UNP JOURNAL OF STATISTICS AND DATA SCIENCE
Volume I Nomor 4 Tahun 2023 Halaman 297 - 305
ISSN(*Online*): 2985-475X

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

Ichlas Djuazva, Dodi Vionanda*, Nonong Amalita, Zilrahmi
Departemen Satatistika, Universitas Negeri Padang, Padang, Indonesia
*Corresponding author: dodi_vionanda@fmipa.unp.ac.id

Submitted : 13 Juni 2023
Revised : 24 Juli 2023
Accepted : 15 Agustus 2023

ABSTRACT

C4.5 is a highly effective decision tree algorithm for classification purposes. Compared to CHAID, Cart, and ID3, C4.5 generates the decision tree faster and is easier to understand. However, C4.5 algorithm is also not exempt from errors in classification, which can impact the accuracy of the resulting model. Model accuracy could be measured by predicting the error rate. One commonly used method for error rate prediction is cross-validation. The cross-validation method divides data into two parts: training set to build model and testing set to test the model. There are several cross-validation techniques commonly used to predict the error rate, such as Leave One Out Cross Validation (LOOCV), Hold Out (HO), and k-fold cross-validation. LOO has unbiased estimation but takes a long time and depends on the data size; HO could avoid overfitting and work faster; and k-folds cross validation has a smaller error rate prediction. This study uses artificially generated data with a normal distribution, including univariate, bivariate, and multivariate datasets with various combinations of mean differences and different correlations. Different correlation structures are applied to see the impact of these different correlations on the error rate prediction method. Considering these factors, this research focuses on comparing three cross-validation methods to predict error rates for the decision tree model generated by C4.5 algorithm. This research found that k-folds cross-validation is the most suitable cross-validation method to apply when testing the model generated by C4.5 algorithm with balanced data.

Keywords: *C4.5, Error Prediction, Hold Out, K-folds, Leave One Out.*

I. PENDAHULUAN

Algoritma C4.5 adalah salah satu algoritma pohon keputusan terbaik yang diketahui dan juga paling luas penggunaanya. Tingkat akurasinya cukup tinggi, terlepas dari volume data yang akan diproses, hal ini disebutkan oleh Lu dkk (2015) pada penelitianya. Salah satu studi terbaru yang dilakukan oleh Hssina (2014) membandingkan pohon keputusan dan algoritma pembelajaran lainnya menunjukkan bahwa C4.5 memiliki kombinasi tingkat kesalahan dan kecepatan yang sangat baik dan menghasilkan pohon keputusan yang lebih kecil daripada metode lain seperti CART, CHAID, dan ID3 sehingga waktu yang dibutuhkan dalam pembentukan pohonnya relatif lebih cepat.

García dkk (2015) juga menyatakan algoritma C4.5 dapat mengatasi *dataset training* yang tidak lengkap, algoritma ini juga dapat mengatasi atribut kontinu. Algoritma C4.5 menggunakan strategi *top down* memilih atribut yang memiliki informasi paling banyak hingga paling sedikit dikombinasikan dengan pendekatan *divide and conquer* memecah data menjadi beberapa kelompok masalah yang lebih kecil dalam membentuk pohon keputusan (H. Liu dan Gegov, 2016). Kombinasi strategi *top down* dan pendekatan *conquer and divide* ini dapat menghasilkan pohon yang sederhana dalam waktu yang cepat serta memberikan hasilyang lebih berfokus kepada permasalahan yang diteliti sehingga penelitian ini akan berfokus pada algoritma C4.5.

Algoritma C4.5 akan menghasilkan model berupa diagram alir yang berbentuk seperti pohon yang dapat terjadi kesalahan dalam pembentukan model maupun dalam klasifikasi. Menurut Dougherty dkk (2010) metode prediksi laju galat (*error prediction*) adalah metode yang umum digunakan untuk mengevaluasi kinerja model. Evaluasi model

DOI: https://doi.org/10.24036/ujsds/vol1-iss4/74

ISSN(*Online*): 2985-475X

perlu dilakukan untuk melihat kemampuan model dalam mlakukan prediksi, serta untuk melihat kecocokan model terhadap data sehingga klasifikasi yang dihasilkan dari model dapat dilihat keakuratanya. Ada dua metode prediksi laju galat yang dapat digunakan dalam menguji kinerja model yaitu *training error rate* dan *testing error rate*. Menurut Mansour dan McAllester (2002) dalam memprediksi laju galat *training error rate* menggunakan data *training* atau data yang telah digunakan untuk membentuk model, hal ini akan mengakibatkan nilai prediksi *error* yang dihasilkan menjadi rendah bahkan nol karena data yang sama digunakan untuk membentuk model dan menguji model. Sedangkan *testing error rate* sendiri membagi data menjadi dua yaitu data *training* yang akan digunakan membentuk model dan data *testing* untuk menguji akurasi dari model, metode ini dapat mengatasi *underestimate* pada model. Peneltian ini menggunakan *cross validation* dalam melakukan prediksi laju galat pada algoritma C4.5. *Cross validation* sendiri merupakan metode *testing error rate* karena metode ini membagi data menjadi *training* dan *testing*. *Cross validation* terdiri dari tiga metode yaitu *Leave One Out Cross Validation* (LOOCV), *Hold Out* (HO), dan *k-folds cross validation*.

Berdasarkan kelebihan yang dimiliki metode *testing error rate* serta diperkuat dengan penelitian yang dilakukan oleh Tougui (2021) yang menyatakan bahwa metode *cross validation* memiliki peforma yang lebih baik dibandingkan dengan metode *testing error rate* lainya yaitu *bootstrap* maka penelitian ini membandingkan peforma ketiga metode *cross validation* tersebut yang digunakan dalam memprediksi laju galat pada algoritma C4.5. Penelitian bertujuan untuk memperoleh metode prediksi laju galat yang cocok digunakan pada algoritma C4.5 dengan membandingkan laju galat yang dihasilkan masing masing metode 3 metode *cross validation* yaitu LOOCV, HO dan *k-folds* yang disajikan ke dalam bentuk boxplot dengan membandingkan *Inter Quartil Range* (IQR). Penggunaan metode prediksi laju galat yang cocok akan menghasilkan model dengan akurasi dan kinerja yang baik. Penelitian ini menggunakan data bangkitan dengan pengaturan beda rataan dengan kasus data univariat, bivariat, dan multivariat serta beda rataan yang dikombinasikan dengan beda struktur korelasi untuk kasus bivariat dan multivariat. Pengaturan beda struktur korelasi diterapkan untuk melihat pengaruh penambahan korelasi antara dua variabel *relevant* dan antara variabel *relevant* dengan *irrelevant* terhadap laju galatnya.

II. METODE PENELITIAN

A. Algoritma C4.5

C4.5 merupakan salah pengembangan algoritma ID3 dengan beberapa peningkatan, salah satunya penanganan atribut kontinu. Atribut tersebut akan diganti dengan atribut diskret menggunakan ambang nilai yang memisahkan data menjadi dua interval (Behera dkk, 2015). Menurut Quinlan (1996) algoritma C4.5 menggunakan kriteria *split* khusus yang telah dimodifikasi yang dinamakan *gain ratio* dalam pemilihan *split* atributnya. Jika prediktor (X) memiliki nilai numerik, maka dilakukan proses *binarization* (membagi data kontinu menjadi dua kategori) 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 nilai terkecil ke terbesar.
2) Melakukan *cut-off* dengan rumus $C_i = \frac{x_i + x_{(i+1)}}{2}$, untuk $i = 1, 2, \ldots, n-1$.
3) Masing-masing nilai *cut-off* dikategorikan ($\leq$ dan $>$) dan dilakukan perhitungan nilai kriteria *Gain*.
4) Nilai *cut-off* dengan kriteria Gain tertinggi merupakan nilai terbaik untuk dipilih sebagai z.

Pada tahapan training, C4.5 menggunakan strategi *top down* yang didasarkan pada pendekatan *divide and conquer* untuk membentuk pohon keputusan (Liu dan Gegov, 2016). Tahapan ini memetakan *training set* dan dengan informasi *gain ratio* sebagai tolak ukur untuk memisahkan atribut dan menghasilkan *nodes* dari akar hingga daun (Dai dan Ji, 2014). *Gain ratio* dapat diperoleh menggunakan rumus dengan S merupakan himpunan kasus, A adalah atribut, sedangkan n jumlah partisi dan p_i adalah proporsi S_i terhadap S.

$$GainRatio\ (S,A) = \frac{Gain\ (S,A)}{SplitInfo\ (S,A)}$$

Gain sendiri dapat diperoleh dengan rumus :

$$Gain(S,A) = Entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} \times Entropy(S_i)$$

dengan

$$Entropy\ (S) = \sum_{i=1}^{n} -p_i \times log_2 p_i$$

Split info dapat diperoleh dengan rumus:

DOI: https://doi.org/10.24036/ujsds/vol1-iss4/74

ISSN(*Online*): 2985-475X

$$SplitInfo(S,A) = -\sum_{i=1}^{n} \frac{S_i}{S} log_2 \frac{S_i}{S}$$

Dalam pembentukan pohon pada algortima C4.5 setiap cabang akan berakhir pada sebuah simpul yang disebut simpul daun (*leaf node)*. Simpul daun yang dihasilkan ini akan mengarah pada kelas data yang dipilih dengan melihat nilai dari kelas pada variabel respon. Jika data pada sebuah cabang tidak berada pada kelas yang sama maka simpul daun akan dibentuk dari kelas terbanyak dan akan dibentuk juga cabang baru atau kasus baru . Jika semua data pada suatu cabang berada di kelas yang sama maka cabang tersebut akan menjadi daun atau keputusan yang dilambangkan dengan $\hat{y}_i$. Algoritma akan dihentikan ketika semua cabang dalam pohon menghasilkan sebuah daun yang mewakili kelas yang sama.

Korelasi pada algoritma C4.5 sangat mempengaruhi pemilihan atribut dalam membentuk pohon. Pada algoritma C4.5 *gain ratio* digunakan untuk memilih atribut pada setiap tahapan dalam metode C4.5. Menurut Zheng dkk (2021) semakin tinggi *gain ratio* suatu atribut semakin besar pula korelasi atribut tersebut dengan atribut yang terpilih sebagai atribut kelas sebelumnya, sehingga semakin besar kemungkinan atribut tersebut untuk terpilih sebagai atribut kelas. Hal ini dapat menimbulkan dampak negatif yaitu atribut yang terpilih akan cenderung memiliki informasi yang sama sementara atribut dengan informasi yang berbeda berkemungkinan tidak terpilih.

B. Cross Validation

Menurut Braga-Neto dan Dougherty (2015) klasifikasi yang baik akan menghasilkan hasil klasifikasi dengan rata-rata *error rate* yang kecil. Prediksi laju galat sangat penting untuk klasifikasi karena validitas model *classifier* yang dihasilkan, didasarkan pada keakuratan prosedur estimasi *error* (Dougherty dan Braga-Neto, 2006). Pada kumpulan data sampel yang besar data dapat dibagi antara data *training* dan *testing*, dengan pengklasifikasi dirancang pada data *training* dan kesalahannya diperkirakan pada data *testing* (E. Dougherty dkk, 2010). Prediksi laju galat dapat dihitung dengan menggunakan *missclassification rate*, dengan rumus berikut:

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

Di mana y_i adalah data aktual amatan *ke-i*, di mana *i*= 1, 2, 3, …, n dan $\hat{y}_i$ adalah data hasil prediksi *ke-i*,yang merupakan kelas data yang dihasilkan yang terdapat pada *node* daun pada pohon, di mana *i* = 1, 2, 3, …, n. Dengan indikator variabelnya jika $y_i \neq \hat{y}_i$ adalah 0, dan $y_i = \hat{y}_i$ adalah 1. Jika 0 maka data diklasifikasikan dengan benar dan jika 1 maka terdapat *error* atau kesalahan klasifikasi dari model. Dalam *cross validation* terdapat tiga metode yaitu HO, LOOCV dan *k-fold cross validation*.

1. Hold Out

Hold out cross validation merupakan metode *cross validation* yang paling sederhana (James dkk., 2013). Metode ini membagi 2/3 data menjadi data *training* dan 1/3 data lainnya menjadi data *testing* (Kohavi, 1995). Perhitungan prediksi laju galat dengan metode *hold out* dapat dilakukan dengan rumus:

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

Dimana $\hat{E}^{HO}$ adalah prediksi laju galat dengan metode *hold out*, n_{uji} adalah banyak amatan pada data uji sedangkan $I(y_i \neq \hat{y}_i)$ indeks akurasi. Indeks akurasi adalah rasio antara jumlah prediksi benar dengan jumlah total prediksi yang dilakukan oleh model pada keseluruhan data. Indeks ini memberikan gambaran umum tentang seberapa baik model dalam melakukan prediksi secara keseluruhan.

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

Menurut Hastie dkk (2013) kesalahan pengujian LOOCV merupakan perkiraan yang tidak bias dari kesalahan prediksi yang sebenarnya, tetapi memiliki variansi yang tinggi karena *set training* praktis sama. Data *training* berisi seluruh pengamatan kecuali satu pengamatan, dan *testing* hanya berisi satu amatan data. *Set training* pertama berisi seluruh amatan kecuali observasi 1, *set training* kedua berisi seluruh amatan kecuali observasi 2, dan seterusnya. Perhitungan galat untuk metode LOOCV dapat diperoleh dengan menggunakan rumus:

$$\hat{E}^{LOOCV} = \frac{1}{n}\sum_{i=1}^{n} (\hat{E}_i)$$

Di mana $\hat{E}^{LOOCV}$ merupakan prediksi laju galat dengan metode *LOOCV*, n banyaknya amatan, dan $\hat{E}_i$ galat pada amatan ke-i yang diperoleh dengan rumus *misclassification rate*. LOOCV memiliki beberapa keunggulan utama dibandingkan *hold out*. Biasnya jauh lebih sedikit akan tetapi LOOCV bisa sangat memakan waktu jika datanya berskala besar, dan jika masing-masing model lambat untuk disesuaikan (James dkk, 2013).

DOI: https://doi.org/10.24036/ujsds/vol1-iss4/74

3. *K-Folds Cross Validation*

Dalam *k-folds cross validation* data set yang tersedia dipartisi menjadi k kelompok data dengan ukuran yang kira-kira sama, "*folds*" mengacu pada jumlah kelompok data yang dihasilkan. Model dilatih menggunakan himpunan bagian $k-1$, yang menjadi *training*. Kemudian model diterapkan ke *subset* yang tersisa, yang dilambangkan sebagai *set testing*, dan kinerjanya diukur. Prosedur ini diulang sampai masing-masing *subset* telah berfungsi sebagai *testing* (Berrar, 2019).

Prediksi laju galat pada *k-folds cross validation* dapat diperoleh dengan menggunakan rumus:

$$\hat{E}^{CV} = \frac{1}{k}\sum_{i=1}^{k}(\hat{E}_i)$$

Dimana $\hat{E}^{CV}$ adalah prediksi laju galat dengan metode *k-fold*, k adalah jumlah kelompok data, dan $\hat{E}_i$ merupakan galat pada iterasi *ke-i*.

C. Jenis dan Sumber Data

Penelitian ini membandingkan kinerja metode prediksi laju galat yang diterapkan dalam pemodelan pohon keputusan menggunakan algoritma C4.5 untuk kasus data yang seimbang. Penelitian ini juga memperhatikan nilai rataan dan variansi galat yang diperoleh dari masing-masing metode prediksi laju galat. Metode prediksi laju galat yang digunakan adalah LOOCV, HO, dan *k-fold cross validation*. Penelitian ini akan menggunakan data yang dibangkitkan dengan *software* R-Studio. Variabel respon untuk data yang dibangkitkan terdiri dari dua kelas yang bernilai 1 dan 2. Untuk variabel prediktor sendiri ada beberapa perlakuan yang akan diterapkan yaitu perbedaan nilai rataan populasi di mana sampel diambil. Pada kasus data bivariat dan multivariat akan diterapkan korelasi antar variabel.

Pada variabel prediktor akan terdiri dari tiga kasus yaitu satu variabel (univariat), dua variabel (bivariat), dan tiga variabel prediktor (multivariat). Perbedaan nilai rataan akan dikaji untuk kasus satu variabel prediktor, sementara kasus dengan dua dan tiga variabel prediktor diterapkan agar simulasi dapat melibatkan struktur korelasi antar variabel pada nilai rataan yang juga berbeda. Untuk perbedaan nilai rataan populasi pada kasus satu variabel prediktor (univariat) data yang dibangkitkan berdistribusi normal dengan uraian pada Tabel 1.

Tabel 1. Ketentuan untuk Data Univariat.

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

Untuk kedua pengaturan pada Tabel 1 data yang dibangkitkan akan berdistribusi normal dengan variansi yang sama.

Untuk kasus bivariat maupun multivariat penerapan pengaturan beda rataan populasi akan diiringi dengan penerapan struktur korelasi. Berikut adalah uraian mengenai pengaturan beda rataan populasi yang diiringi dengan struktur korelasi pada Tabel 2.

Tabel 2. Ketentuan Struktur Rataan untuk Data Bivariat dan 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	$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{pmatrix}2\\0\end{pmatrix}$	$\begin{pmatrix}0\\0\\0\end{pmatrix}$	$\begin{pmatrix}1\\0\\0\end{pmatrix}$
5	-	-	$\begin{pmatrix}0\\0\\0\end{pmatrix}$	$\begin{pmatrix}0\\0\\1\end{pmatrix}$

DOI: https://doi.org/10.24036/ujsds/vol1-iss4/74

ISSN(*Online*): 2985-475X

Pengaturan	Bivariat		Multivariat	
	$\mu^{(1)}$	$\mu^{(2)}$	$\mu^{(1)}$	$\mu^{(2)}$
6	-	-	$\begin{pmatrix}0\\0\\0\end{pmatrix}$	$\begin{pmatrix}2\\2\\2\end{pmatrix}$
7	-	-	$\begin{pmatrix}0\\0\\0\end{pmatrix}$	$\begin{pmatrix}2\\2\\0\end{pmatrix}$
8	-	-	$\begin{pmatrix}0\\0\\0\end{pmatrix}$	$\begin{pmatrix}2\\0\\2\end{pmatrix}$
9	-	-	$\begin{pmatrix}0\\0\\0\end{pmatrix}$	$\begin{pmatrix}2\\0\\0\end{pmatrix}$
10	-	-	$\begin{pmatrix}0\\0\\0\end{pmatrix}$	$\begin{pmatrix}0\\0\\2\end{pmatrix}$

Dari uraian Tabel 2 pada kasus 1 untuk data bivariat dinamakan variabel *relevant*, hal ini karena kasus satu variabel memuat informasi tentang perbedaan kelas pada kedua variabel penjelas. Sedangkan pengaturan 2 hanya variabel pertama yang memuat informasi tentang perbedaan kelas sementara variabel kedua tidak yang disebut variabel *irrelevant*. Hal serupa juga terdapat pada kasus data multivariat di mana terdapat variabel *relevant* juga variabel *irrelevant*. Untuk pengaturan struktur korelasi pada kasus data bivariat akan dijelaskan pada Tabel 3.

Tabel 3. Ketentuan Struktur Korelasi Kasus 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	$\begin{bmatrix}1 & 0.9\\0.9 & 1\end{bmatrix}$

Tabel 3 menjelaskan tiga struktur korelasi di mana 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 *relevant* dan antara variabel *relevant* dengan variabel *irrelevant* dengan mengombinasikan penerapan pengaturan beda rataan populasi dan struktur korelasi antar variabel. Hal serupa juga dikaji pada kasus multivariat dengan struktur korelasi yang akan dijelaskan dengan Tabel 4.

Tabel 4. Ketentuan Struktur Korelasi Kasus Multivariat

Pengaturan	Struktur korelasi
1	$\begin{bmatrix}1 & 0 & 0\\0 & 1 & 0\\0 & 0 & 1\end{bmatrix}$
2	$\begin{bmatrix}1 & 0.5 & 0\\0.5 & 1 & 0\\0 & 0 & 1\end{bmatrix}$
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	$\begin{bmatrix}1 & 0.9 & 0.9\\0.9 & 1 & 0.9\\0.9 & 0.9 & 1\end{bmatrix}$

Karena penelitian ini mengkaji untuk kasus data seimbang maka variabel respon yang terdiri dari dua kelas dengan nilai 1 dan 2 akan dibangkitkan dengan proporsi 50:50.

DOI: https://doi.org/10.24036/ujsds/vol1-iss4/74

D. Langkah Analisis

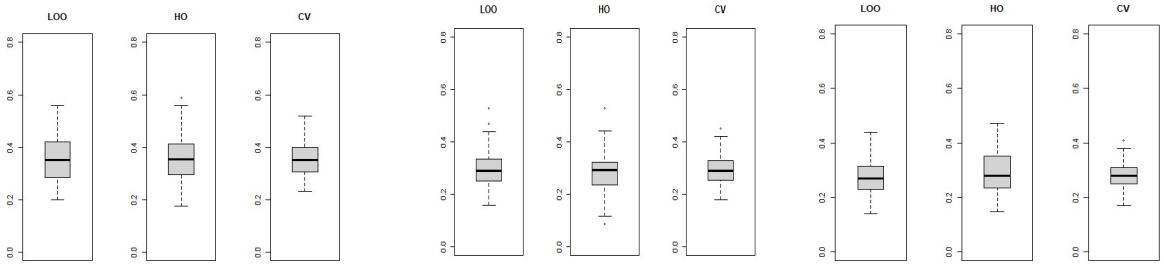
Analisis pada penelitian ini dilakukan menggunakan software r-studio dengan langkah-langkah sebagai berikut:

1. Membangkitan data acak berdistribusi normal dengan *function data_generating* yang terdiri dari data univariat, bivariat, dan multivariat dengan ketentuan rataan, korelasi, dan simpangan baku yang telah ditampilkan diatas.
2. Membagi data yang telah dibangkitakan menjadi data training dan data testing menggunakan ketiga metode cross validation yaitu hold out, LOOCV, dan k-folds cross validation.
3. Membentuk model menggunakan data training lalu model diuji dengan menggunakan data testing.
4. Melakukan pengulangan proses 1 sampai 3 sebanyak 100 kali pengulangan agara hasil yang diperoleh lebih akurat.
5. Melakukan perbandingan hasil yang diperoleh dari masing masing prediksi galat yang berupa boxplot.
6. Menarik kesimpulan dari hasil perbandingan yang telah dilakukan.

III. HASIL DAN PEMBAHASAN

A. Hasil

Berdasarkan tujuan penelitian ini yaitu membandingkan LOO, HO, dan *k-folds* dalam memprediksi laju galat pada algoritma C4.5. Pada penelitian ini model dibentuk dengan menggunakan algoritma C4.5 menggunakan data *training.* Hasil yang diperoleh dari penelitian ini akan ditampilkan dalam bentuk *boxplot* yang dibentuk dari galat yang diperoleh masing masing metode dari 100 kali pengulangan, perbandingan ketiga metode tersebut *cross validation* akan dilakukan dengan melihat nilai *Inter Quartil Range* (IQR). Metode prediksi laju galat yang cocok akan menghasilkan nilai IQR yang lebih kecil diantara ketiga metode. Metode prediksi laju galat dengan IQR terendah akan menjadi metode yang paling cocok diterapkan pada algoritma C4.5. Hasil *boxplot* perbandingan metode prediksi laju galat yang dieroleh ditampilkan pada Gambar 1.



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

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

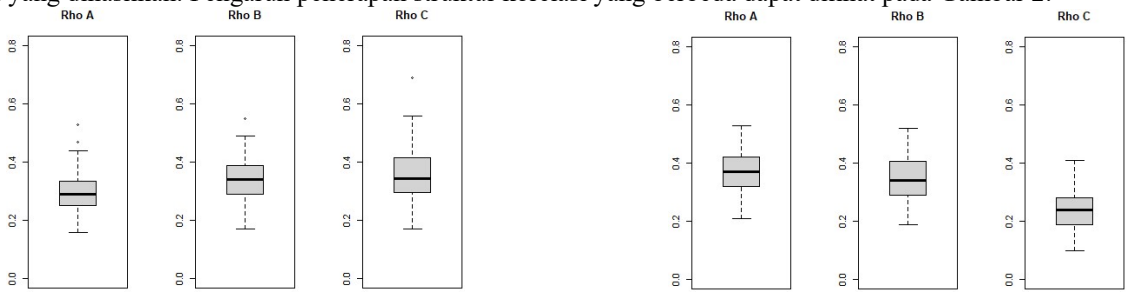
Gambar 1 menampilkan hasil prediksi laju galat masing masing metode untuk data univariat, bivariat, dan multivaiat. Metode *k-folds cross validation* memiliki nilai prediksi laju galat yang kecil untuk setiap jenis data. Pada data univariat metode LOOCV dan HO menghasilkan prediksi laju galat yang cenderung sama sedangkan *k-folds cross validation* menghasilkan *boxplot* dengan IQR yang lebih kecil dibanding LOOCV dan HO. Pada data bivariat *k-folds* juga menghasilkan prediksi laju galat yang paling kecil diikuti LOOCV dengan IQR yang lebih kecil daripada HO yang memiliki IQR terbesar. Pada data multivariat *k-folds* juga menghasilkan IQR yang terkecil dibandingkan metode LOO dan HO. Dari tiga kasus data metode *kfolds* selalu menghasilkan IQR terkecil, hal ini berarti variansi *error rate* yang dihasilkan kecil dan stabil serta terkonsentrasi disekitar median. Dapat disimpulkan bahwa metode k-fold *cross validation* merupakan metode yang paling cocok diterapkan pada algoritma C4.5 dengan data seimbang. Algoritma C4.5 merupakan metode utama yang digunakan untuk membangun model atau pohon keputusan. Model yang dihasilkan tersebut akan diuji laju galat nya dengan menggunakan ketiga metode prediksi galat *cross validation.*

Pengaturan nilai rataan pada data bivariat dan multivariat menghasilkan dua jenis variabel yaitu variabel *relevant* yang menjelaskan pebedaan informasi antara dua kelas data dan variabel *irrelevant* yang tidak mengandung perbedaan informasi pada dua kelas data. Struktur korelasi berbeda yaitu tanpa korelasi (Rho A), korelasi sedang (Rho B), dan korelasi tinggi (Rho C) yang diterapkan antara variabel *relevant* dengan variabel *relevant* dan variabel

DOI: https://doi.org/10.24036/ujsds/vol1-iss4/74

ISSN(*Online*): 2985-475X

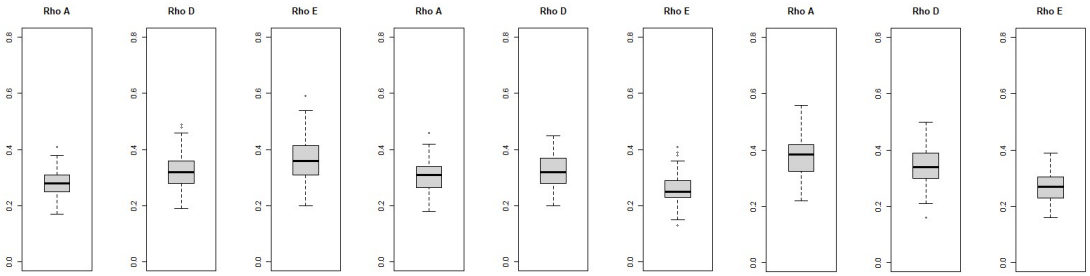
relevant dengan *irrelevant* untuk data bivariat serta penambahan korelasi sedang (Rho D) dan korelasi tinggi (Rho E) antara tiga variabel untuk data multivariat yang dibangkitkan dapat dilihat pengaruhnya terhadap laju prediksi laju galat yang dihasilkan. Pengaruh penerapan struktur korelasi yang berbeda dapat dilihat pada Gambar 2.



(a) **(b)**

Gambar 2. Prediksi laju galat metode *k-folds* dengan korelasi berbeda untuk (a) kasus 1 variabel *relevant* dan (b) kasus 2 variabel *irrelevant* untuk data bivariat.

Gambar 2 menunjukkan pengaruh penambahan beda struktur korelasi pada variabel *relevant* dengan variabel *relevant* dan variabel *relevant* dengan *irrelevant*. Ketika korelasi ditambahkan pada dua variabel *relevant* yaitu pada pengaturan beda rataan 1 pada tabel 2 data bivariat laju prediksi laju galat yang dihasilkan akan bertambah besar dapat dilihat dari boxplot (a) pada Rho A atau tidak ada korelasi galat yang dihasilkan kecil ketika korelasi ditambahkan pada Rho B dengan korelasi sedang prediksi laju galat yang dihasilkan jadi lebih besar, saat korelasi ditambah menjadi korelasi tinggi dilambangkan dengan Rho C prediksi laju galat yang dihasilkan menjadi lebih besar dari kedua korelasi sebelumnya. Sedangkan untuk pengaturan beda korelasi yang ditambahkan antara variabel *relevant* dengan variabel *irrelevant* tidak memberi banyak pengaruh pada hasil prediksi laju galatnya tetapi untuk korelasi tinggi memiliki hasil prediksi laju galat yang sedikit lebih kecil dibandingkan struktur tanpa korelasi maupun korelasi sedang. Pada kasus bivariat nilai korelasi yang ditambahkan memberikan dampak yang berbeda antara dua variabel *relevant* dan antara variabel *relevant* dengan variabel *irrelevant.* Korelasi yang diberikan antara dua variabel menghasilkan IQR yang semakin besar seiring semakin besar korelasi yang ditambahkan sementara ketika pada data bivariat dengan salah satu variabel *irrelevant* korelasi yang debrikan menghasilkan IQR yang semakin kecil ketika nilai korelasi yang ditambahkan semakin besar. Pengaturan beda struktur korelasi ini juga diterapkan pada kasus multivariat yang ditampilkan Gambar 3.



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

Gambar 3. Perbandingan prediksi laju galat data multivariat berdasarkan korelasi pada metode *k-folds* untuk (a) 3 variabel *relevant*, (b) 2 variabel *relevant* dengan 1 *irrelevant*, dan (c) 1 variabel *relevant* dengan 2 *irrelevant* .

Gambar 3 menampilkan *boxplot* yang menampilkan pengaruh penambahan beda struktur korelasi yang diterapakan antara 3 variabel *relevant*, 2 variabel *relevant* dengan 1 *irrelevant*, dan 1 variabel *relevant* dengan 2 *irrelevant* terhadap laju galatnya pada metode *k-folds*. Korelasi yang ditambahkan ditunjukan Rho A, Rho D, dan Rho E yang merupakan beda struktur korelasi tanpa korelasi, korelasi sedang, dan korelasi tinggi berturut-turut antara 3 variabel. Penambahan korelasi pada 3 variabel *relevant* akan mempengaruhi prediksi laju galat dimana semakin tinggi korelasi yang diberikan semakin besar hasil prediksi laju galatnya yang ditunjukkan Rho A, D, dan E. Sementara

DOI: https://doi.org/10.24036/ujsds/vol1-iss4/74

penambahan korelasi untuk 2 variabel *relevant* dengan 1 *irrelevant* menghasilkan hasil prediksi laju galat yang semakin kecil dapat dilihat pada Rho D dan E pada *boxplot* yang menghasilkan IQR yang lebih kecil, hal ini juga berlaku pada penambahan korelasi antara 1 variabel *relevant* dengan 2 *irrelevant*. Semakin besar korelasi yang ditambahkan nilai laju galatnya menjadi semakin kecil seperti terlihat pada Gambar 3. Korelasi ada kasus multivariat memeberikan hasil yang sama dengan kasus bivariat dimana ketika korelasi ditambahkan antara variabel *relevant* IQR yang dihasilkan meningkat dan ketika korelasi ditambahkan pada antara variabel dengan salah satu variabel adalah variabel *irrelevant* IQR dari *error rate* yang dihasilkan semakin kecil seiring penambahan korelasinya.

B. Pembahasan

Penelitian ini menghasilkan metode *k-folds* dengan 10-*folds* sebagai metode terbaik dengan variansi terkecil yang dilihat melalui IQR dari *boxplot* yang dihasilkan. Dari tiga kasus data yang dibangkitkan pada penelitian ini dengan struktur rataan yang berbeda metode *k-folds* dengan 10-*folds* hampir selalu menghasilkan variansi *error rate* terkecil yang dilihat dari nilai IQR yang dihasilkan selalu lebih kecil dari metode LOO dan Ho. Secara keseluruhan median dan variansi galat yang dihasilkan metode HO cenderung tidak stabil dan lebih besar dibandingkan LOOCV dan *k-folds*. Hal ini dikarenakan dalam pembagian data *training* dan *testing* HO menggunakan lebih sedikit data untuk training dibandingkan LOOCV dan *k-folds* dengan hanya 2/3 data yang digunakan sebagai *training*. Hal ini menyebabkan model dibangun menggunakan lebih sedikit data dibandingkan LOOCV dan *k-folds*.

Penambahan korelasi yang diterapkan pada kasus bivariat dan multivariat pada penelitian ini memberikan dampak yang berbeda. Dampak yang dihasilkan ketika korelasi yang diterapkan semakin besar berbeda ketika data terdiri dari variabel-variabel *relevant* (terdapat informasi perbedaan kelas) yang menghasilkan variansi *error rate* yang semakin besar sementara ketika data terdiri dari variabel *relevant* dan *irrelevant* (tidak terdapat informasi perbedaan kelas) variansi *error rate* yang dihasilkan justru semakin kecil. Hal ini dimungkinkan karena ketika variabel korelasi ditambahkan pada dua variabel *relevant* akan meningkatkan kemungkinan variabel tersebut terpilih sebagai atribut sehingga model akan terbentuk dari variabel yang mengandung informasi yang hampir sama. Hal ini akan mengkatkan kemungkinan model salah dalam melakukan prediksi data atau *error rate* yang dihasilkan meningkat.

IV. KESIMPULAN

Berdasarkan perbandingan yang dilakukan terhadap tiga metode prediksi laju galat yang diterapkan untuk melihat akurasi model yang dihasilkan algoritma C4.5, metode prediksi laju galat *k-folds cross validation* merupakan metode prediksi laju galat yang paling cocok diterapkan pada algoritma C4.5. Penambahan korelasi pada variabel *relevant* dan *irrelevant* juga memberikan dampak terhadap hasil prediksi laju galat yang dihasilkan, penambahan korelasi memberikan dampak yang sama untuk data bivariat dan multivariat. Semakin besar struktur korelasi yang diberikan antara dua variabel *relevant* mengakibatkan kenaikan nilai prediksi laju galatnya, sementara pemberian korelasi antara variabel *relevant* dengan *irrelevant* memberi dampak pada penurunan nilai prediksi laju galat.

Berdasarkan hasil penelitian yang telah diapaparkan direkomendasikan penggunaan metode prediksi laju galat *k-folds cross validation* dalam memprediksi laju galat pada algoritma C4.5. Peneliti juga merekomendasikan untuk melakukan penanganan korelasi terlebih dahulu sebelum melakukan klasifikasi pada data dengan kondisi korelasi yang tinggi antara variabel, khususnya untuk data yang terdiri dari dua atau lebih variabel *relevant* dengan korelasi yang tinggi. Pada penelitian selanjutnya dapat membandingkan metode *k-folds cross validation* dengan jumlah *folds* yang berbeda untuk melihat *folds* yang cocok diterapkan dalam prediksi laju galat algoritma C4.5 maupun algoritma lain dengan data asli.

DAFTAR PUSTAKA

Behera, H. S., Jain, L. C., Mandal, J. K., & Mohapatra, D. P. (2015), Computational Intelligence in Data Mining - Volume 1: Proceedings of the International Conference on CIDM, 20-21 December 2014 (1st ed. 2015) Eds: Robert J. Howlett, KES International, Shoreham-by-Sea, UK.

Berrar, D. (2019), "Cross-Validation", Dalam Encyclopedia of Bioinformatics and Computational Biology, Elsevier, hal. 542–545.

Braga-Neto, U. de M., & Dougherty, E. R. (2015), Error estimation for pattern recognition, IEEE Press ; Wiley.

Dai, W., & Ji, W. (2014), "A MapReduce Implementation of C4.5 Decision Tree Algorithm", International Journal of Database Theory and Application, Vol. 7, No. 1, hal. 49–60.

DOI: https://doi.org/10.24036/ujsds/vol1-iss4/74

Dougherty, E. R., & Braga-Neto, U. (2006), "Epistemology of Computational Biology: Mathematical Models and Experimental Prediction as the Basis of Their Validity", Journal of Biological Systems, Vol. 14, No. 01, hal. 65–90.

Dougherty, E., Sima, C., Hua, Hanczar, B., & Braga-Neto, U. (2010), "Performance of Error Estimators for Classification", Current Bioinformatics, Vol. 5, No. 1, hal. 53–67.

García-Laencina, P. J., Abreu, P. H., Abreu, M. H., & Afonoso, N. (2015), "Missing data imputation on the 5-year survival prediction of breast cancer patients with unknown discrete values", Computers in Biology and Medicine, vol. 59, hal. 125–133.

Hastie, T., Tibshirani, R., & Friedman, J. (2013), The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer.

Hssina, B., Merbouha, A., Ezzikouri, H., & Erritali, M. (2014), "A comparative study of decision tree ID3 and C4.5", International Journal of Advanced Computer Science and Applications, Vol. 4, No. 2.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013), An Introduction to Statistical Learning, Vol. 103, Springer New York.

Kohavi, R. (1995), "A study of cross-validation and bootstrap for accuracy estimation and model selection" International Joint Conference on Artificial Intelligence, Vol. 14, No. 2, hal. 1137–1145.

Liu, H., & Gegov, A. (2016), "Induction of Modular Classification Rules by Information Entropy Based Rule Generation", dalam Innovative Issues in Intelligent Systems, eds. V. Sgurev, R. R. Yager, J. Kacprzyk, & V. Jotsov, Springer International Publishing, Vol. 623, hal. 217–230.

Lu, Z., Wu, X., & Bongard, J. C. (2015), "Active Learning through Adaptive Heterogeneous Ensembling", IEEE Transactions on Knowledge and Data Engineering, Vol. 27, No. 2, hal 368–381.

Mansour, Y., McAllester, D. (2002), "Boosting using branching programs", Journal ofComputer and System Sciences, Vol. 64, hal. 103–112.

Quinlan, J. R. (1996), "Improved Use of Continuous Attributes in C4.5". Journal of Artificial Intelligence Research, Vol. 4, hal. 77–90.

Refaeilzadeh, P., Tang, L., & Liu, H. (2016), "Cross-Validation", dalam Encyclopedia of Database Systems, eds. L. Liu & M. T. ÖzsuSpringer New York, hal. 1-7.

Tougui, I., Jilbab, A., & El Mhamdi, J. (2021), "Impact of the choice of cross-validation techniques on the results of machine learning-based diagnostic applications", Healthcare informatics research, Vol. 27, No. 3, hal. 189-199.

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

DOI: https://doi.org/10.24036/ujsds/vol1-iss4/74