fold* CV pada pemodelan algoritma C4.5 memperlihatkan metode prediksi laju galat *K-Fold CV* memiliki kinerja prediksi yang paling baik. Penelitian lain dilakukan oleh Molinaro, dkk. (2005) yang membandingkan metode prediksi laju galat *Split sample, K-Fold CV, LOOCV, Monte Carlo cross-validation (MCCV),* dan *Bootstrap* pada algoritma LDA, DDA, NN, dan CART juga menunjukkan bahwa metode prediksi laju galat *K-Fold CV* menghasilkan kinerja prediksi yang sangat baik. Kedua penelitian tersebut dilakukan menggunakan data seimbang serta tidak melihat pengaruh beda korelasi antar variabel prediktor yang digunakan. Sementara itu, dalam klasifikasi banyak kasus nyata dijumpai data tidak seimbang dengan berbagai tingkat ketidakseimbangan data. Algoritma C4.5 juga sensitif terhadap kekuatan korelasi antar variabel penjelas sebab dapat mempengaruhi perolehan Rasio Gain. Oleh karena itu, dilakukan penelitian perbandingan kinerja metode prediksi laju galat HO, LOOCV, dan *K-Fold CV* diberbagai pengaturan data simulasi dengan harapan dapat bermanfaat untuk pertimbangan penggunaan metode prediksi laju galat algoritma C4.5 yang tepat sesuai kondisi data nyata yang ditemui nantinya. Tujuan penelitian untuk mengetahui metode prediksi laju galat yang cocok diterapkan pada algoritma C4.5 kasus data tidak seimbang serta pengaruh beda korelasi yang diberikan.

II. METODE PENELITIAN

A. Algoritma C4.5

Algoritma C4.5 membentuk model pohon keputusan berdasarkan kriteria Entropy dan informasi Rasio Gain (Prasetyo 2014: 60). Model pohon keputusan dibangun menggunakan data *training* yang telah disiapkan. Hal utama yang perlu dilakukan adalah memilih simpul akar (*Root*). Menurut Quinlan (1993: 20-25) langkah penentuan simpul akar dimulai dengan menghitung nilai *Entropy* total atau *Entropy(S)* dengan rumus berikut.

$$Entropy(S) = -\sum_{j=1}^{k} \frac{freq(C_j,S)}{|S|} \times Log_2\left(\frac{freq(C_j,S)}{|S|}\right)$$

dimana.

S = Himpunan kasus.
|S| = Jumlah kasus dalam himpunan S.
$freq(C_j, S)$ = Jumlah kasus S dalam kelas C_j, dimana C_j satu-satunya kelas di S.
k = Jumlah kelas data.

Jika variabel X bernilai numerik, maka dilakukan uji biner dengan hasil $x \leq z$ dan $x > z$. Nilai z merupakan ambang batas (*thresholds*) terbaik dengan pemilihan sebagai berikut.

1. Data pada prediktor X diurutkan dari yang terkecil hingga terbesar.
2. Melakukan *cut-off* dengan rumus $c_i = \frac{x_i + x_{(i+1)}}{2}$, $untuk\ i = 1,2,\cdots,n-1$.
3. Masing-masing nilai *cut-off* dikategorikan ($\leq$ dan $>$) dan dilakukan perhitungan nilai Rasio Gain.
4. Nilai *cut-off* dengan kriteria Rasio *Gain* tertinggi merupakan nilai terbaik untuk dipilih sebagai nilai z.

Selanjutnya melakukan perhitungan nilai *Entropy* masing-masing kategori (S_i), dilanjutkan dengan menghitung nilai Entropy variabel X untuk n adalah jumlah kategori variabel X, dan menghitung nilai *Gain* dengan rumus sebagai berikut.

$$Entropy(S_i) = -\sum_{j=1}^{k} \frac{freq(C_j,S_i)}{|S_i|} \times Log_2\left(\frac{freq(C_j,S_i)}{|S_i|}\right)$$

$$Entropy_X(S) = \sum_{i=1}^{n} \frac{|S_i|}{S} \times Entropy(S_i)$$

$$Gain(X) = Entropy(S) - Entropy_X(S)$$

Sementara itu, dilakukan pula perhitungan Split Entropy variabel prediktor sebagai berikut.

$$split\ Entropy(X) = -\sum_{i=1}^{n} \frac{|S_i|}{S} \times log_2\left(\frac{|S_i|}{S}\right)$$

Nilai *Gain* dan nilai *Split Entropy* variabel prediktor yang telah didapatkan dapat digunakan untuk menghitung *Rasio Gain* dengan rumus berikut.

$$Rasio\ gain(X) = \frac{Gain(X)}{Split\ Entropy(X)}$$

Pemilihan simpul akar didasarkan oleh perolehan nilai *Rasio Gain*, fitur yang memiliki nilai Rasio Gain tertinggi dipilih sebagai simpul akar dengan hasil $x \leq z$ dan $x > z$ sebagai lengan *splitting*. Langkah dilakukan secara rekursif untuk setiap cabang atau simpul internal selanjutnya. Pemisahan berhenti setelah seluruh amatan berada dalam *thresholds* atau semua data dalam setiap simpul hanya memberikan satu label kelas yang tidak dapat dipecah (simpul

DOI: https://doi.org/10.24036/ujsds/vol1-iss4/89

daun) yang berisi keputusan (Rokach & Maimon, 2005). Pada simpul daun itulah terdapat $\hat{y}_i$ yaitu hasil prediksi setiap amatan variabel prediktor x_i masuk pada variabel respon kategori 0 atau 1.

Pada algoritma C4.5, korelasi mempengaruhi besarnya perolehan nilai Rasio Gain. Semakin kuat korelasi antar variabel prediktor maka semakin besar nilai Rasio Gain yang didapatkan. Nilai Rasio Gain yang tinggi dapat meningkatkan potensi klasifikasi algoritma C4.5 sebab pemilihan fitur simpul yang membentuk pohon keputusan didasarkan pada nilai Rasio Gain tertinggi dari fitur yang tersedia, dan besar kemungkinan fitur yang telah terpilih sebelumnya dapat terpilih kembali dengan *splitting* yang berbeda. Hal ini dapat menurunkan akurasi prediksi algoritma C4.5 (Zheng, dkk., 2021).

B. Prediksi Laju Galat

Prediksi laju galat merupakan usaha yang dilakukan untuk mengestimasi besarnya kesalahan dalam memprediksi. Menurut Efron & Tibsirani (1993: 237) laju galat didefinisikan sebagai probabilitas klasifikasi yang salah atau *Misclassification Rate.* James, dkk. (2013: 184) mendeskripsikan laju galat tersebut sebagai berikut.

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

dimana.

$\hat{E}$ = Prediksi laju galat
y_i = Data aktual atau sebenarnya amatan ke- i untuk i = 1,2,…, n
$\hat{y}_i$ = Data hasil prediksi amatan ke-i untuk i = 1,2,…, n
n = Jumlah amatan
Err_i = *Error* atau galat

Kriteria galat pada klasifikasi dapat dilihat dengan fungsi indikator seperti berikut.

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

dengan syarat :

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

dimana, jika $I(y_i \neq \hat{y}_i) = 1$, artinya terjadi kesalahan klasifikasi pada amatan ke-i. Jika $I(y_i \neq \hat{y}_i) = 0$, artinya amatan ke-i terklasifikasi dengan benar (James, dkk., 2013: 268).

C. Metode Prediksi Laju Galat *Hold Out*

Metode kerja HO yaitu membagi data set kedalam 2 bagian yaitu data *training* dan data *testing* secara acak (Refaeilzadeh, dkk., 2009). Menurut Kohavi (1995), pembagian jumlah antara data *training* dan data *testing* yaitu 2/3 data *training* dan 1/3 data *testing*. Prediksi laju galat pada metode *HO* dihitung menggunakan Persamaan 1 berikut.

$$\hat{E}^{HO} = \frac{1}{n_{test}}\sum_{i=1}^{n_{test}} Err_i \quad (1)$$

dimana.

$\hat{E}^{HO}$ = Prediksi laju galat metode *Hold Out.*
n_{test} = Jumlah amatan pada data *testing.*

Sifat-sifat metode prediksi laju galat HO adalah *independent training and test* yang tidak memakan waktu lama karena tidak ada proses iterasi serta dapat mengatasi *overfitting*. Akan tetapi, hasil prediksi sangat bergantung pada pilihan untuk pemisahan antara data *training* dan data *testing*. Semakin besar data *training* yang digunakan, maka hasil akurasi semakin menunjukkan hasil yang tinggi (Refaeilzadeh, dkk., 2009).

D. Metode Prediksi Laju Galat *Leave One Out Cross Validation*

Menurut James, dkk. (2013: 178-179) metode LOOCV hanya menggunakan 1 amatan sebagai data *testing* dan n-1 amatan sisanya sebagai data *training*. Pada metode LOOCV dilakukan proses perhitungan sebanyak n iterasi yang tidak menggunakan sistim pengacakan dalam membagi datanya, sebab dalam setiap iterasi hanya menyisakan 1 amatan sebagai data *testing.* Prediksi laju galat metode LOOCV dihitung menggunakan Persamaan 2 berikut.

$$\hat{E}^{LOO} = \frac{1}{n}\sum_{i=1}^{n} Err_i \quad (2)$$

dimana.

$\hat{E}^{LOO}$ = Prediksi laju galat metode *LOOCV*
n = Jumlah data

DOI: https://doi.org/10.24036/ujsds/vol1-iss4/89

Sifat-sifat metode LOOCV menurut Refaeilzadeh, dkk. (2009) dan James, dkk. (2013: 179) yaitu, memiliki estimasi akurasi yang hampir tidak bias dan n-1 amatan sebagai data *training* dapat meningkatkan akurasi sebab data *training* lebih besar daripada data *testing.* Selain itu, tidak adanya pengacakan memberikan hasil sama ketika LOOCV diulang. Akan tetapi, metode LOOCV memerlukan waktu lebih lama dalam komputasi sebab terdapat iterasi sebanyak n.

E. Metode Prediksi Laju Galat *K-Fold Cross Validation*

Menurut James, dkk. (2013: 181) metode *K-Fold CV* merupakan metode prediksi yang membagi secara acak dataset kedalam K kelompok dengan ukuran yang sama. Data *training* terdiri dari K-1 kelompok amatan, dan 1 kelompok amatan lain digunakan sebagai data *testing.* Pada *K-Fold CV* dilakukan iterasi sebanyak *K*. Setiap iterasi dihitung laju galat kelompok *testing* dan laju galat *K-Fold CV* menggunakan Persamaan 3 berikut.

$$\hat{E}^{CV} = \frac{1}{K}\sum_{i=1}^{K}\frac{1}{n_K}\sum_{i=1}^{n_K} Err_i \quad (3)$$

dimana.

$\hat{E}^{CV}$ = Prediksi laju galat metode *K-Fold CV*

K = Jumlah iterasi

n_K = Jumlah data dalam kelompok *testing*

Sifat-sifat metode prediksi laju galat K-Fold CV yaitu sering memberikan perkiraan *error rate* yang lebih akurat daripada metode LOOCV, artinya K-Fold CV memiliki estimasi kinerja yang lebih baik daripada metode LOOCV namun tidak ada jaminan hasil ini selalu didapat. Faktor perbedaan prinsip kerja algoritma pembentuk model maupun kondisi dataset yang digunakan dapat mempengaruhi hasil perkiraan *error rate*. Sementara itu, tidak ada jaminan tiap *fold* menghasilkan laju klasifikasi (*misclassification rate*) yang sama dan tidak optimal pada dataset berukuran kecil.

F. Jenis dan Sumber Data

Jenis data yang digunakan merupakan data simulasi dengan sumber data dari proses bangkitan data *Software R Studio*. Penelitian membandingkan metode prediksi laju galat HO, LOOCV, dan K-Fold CV dengan melihat kestabilan nilai median *error rate* dan variasi *error rate* dari prediksi laju galat yang dihasilkan. Pada penelitian ini, masing-masing metode prediksi laju galat menghasilkan 100 nilai prediksi *error rate* yang diperoleh dari hasil iterasi 100 amatan seperti dalam poin 5 langkah analisis. Nilai prediksi inilah yang nantinya digunakan dalam membentuk grafis *boxplot* untuk melihat kestabilan nilai median *error rate* dan variasi *error rate*.

Pemodelan pohon keputusan dengan algoritma C4.5 menggunakan variabel respon bertipe data kategorik yang dibangkitkan dengan kategori 0 dan 1. Sedangkan untuk variabel prediktor akan dibangkitkan dengan beberapa pengaturan berikut.

1. Pengaturan jumlah variabel penjelas.

Pengaturan data univariat menggunakan satu buah variabel prediktor sedangkan pengaturan data bivariat menggunakan dua buah variabel prediktor. Pengaturan data bivariat dikaji untuk kasus yang memungkinkan simulasi yang melibatkan struktur korelasi dan struktur rataan yang berbeda.

2. Pengaturan perbedaan rataan populasi

Data pada kasus univariat dibangkitkan dengan distribusi normal univariat N($\boldsymbol{\mu}, \boldsymbol{\sigma}$) menggunakan pengaturan rataan populasi $\boldsymbol{\mu_1} = 0$ untuk kategori Y=0, $\boldsymbol{\mu_2} = 1$ untuk kategori Y=1, dan $\boldsymbol{\sigma} = 1$.. Pada kasus data bivariat, pengaturan rataan populasi diuraikan dalam Tabel 1.

Tabel 1. Pengaturan Rataan Populasi Data Bivariat

Pengaturan	$\underline{\boldsymbol{\mu_1}}$	$\underline{\boldsymbol{\mu_2}}$	$\boldsymbol{\sigma}$
1	$\begin{bmatrix}0\\0\end{bmatrix}$	$\begin{bmatrix}1\\1\end{bmatrix}$	$\begin{bmatrix}1\\1\end{bmatrix}$
2	$\begin{bmatrix}0\\0\end{bmatrix}$	$\begin{bmatrix}1\\0\end{bmatrix}$	$\begin{bmatrix}1\\1\end{bmatrix}$

Berdasarkan Tabel 1, pengaturan pertama merupakan kondisi dua variabel penjelas yang memuat informasi tentang beda rataan. Artinya, setiap kategori (Y=0 dan Y=1) memiliki pengaturan rataan populasi yang berbeda yaitu $\boldsymbol{\mu_1} = 0$ untuk kategori Y=0, $\boldsymbol{\mu_2} = 1$ untuk kategori Y=1. Kondisi ini disebut kasus data bivariat antar variabel relevan. Sementara itu, pengaturan kedua merupakan kondisi dua variabel penjelas dengan variabel penjelas ke-1 memuat informasi beda rataan dengan $\boldsymbol{\mu_1} = 0$ untuk kategori Y=0 dan $\boldsymbol{\mu_2} = 1$ untuk kategori Y=1. Sedangkan variabel penjelas ke-2 tidak memuat informasi beda rataan sebab rataan populasi tiap kategori sama, yaitu $\boldsymbol{\mu_1} = 0$ untuk

DOI: https://doi.org/10.24036/ujsds/vol1-iss4/89

ISSN(*Online*): 2985-475X

kategori Y=0 dan $\boldsymbol{\mu_2} = 0$ untuk kategori Y=1. Kondisi ini disebut kasus data bivariat variabel relevan dengan variabel irrelevan.

3. Pengaturan korelasi antar variabel penjelas untuk kasus data bivariat.

Pada pengaturan korelasi antar variabel penjelas kasus data bivariat, digunakan 3 pengaturan korelasi seperti dalam Tabel 2.

Tabel 2. Pengaturan Struktur Korelasi Data Bivariat

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

Pada Tabel 2**,** pengaturan 1 menunjukkan kasus data bivariat tanpa korelasi**,** pengaturan 2 menunjukkan kasus data bivariat dengan korelasi sedang, dan pengaturan 2 menunjukkan kasus data bivariat dengan korelasi tinggi.

4. Pengaturan Proporsi Data Tidak Seimbang

Proporsi kelas menunjukkan tingkat ketidakseimbangan data yang digunakan, yaitu perbandingan banyak amatan pada kategori Y_0 dan Y_1. Menurut Vluymans (2019: 83) ketidakseimbangangan data diukur dengan rasio ketidakseimbangan (IR). IR = 1 berarti data seimbang sempurna, IR >1,5 dianggap data tidak seimbang, dan IR = 9 berarti data sangat tidak seimbang. Oleh karena itu, digunakan pengaturan proporsi kelas data seperti Tabel 3.

Tabel 3. Pengaturan Proporsi Kelas Data

Pengaturan	Proporsi kelas
1	50 : 50
2	60 : 40
3	90 : 10

Pada Tabel 3, pengaturan 1 mewakili IR=1, pengaturan 2 mewakili IR > 1,5, dan pengaturan 3 mewakili IR =9.

Empat poin pengaturan tersebut digunakan dalam pembangkitan dataset yang akan digunakan dalam penelitian pada langkah awal analisis data. Dataset bivariat dalam penelitian ini mengkombinasikan struktur beda rataan pada poin 2 dan struktur korelasi pada poin 3. Tujuan kombinasi yaitu mengkaji kasus data tidak seimbang pada poin 4 dengan kondisi tanpa korelasi, korelasi sedang, serta korelasi tinggi antar variabel relevan maupun antar variabel relevan dan variabel irrelevan untuk kasus data tidak seimbang.

G. Langkah Analisis

Penelitian dilakukan dengan langkah-langkah sebagai berikut.

1. Membangkitkan dataset bilangan acak menggunakan *function data.generating* dengan ketentuan pengaturan data yang telah ditetapkan.
2. Memprediksi nilai laju galat dengan metode prediksi laju galat HO dengan langkah sebagai berikut.
   a. Dataset sebanyak 100 amatan dibagi menjadi 2/3 data *training* dan 1/3 data *testing*.
   b. Membentuk model algoritma C4.5 dengan struktur tree dengan data *training*.
   c. Melakukan prediksi dengan data *testing* dan menghitung *error*.
   d. Menghitung *error rate* HO dengan Persamaan 1 .
3. Memprediksi nilai laju galat dengan metode prediksi laju galat LOOCV dengan langkah sebagai berikut.
   a. Dataset sebanyak 100 amatan dibagi menjadi n-1 data *training* dan 1 amatan sisanya sebagai data *testing*.
   b. Membentuk model algoritma C4.5 dengan struktur *tree* dengan data *training*.
   c. Menguji prediksi dengan data *testing* dan menghitung *error*.
   d. Melakukan iterasi sebanyak 100 amatan.
   e. Menghitung *error rate* LOOCV dengan Persamaan 2.
4. Memprediksi nilai galat dengan metode prediksi laju galat *5-Fold CV* dengan langkah berikut.
   a. Dataset 100 amatan dibagi acak dalam 5 kelompok (4 kelompok data *training* dan 1 kelompok data *testing*).
   b. Membentuk model algoritma C4.5 dengan struktur *tree* dengan data *training*.

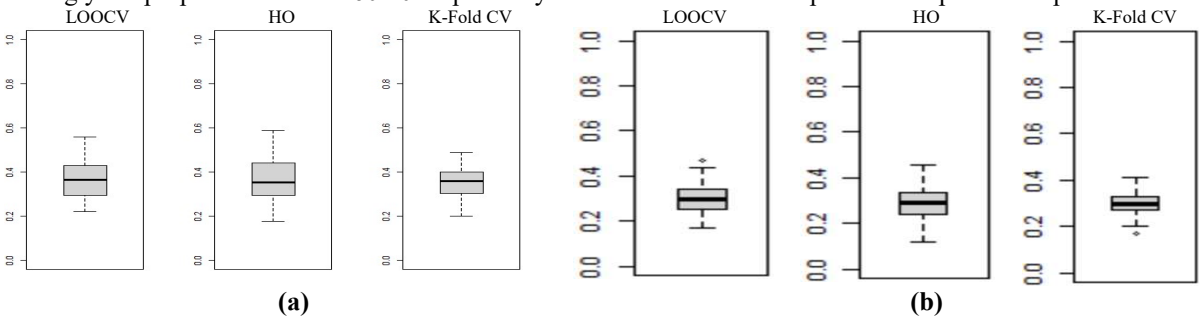
DOI: https://doi.org/10.24036/ujsds/vol1-iss4/89

c. Menguji prediksi dengan *data testing* dan menghitung *error*.
d. Menghitung *error rate* K-Fold CV dengan Persamaan 3.

5. Melakukan iterasi sebanyak 100 kali pada masing-masing metode prediksi laju galat.
6. Membuat *boxplot* masing-masing metode prediksi galat untuk melihat variasi *error rate* dan median *error rate*.
7. Membandingkan ketiga metode prediksi laju galat dan memilih metode prediksi laju galat terbaik.
8. Membuat kesimpulan akhir.

III. HASIL DAN PEMBAHASAN

Model pohon keputusan dibangun setiap dataset telah dibagi dalam data *training* dan *testing* pada setiap metode prediksi laju galat, sehingga pembentukan pohon keputusan terjadi berkali-kali. Setiap metode prediksi laju galat, perhitungan *error rate* dilakukan setelah model pohon keputusan selesai dibangun. Bedanya, dalam metode prediksi laju galat HO satu nilai *error rate* HO didapat tanpa melakukan iterasi. Pada metode LOOCV, untuk memperoleh satu nilai *error rate* LOOCV harus merata-ratakan 100 *error rate* karena metode LOOCV memiliki n buah iterasi. Sementara itu, pada metode K-Fold CV dengan K = 5, satu nilai *error rate* 5-Fold CV didapat dari merata-ratakan 5 buah *error rate* karena memiliki k buah iterasi. Satu nilai *error rate* dari masing-masing metode prediksi laju galat tidak dapat digunakan untuk membandingkan kinerja metode prediksi laju galat. Perbandingkan kinerja metode prediksi laju galat HO, LOOCV, dan *K-Fold CV* pada pemodelan klasifikasi algoritma C4.5 untuk kasus data tidak seimbang dilihat berdasarkan tampilan grafis *boxplot*. Oleh karena itu, diperlukan poin 5 langkah analisis untuk menghasilkan 100 buah nilai prediksi *error rate* setiap metode prediksi laju galat. Hasil ke-100 buah nilai prediksi *error rate* setiap metode prediksi laju galat itulah yang digunakan untuk membentuk grafis *boxplot* tiap metode prediksi laju galat. Grafis *boxplot* akan memberikan gambaran kestabilan nilai median *error rate* dan variasi *error rate dari 100* nilai prediksi *error rate* setiap metode prediksi laju galat. Variasi *error rate* dapat dilihat dari besarnya kotak boxplot atau nilai IQR(*InterQuartileRange*) *boxplot* yang merupakan jangkauan dalam kuartil yang dapat dihitung dengan cara ($Q_3 - Q_1$), sedangkan median *error rate* dilihat dari garis tengah kotak *boxplot* atau nilai mediannya. Metode prediksi laju galat yang paling cocok digunakan pada pemodelan klasifikasi algoritma C4.5 untuk kasus data tidak seimbang diindikasi memiliki nilai median *error rate* yang stabil serta variasi *error rate* terkecil. Kinerja metode prediksi laju galat HO, LOOCV, dan K-Fold CV dilihat dari jumlah variabel prediktor dalam pengaturan data tidak seimbang yaitu proporsi kelas data 60:40 tanpa adanya korelasi antar variabel prediktor dapat dilihat pada Gambar 1.



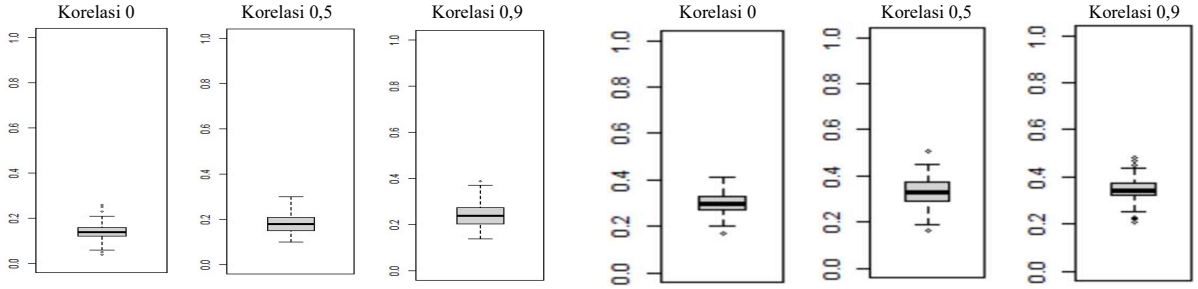
Gambar 1. *Boxplot* Proporsi 60:40 tanpa korelasi (a) Data Univariat (b) Data Bivariat

Tampilan Gambar 1 menunjukkan bahwa pada pengaturan proporsi kelas data tidak seimbang tanpa adanya korelasi antar variabel prediktor, metode prediksi laju galat K-Fold CV menjadi metode prediksi laju galat terbaik sebab memiliki kinerja prediksi *error rate* yang lebih optimal dibanding metode LOOCV dan HO baik pada data univariat maupun data bivariat. Pada keseluruhan simulasi yang telah dilakukan, metode prediksi laju galat K-Fold CV memiliki kestabilan nilai median *error rate* yang sangat baik dan konsisten memperlihatkan variasi *error rate* terkecil dibandingkan metode prediksi galat yang lain. Tiga faktor yang mendukung hal tersebut yaitu model pohon keputusan dibangun dari data *training* yang cukup besar, faktor iterasi, dan faktor keacakan. Model pohon keputusan yang dibangun dengan data *training* yang cukup besar menyebabkan model memiliki cukup informasi untuk melakukan prediksi pada data baru, sehingga dapat memperkecil laju *error* yang dihasilkan. Faktor iterasi berguna untuk mengevaluasi hasil *error rate* sebab setiap iterasi membentuk pohon keputusan yang baru dengan data *training* yang berbeda dengan sebelumnya. Faktor keacakan dalam memilih data *training* dan *testing* menyebabkan model pohon keputusan dibentuk dari amatan yang tidak identik sehingga dapat memperkecil variasi *error rate*. Metode prediksi

DOI: https://doi.org/10.24036/ujsds/vol1-iss4/89

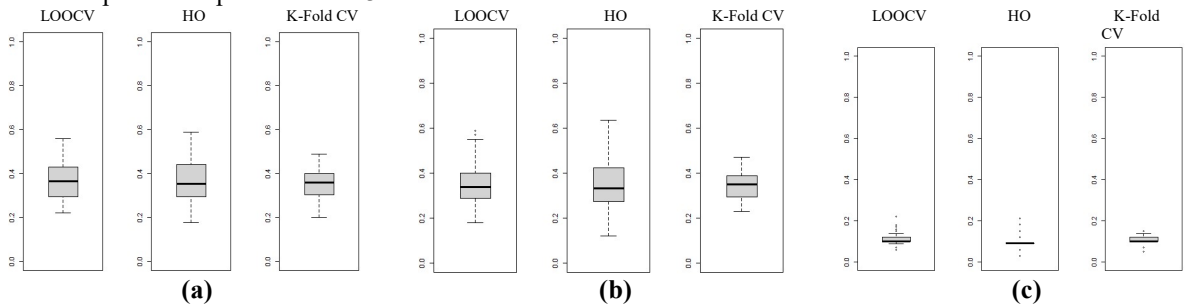
ISSN(*Online*): 2985-475X

laju galat LOOCV memiliki data *training* dan iterasi yang besar. Akan tetapi, dalam iterasinya, model pohon keputusan dibentuk dari amatan yang identik sehingga variasi *error rate* cenderung lebih besar dibanding K-Fold CV. Metode prediksi laju galat HO memiliki nilai median *error rate* cenderung tidak stabil dengan variasi *error rate* yang besar, sebab model pohon keputusan hanya dibangun satu kali dengan data *training* yang lebih kecil daripada metode prediksi laju galat LOOCV dan K-Fold CV. Selanjutnya, tampilan Gambar 2 merupakan kinerja metode prediksi laju galat K-Fold CV dalam kasus data bivariat antar variabel relevan dan kasus data bivariat variabel relevan dengan variabel irrelevan dilihat dari perbedaan korelasi antar variabel prediktor pengaturan proporsi kelas data 60:40.



Gambar 2. *Boxplot* Data Bivariat Prediksi Laju Galat K-Fold CV Proporsi 60:40 (a) Antar Variabel Relevan (b) Variabel Relevan dan Irrelevan

Tampilan Gambar 2 menunjukkan bahwa pada kasus data bivariat antar variabel relevan, korelasi tinggi cenderung memperbesar variasi *error rate* metode prediksi laju galat K-Fold CV diproporsi data tidak seimbang. Pola yang sama juga terlihat pada pengaturan proporsi kelas data dan metode prediksi laju galat yang lainnya. Tingginya korelasi juga cenderung mempengaruhi besar median *error rate* yang dihasilkan oleh metode prediksi laju galat K-Fold CV. Semakin besar korelasi yang diberikan maka semakin besar pula nilai median *error rate* yang diperoleh. Pola yang sama juga ditunjukkan data bivariat antar variabel relevan pada metode prediksi laju galat HO dan LOOCV dipengaturan proporsi kelas data 50:50 dan 60:40. Akan tetapi pola ini tidak berlaku pada pengaturan proporsi kelas data yang ekstrim atau pada proporsi 90:10, sebab nilai median *error rate* yang dihasilkan metode prediksi laju galat HO, LOOCV, dan K-Fold CV sama disemua korelasi yang diberikan. Sementara itu, pada kasus data bivariat variabel relevan dengan variabel irrelevan semakin besar korelasi yang diberikan, nilai median *error rate* yang dihasilkan tidak selalu mengalami peningkatan. Artinya semakin besar korelasi yang diberikan nilai median *error rate* tidak konsisten menunjukkan nilai semakin tinggi namun juga menunjukkan semakin besar korelasi yang diberikan nilai median *error rate* semakin menurun. Kinerja ketiga metode prediksi laju galat pada data univariat dilihat dari pengaturan proporsi kelas data dapat dilihat pada Gambar 3.



Gambar 3. *Boxplot* Data Univariat (a) Proposi 50:50 (b) Proporsi 60:40 (c) Proporsi 90:10

Pada Gambar 3 bagian (a) terlihat bahwa metode prediksi laju galat K-Fold CV memiliki variasi *error rate* terkecil dibandingkan metode prediksi laju galat LOOCV dan HO. Artinya metode K-Fold CV berkinerja baik dalam pengaturan dataset univariat dengan proporsi kelas 50:50. Pada bagian (b), terlihat bahwa metode prediksi laju galat K-Fold CV juga memiliki variasi *error rate* terkecil dibandingkan metode prediksi laju galat HO dan LOOCV yang berarti metode K-Fold CV berkinerja baik dalam pengaturan dataset univariat dengan proporsi kelas data tidak seimbang yaitu proporsi kelas data 60:40. Sedangkan pada bagian (c), terlihat bahwa metode prediksi laju galat HO memiliki variasi *error rate* terkecil dibandingkan metode prediksi laju galat K-Fold CV dan LOOCV. Artinya metode

DOI: https://doi.org/10.24036/ujsds/vol1-iss4/89

HO berkinerja baik dalam pengaturan dataset univariat dengan proporsi ambang batas kelas data sangat tidak seimbang yaitu 90:10. Pola yang sama juga ditunjukkan pada data bivariat baik pada kasus data bivariat antar variabel relevan ataupun pada kasus data bivariat variabel relevan dengan variabel irrelevan. Hasil simulasi secara keseluruhan memperlihatkan pada proporsi data seimbang hingga data tidakseimbang metode prediksi laju galat K-Fold CV cenderung memiliki variasi *error rate* terkecil dibandingkan metode prediksi laju galat LOOCV dan HO. Akan tetapi, pada ketidakseimbangan data yang ekstrim, ketiga metode prediksi laju galat tidak bekerja secara optimal dibuktikan dengan perolehan nilai median *error rate* yang konstan disemua pengaturan simulasi data univariat dan bivariat. Selain itu, didapati metode prediksi laju galat HO yang cenderung menghasilkan variasi *error rate* terkecil.

IV. KESIMPULAN

Metode prediksi laju galat K-Fold CV memiliki kinerja paling baik pada algoritma C4.5 dalam kasus data tidak seimbang dibandingkan metode HO dan LOOCV, sebab cenderung memiliki nilai median *error rate* yang stabil dan variasi *error rate* terkecil. Pada data bivariat, pengaruh korelasi lebih terlihat pada kasus data bivariat antar variabel relevan. Korelasi yang tinggi cenderung memperbesar nilai median *error rate* dan variasi *error rate* yang dihasilkan metode prediksi laju galat. Sehingga korelasi yang tinggi menjadi salah satu faktor yang menyebabkan kinerja metode prediksi laju galat menurun selain faktor ketidakseimbangan data. Pada kasus data bivariat variabel relevan dengan variabel irrelevan, korelasi yang tinggi dapat meningkatkan dan menurunkan nilai median *error rate* dan variasi *error rate* yang dihasilkan. Peneliti merekomendasikan penggunaan metode prediksi laju galat K-Fold CV untuk evaluasi kinerja model yang dibentuk menggunakan metode klasifikasi algoritma C4.5. Selain itu, direkomendasikan pula melakukan penanganan data tidak seimbang terlebih dahulu khususnya pada data dengan ketidakseimbangan ekstrim, sebab ketidakseimbangan data yang ekstrim membuat metode prediksi galat bekerja tidak optimal.

DAFTAR PUSTAKA

Chawla, N. V. (2009). Data mining for imbalanced datasets: An overview. *Data mining and knowledge discovery handbook*, 875-886.

Efron, B. & Tibsirani, R. J. 1993. *An Introduction to the Bootstrap.* New York: Chapman & Hall.

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

Kohavi, R. (1995, August). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai* (Vol. 14, No. 2, pp. 1137-1145).

Kotsiantis, S., Kanellopoulos, D., & Pintelas, P. (2006). Handling imbalanced datasets: A review. *GESTS international transactions on computer science and engineering*, *30*(1), 25-36.

Ling, C. X., Huang, J., & Zhang, H. (2003). AUC: a better measure than accuracy in comparing learning algorithms. In *Advances in Artificial Intelligence: 16th Conference of the Canadian Society for Computational Studies of Intelligence, AI 2003, Halifax, Canada, June 11–13, 2003, Proceedings 16* (pp. 329-341). Springer Berlin Heidelberg.

Mingers, J. (1989). An empirical comparison of pruning methods for decision tree induction. *Machine learning*, *4*(2), 227-243.

Molinaro, A. M., Simon, R., & Pfeiffer, R. M. (2005). Prediction error estimation: a comparison of resampling methods. *Bioinformatics*, *21*(15), 3301-3307.

Prasetyo, E. (2014). *Data Mining Mengolah Data Menjadi Informasi Menggunakan Matlab*. Yogyakarta: Penerbit Andi.

Quinlan, J. R. 1993. *C4.5: Program for Machine Learning*. San Mateo: Morgan Kaufmann

Refaeilzadeh, P., Tang, L., & Liu, H. (2009). Cross-validation. *Encyclopedia of database systems*, *5*, 532-538.

Rokach, L., & Maimon, O. (2005). Top-down induction of decision trees classifiers-a survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *35*(4), 476-487.

Simon, R., Radmacher, M. D., Dobbin, K., & McShane, L. M. (2003). Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *Journal of the National Cancer Institute*, *95*(1), 14-18.

Vluymans, S. 2019. *Dealing with imbalanced and weakly labelled data in machine learning using fuzzy and rough set methods* (Vol. 807). Belgium: Springer

Zheng, X., Feng, W., Huang, M., & Feng, S. (2021). Optimization of PBFT algorithm based on improved C4. 5. *Mathematical Problems in Engineering*, *2021*, 1-7

DOI: https://doi.org/10.24036/ujsds/vol1-iss4/89