UNP JOURNAL OF STATISTICS AND DATA SCIENCE
Volume 1 Nomor 5 Tahun 2023 Halaman 456-463.

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

Comparison of Error Prediction Methods in Classification Modeling with Chi-Squared Automatic Interaction Detection (CHAID) Methods for Balanced Data

Findri Wara Putri, Dodi Vionanda*, Atus Amadi Putra, Fadhilah Fitri
Departemen Statistika, Universitas Negeri Padang, Padang, Indonesia
*Corresponding author: dodi_vionanda@fmipa.unp.ac.id

Submitted : 07 Oktober 2023
Revised : 24 Oktober 2023
Accepted : 26 Oktober 2023

ABSTRACT

CHAID (Chi-Squares Automatic Interaction Detection) is one of the classification algorithms in the decision tree method. The classification result are displayed in the form of a tree diagram model. After the model is formed, it is necessary to calculate the accuracy of the model. The goal is to see the performance of the model. The accuracy of this model can be determined by calculating the level of prediction error in the model. The error rate prediction method works by dividing data into training data and testing data. There are three methods in the error rate prediction method, such as Leave one out cross validation (LOOCV), Hold out, and k-fold cross validation. These methods have different performance in dividing data into training data and test data. The aim of this research is to identify the performance of each error rate prediction method for balanced data and to find out which error rate prediction method is most suitable to be applied to the CHAID method. This comparison is carried out by considering several factors, namely the marginal probability matrix and different correlations. The comparison results will be observed using a boxplot by looking at the median error rate and lowest variance. This research found that k-fold cross validation is the most suitable error rate prediction method applied to the CHAID method for balanced data.

Keywords: *CHAID, Hold Out, K-Fold Cross Validation, Leave One Out*

I. PENDAHULUAN

Metode CHAID merupakan metode eksplorasi untuk mengklasifikasi data dengan cara membangun pohon keputusan. Pohon keputusan pada metode CHAID dapat memberikan informasi berupa variabel independen yang berpengaruh secara signifikan terhadap variabel dependen (DuToit, 2012). Analisis ini menghasilkan output berupa diagram pohon dan perlu dinilai akurasinya. Salah satu metode yang dapat digunakan untuk menilai akurasi yaitu metode prediksi laju galat. Dalam prediksi laju galat, nilai *error* terkecil dapat digunakan sebagai ukuran akurasi model yang baik.

Metode prediksi laju galat yang paling sering digunakan adalah *Cross Validation* (CV). CV menghasilkan data yang menunjukkan keakuratan atau kesesuaian pengukuran dengan data nyata yang memungkinkan pengujian menggunakan data *testing* dan data *training*. Pembentukan data *training* dan data *testing* dibagi kedalam tiga metode yaitu *Leave One Out* (LOO), *Hold Out*, *k-fold Cross Validation*.

Pada LOO, setiap pengamatan berperan sebagai data *training* dan data *testing* (Wong, 2015). Kelebihan LOO yaitu sistem beraturan pada analisis ini dan tidak adanya pengacakan untuk setiap pengambilan data *testing* dan *training* yang menyebabkan hasil akurasi rataan yang selalu konstan (Wong, 2015). Kelemahan metode ini yaitu analisis yang memakan banyak waktu dan biaya jika digunakan pada data berukuran besar dan memiliki hasil perkiraan nilai varians yang sangat besar (Efron, 1983).

Metode *Hold Out*, data dibagi dua secara acak dimana dua pertiga data dijadikan sebagai data *training* dan satu pertiga lainnya digunakan sebagai data *testing.* Pembagian pada metode *Hold Out* selalu jumlah data *training* lebih besar daripada jumlah data *testing*, dengan tujuan agar dapat mempertahankan rasio antarkelas (James, dkk, 2013). Kelebihan metode ini yaitu pekerjaan yang dilakukan sangat ringkas apabila dibandingkan dengan metode lainnya dikarenakan data hanya perlu dipisah menjadi data *training* dan data *testing*. Akan tetapi hal tersebut dapat menjadi

DOI: https://doi.org/10.24036/ujsds/vol1-iss5/116

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

kekurangan dari analisis ini yaitu hasil akhir yang diperoleh sangat bergantung pada pembagian data yang dilakukan (Kohavi, 1995).

K-fold cross validation mengelompokkan data secara acak kedalam k kelompok kemudian membagi kelompok menjadi data *training* dan data *testing* dan melakukan perulangan sebanyak k kali dengan meninggalkan satu kelompok sebagai data *testing* di setiap perulangan. Pembagian kelompok pada *k-fold* dilakukan pada masing-masing kelas dengan membagi data hampir sama rata setiap kelompoknya. Jumlah k yang biasa digunakan adalah sebanyak 5 dan 10 kelompok (James, dkk, 2013).

Tujuan penelitian ini adalah untuk mengidentifikasi kinerja dari masing-masing metode prediksi laju galat untuk data seimbang dan mengetahui metode prediksi laju galat yang paling sesuai diterapkan pada metode CHAID. Manfaat penelitian ini adalah sebagai tambahan ilmu pengetahuan terhadap perbandingan metode prediksi laju galat dan metode CHAID.

II. METODE PENELITIAN

A. *Chi-Squared Automatic Interaction Detection* (CHAID)

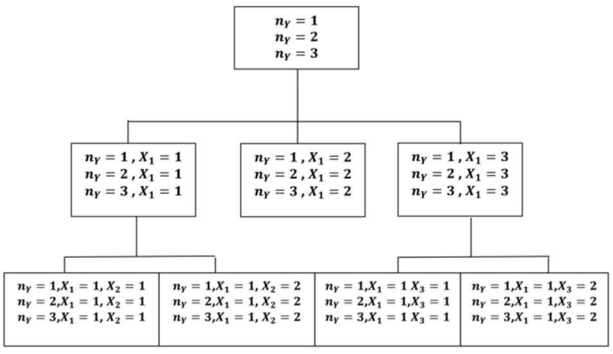
CHAID merupakan teknik iteratif yang menguji satu-persatu variabel independen yang digunakan dalam klasifikasi, dan menyusunnya berdasarkan tingkat signifikansi statistik uji *chi-square* terhadap variabel dependen (Gallagher, 2000). Metode CHAID digunakan untuk membentuk klasifikasi yang membagi sebuah sampel menjadi dua kelompok atau lebih yang berbeda-beda berdasarkan pada kriteria tertentu. Metode ini merupakan metode pengklasifikasian data yang berupa data kategorik dimana metode ini bertujuan untuk membagi kumpulan data yang menjadi sub kelompok berdasarkan variabel dependen (Lehmann dan Eherler, 2001). Statistik uji:

$$x^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{ij} - E_{ij})^2}{E_{ij}} \quad (1)$$

Statistik uji *chi-square* digunakan dalam dua cara pada analisis CHAID. Pertama untuk menentukan apakah kategori-kategori dalam sebuah variabel independen bersifat sama dan bisa digabungkan menjadi satu. Kedua, ketika semua variabel independen sudah diringkas menjadi bentuk signifikan dan tidak mungkin digabung lagi, maka statistik uji-*chi-square* digunakan untuk menentukan variabel independen mana yang paling signifikan untuk membagi kategori-kategori dalam variabel dependen (Gallagher, 2000).

Menurut Bagozzi (1994), langkah-langkah analisis CHAID secara garis besar dibagi menjadi tiga tahap, yaitu: Penggabungan (Merging), dalam tahap pertama akan diperiksa signifikansi dari masing-masing kategori variabel independen terhadap variabel dependen. Pemisahan (Splitting), dalam tahap kedua ini memilih variabel independen yang mana akan digunakan sebagai split node (pemisah simpul) yang terbaik, pemilihan dikerjakan dengan membandingkan p-value (dari tahap merging) pada setiap variabel independen. Penghentian (Stopping), pada tahap ketiga ini dilakukan jika proses pertumbuhan pohon harus dihentikan jika tidak ada lagi variabel independen yang signifikan menunjukkan perbedaan terhadap variabel dependen.

CHAID akan menghasilkan sebuah diagram pohon klasifikasi yang menggambarkan pembentukan segmen. Diagram CHAID terdiri dari batang pohon (*tree trunk*) dengan membagi menjadi lebih kecil berupa cabang-cabang (*brances*) yang dimulai dari simpul akar melalui tiga tahap pada setiap simpul yang terbentuk dan secara berulang. Berikut gambar diagram pohon CHAID secara umum (Lehmann dan Eherler, 2001).



Gambar 1. Diagram Pohon Klasifikasi CHAID

DOI: https://doi.org/10.24036/ujsds/vol1-iss5/116

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

Pada pohon keputusan CHAID terdapat nilai sebenarnya (yi) dan nilai yang diharapkan ($\hat{y}_i$). Nilai sebenarnya adalah nilai aktual dari variabel dependen yang ingin diprediksi oleh model, sedangkan nilai yang diharapkan merupakan nilai yang diprediksi oleh model untuk variabel dependen berdasarkan kondisi pada simpul pohon keputusan.

B. Prediksi Laju Galat

Galat atau *error* merupakan selisih antara nilai yang diharapkan dengan nilai sebenarnya (Purwati&Erawati, 2020). Pada klasifikasi untuk mengukur galat atau *error* menggunakan indikator variabel dengan rumus sebagai berikut (James, dkk, 2013).

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

dimana,

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

Prediksi laju galat adalah suatu metode tambahan dari teknik data mining yang bertujuan utuk memperoleh hasil akurasi yang maksimal. Prediksi laju galat terhadap model adalah mengukur kinerja model dengan menghitung segala bentuk tingkat kesalahan prediksi pada model.

Salah satu metode yang dapat digunakan untuk memprediksi laju galat adalah metode *cross validation* (CV). CV merupakan teknik penguji keefektifan dari model yang dibentuk dengan melakukan penyusunan ulang (*resampling*) pada data untuk membaginya menjadi dua bagian yaitu data *training* dan data *testing*. Data *training* akan dipakai untuk melatih model sehingga dapat memahami pola data sedangkan untuk melakukan validasi terhadap model tersebut, akan digunakan data *testing* sebagai pengujinya. CV memiliki beberapa algoritma pembelajaran dan yang paling umum digunakan adalah *leave one out, hold out, dan k-fold* (James, dkk, 2013).

1. *Leave One Out*

Leave One Out (LOO) merupakan salah satu metode validasi model terbaik dengan cara membagi data menjadi dua bagian yaitu data *training* dan data *testing* berdasarkan jumlah pengamatan. Dimana untuk data *training* sebanyak n-1 pengamatan dan sisa 1 pengamatan digunakan sebagai data *testing*. Proses perhitungan dapat diulang sebanyak k kali dimana selalu meninggalkan satu pengamatan untuk data *testing*. Pemilihan model terbaik dilihat dari waktu perhitungan, semakin cepat waktu perhitungan maka model dikatakan baik walaupun hasil yang didapatkan tidak akurat (James, dkk, 2013). Namun pembagian ini memiliki kelemahan dalam segi komputasi, data yang digunakan berjumlah sangat besar maka akan menimbulkan permasalahan dalam komputasi. Tingkat akurasi yang diperoleh LOO hampir tidak bias tetapi variansi yang dihasilkan tinggi (Rafaeilzadeh, dkk, 2016). Untuk menghitung nilai error pada algoritma LOO dapat menggunakan rumus sebagai berikut.

$$\hat{E}^{LOOCV} = \frac{1}{n}\sum_{i=1}^{n} I(y_{(i)} \neq \hat{y}_{(i)}) \quad (2)$$

2. *Hold Out*

Hold out merupakan metode *cross validation* yang paling sederhana. Metode ini membagi data menjadi dua bagian yaitu data *training* dan data *testing* (James, dkk, 2013). Metode ini membagi data menjadi 2/3 data sebagai data *training* dan sisa 1/3 data menjadi data *testing*. Pada metode ini data *testing* diambil dan tidak dimasukkan selama proses *training*. *Hold out* menghindari tumpang tindih antara data *training* dan data *testing*, menghasilkan perkiraan kinerja algoritma secara lebih akurat.

Perhitungan prediksi galat dengan metode *hold out* dapat dilakukan dengan rumus.

$$\hat{E}^{HO} = \frac{1}{n_{uji}}\sum_{i=1}^{n_{uji}} I(y_i \neq \hat{y}_i) \quad (3)$$

3. *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 rumus berikut.

DOI: https://doi.org/10.24036/ujsds/vol1-iss5/116

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

$$\hat{E}^{CV} = \frac{1}{k}\sum_{i=1}^{k} I(y_i \neq \hat{y}_i) \quad (4)$$

C. Jenis Penelitian dan Sumber Data

Jenis penelitian yang digunakan adalah penelitian eksperimen. Penelitian eksperimen bertujuan untuk melakukan perbandingan suatu akibat perlakuan tertentu dengan suatu perlakuan lain yang berbeda atau menguji pengaruh hubungan sebab akibat antara suatu variabel dengan variabel lainnya. Penelitian ini bertujuan untuk membandingkan algoritma prediksi laju galat dalam pemodelan klasfikasi dengan metode CHAID untuk data seimbang. Algoritma prediksi laju galat yang dibandingkan yaitu LOOCV, *hold out*, dan *k-fold cross validation*. Hal yang dilihat adalah nilai median dan nilai variasi atau keragaman dari nilai galat yang didapatkan.

Penelitian ini menggunakan jenis data simulasi yang diperoleh dari proses bangkitan data pada *software RStudio*. Data simulasi yang dibangkitkan terdiri dari variabel dependen (Y) dan variabel independen (X). Kedua variabel bersifat kategorik dengan skala data rasio atau interval. Jumlah pengamatan yang dibangkitkan adalah sebanyak 100 sampel. Variabel Y berbentuk ordinal dengan nilai 1 dan 2, sedangkan variabel X dibangkitkan berdasarkan jumlah variabelnya yang terdiri dari satu variabel (univariat), dua variabel (bivariat) dan tiga variabel (multivariat).

Perbedaan matriks peluang marginal kasus univariat dibangkitkan dari distribusi binomial yang dijabarkan pada Tabel 1 berikut.

Tabel 1. Ketentuan Data Bangkitan untuk Data Univariat

Jumlah Variabel	Matriks Peluang Marginal	
	$m^{(1)}$	$m^{(2)}$
Univariat	0.9	0.1

Tabel 1 merupakan ketentuan aturan data bangkitan untuk jenis data univariat. Sementara itu untuk kasus data bivariat dan multivariat dibangkitkan berdasarkan matriks peluang marginal dengan ketentuan pada Tabel 2 berikut.

Tabel 2. Ketentuan Data Bangkitan untuk Data Bivariat dan Multivariat

Jumlah Variabel	Matriks Peluang Marginal
Bivariat	$\begin{bmatrix} 0.9 & 0.8 \\ 0.1 & 0.2 \end{bmatrix}$
Multivariat	$\begin{bmatrix} 0.15 & 0.20 & 0.05 \\ 0.85 & 0.80 & 0.25 \\ 0 & 0 & 0.70 \end{bmatrix}$

Untuk kasus bivariat maupun multivariat penerapan pengaturan beda matriks peluang marginal akan diiringi dengan penerapan struktur korelasi. Berikut adalah uraian mengenai pengaturan struktur korelasi pada kasus data bivariat akan dijelaskan pada Tabel 3.

Tabel 3. Ketentuan Struktur Korelasi Kasus Univariat

Pengaturan	Struktur Korelasi
1	$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$
2	$\begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$
3	$\begin{bmatrix} 1 & 0.6 \\ 0.6 & 1 \end{bmatrix}$

Tabel 3 menjelaskan 3 struktur korelasi 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 dengan mengombinasikan penerapan pengaturan beda matriks peluang marginal dan struktur korelasi antar variabel. Hal serupa juga dikaji pada kasus multivariat dengan struktur korelasi yang akan dijelaskan dengan Tabel 4.

DOI: https://doi.org/10.24036/ujsds/vol1-iss5/116

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

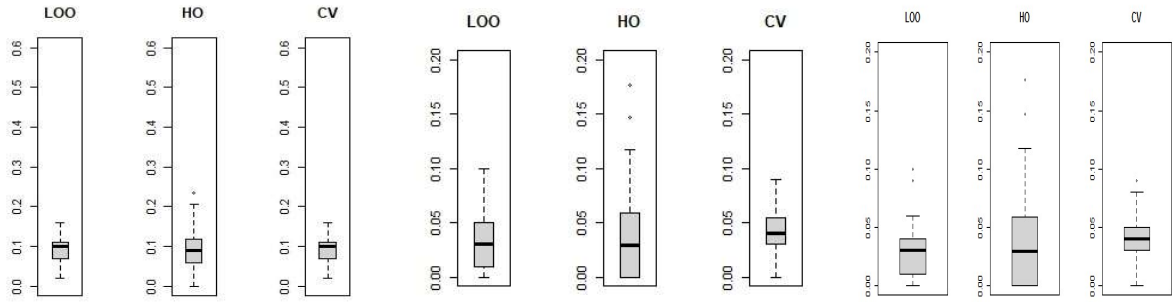
Tabel 4. Ketentuan Struktur Korelasi Kasus Multivariat

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	$\begin{bmatrix} 1 & 0.6 & 0 \\ 0.6 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$
4	$\begin{bmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{bmatrix}$
5	$\begin{bmatrix} 1 & 0.6 & 0.6 \\ 0.6 & 1 & 0.6 \\ 0.6 & 0.6 & 1 \end{bmatrix}$

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

III. HASIL DAN PEMBAHASAN

Data pada penelitian ini menggunakan variasi yang berbeda. Variasi yang berbeda dapat memberikan pengaruh yang berbeda-beda untuk prediksi laju galat yang dihasilkan. Berdasarkan tujuan penelitian ini yaitu membandingkan LOO, HO, dan *k-folds* dalam memprediksi laju galat pada algoritma CHAID, hasil yang diperoleh dari penelitian ini akan ditampilkan dalam bentuk *boxplot*, perbandingan ketiga metode *cross validation* tersebut akan dilakukan dengan melihat nilai *Inter Quartil Range* (IQR). Metode prediksi galat yang cocok akan menghasilkan nilai IQR yang lebih kecil dari ketiga metode. Metode prediksi laju galat dengan nilai IQR terendah akan menjadi metode yang paling cocok untuk diterapkan pada algotirma CHAID. Berikut hasil *boxplot* perbandingan metode prediksi galat pada Gambar 2.



(a) (b) (c)

Gambar 2. (a) Prediksi laju galat dari data acak univariat tanpa korelasi (b) Prediksi laju galat dari data acak bivariat tanpa korelasi (c) Prediksi laju galat dari data acak multivariat tanpa korelasi

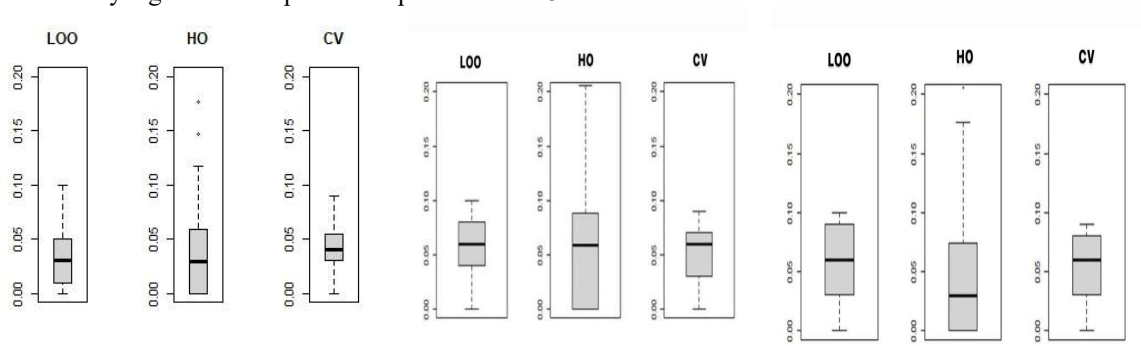
Gambar 2 merupakan hasil dari prediksi laju galat dari tiga jenis data yang berbeda yaitu univariat, bivariat dan multivariat menggunakan pengaturan 1 pada Tabel 3. Pada Gambar 2 tampak algoritma *hold out* memiliki nilai variasi *error rate* paling besar diantara ketiga algoritma prediksi laju galat. Namun, nilai median *error rate* yang dihasilkan cenderung lebih kecil dibandingkan algoritma lainnya. Untuk algoritma LOOCV dan *k-fold* memiliki kinerja yang tampak sepadan dalam memprediksi laju galat dengan nilai variasi yang hampir sama, namun jika diperhatikan dengan lebih teliti variasi *error rate* untuk algoritma *k-fold* lebih kecil dibandingkan LOOCV. Berdasarkan

DOI: https://doi.org/10.24036/ujsds/vol1-iss5/116

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

pemaparan ini, dapat disimpulkan bahwa metode *k-fold cross validation* merupakan metode prediksi laju galat yang lebih baik dan cocok digunakan dalam pemodelan klasifikasi dengan metode CHAID untuk data seimbang.

Perlakuan berbeda yang diterapkan dalam proses pembangkitan data dapat dilihat pengaruhya dalam laju prediksi galat. Pada kasus ini pengaruh penambahan korelasi akan dilihat pada data bivariat. Pengaruh penerapan struktur korelasi yang berbeda dapat dilihat pada Gambar 3.

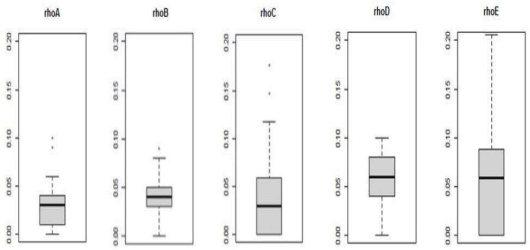


(a) (b) (c)

Gambar 3. Prediksi Galat data Bivariat dengan korelasi berbeda (a) tanpa korelasi (b) korelasi sedang (c) korelasi tinggi

Gambar 3 menunjukkan pengaruh penambahan beda struktur korelasi pada data dengan dua variabel prediktor. Gambar 3(a) merupakan hasil boxplot tanpa korelasi pada data bivariat. Ketika korelasi ditambahkan menjadi korelasi sedang pada Gambar 3(b) laju prediksi galat yang dihasilkan berbeda dibandingkan dengan gambar 3(a), dapat dilihat dari boxplot 3(b) pada LOO, galat yang dihasilkan semakin kecil dibandingkan tanpa korelasi, akan tetapi pada HO prediksi galat yang dihasilkan lebih besar, kemudian pada CV prediksi galat yang dihasilkan menjadi lebih kecil dari boxplot 3(a). Sedangkan untuk korelasi tinggi pada Gambar 3(c) memiliki hasil boxplot yang hampir sama dengan korelasi sedang tetapi sedikit mengalami penurunan pada HO.

Pengaturan beda struktur korelasi ini juga diterapkan pada kasus multivariat. Struktur korelasi berbeda yaitu tanpa korelasi (rhoA), korelasi sedang dua variabel (rhoB), korelasi tinggi dua variabel (rhoC), korelasi sedang tiga variabel (rhoD), dan korelasi tinggi tiga variabel (rhoE) yang ditampilkan pada Gambar 4..



Gambar 4. Perbandingan prediksi galat data multivariat dengan korelasi berbeda

Gambar 4 menampilkan *boxplot* yang menampilkan pengaruh penambahan beda struktur korelasi untuk data multivariat (3 variabel). RhoA, rhoB, dan rhoC merupakan beda struktur korelasi tanpa korelasi, korelasi sedang, dan korelasi tinggi dua variabel secara berurutan, sedangkan rhoD dan rhoE merupakan korelasi sedang dan korelasi tinggi tiga variabel. Penambahan korelasi pada dua variabel akan mempengaruhi prediksi galat dimana semakin besar korelasi diberikan maka semakin besar hasil prediksi galat yang dihasilkan. Penambahan korelasi untuk tiga variabel menghasilkan prediksi galat yang lebih kecil yaitu pada rhoD, akan tetapi pada rhoE nilai prediksi galat kembali meningkat.

DOI: https://doi.org/10.24036/ujsds/vol1-iss5/116

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

IV. KESIMPULAN

Berdasarkan perbandingan yang dilakukan terhadap tiga metode prediksi laju galat yang diterapkan pada algoritma CHAID, metode prediksi galat *k-fold cv* menghasilkan nilai prediksi galat yang lebih kecil dibandingkan metode LOOCV dan HO untuk masing-masing data univariat, bivariat dan multivariat. Nilai IQR terkecil pada *k-fold cross validation* menjadikannya sebagai metode prediksi galat yang paling cocok diterapkan pada CHAID. Penambahan nilai korelasi juga memberikan dampak terhadap hasil prediksi galat yang dihasilkan. Pada korelasi sedang, laju galat mengalami perubahan yang signifikan apabila diperlakukan berbeda berdasarkan nilai korelasi yang digunakan. Secara keseluruhan, median dan variansi galat yang dihasilkan oleh metode HO cenderung tidak stabil dan lebih besar dibandingkan LOO dan *k-fold*. Hal ini dikarenakan dalam pembagian data *training* dan *testing* HO menggunakan lebih sedikit data untuk *training* dibandingkan dua metode lainnya yaitu hanya 2/3 data yang digunakan sebagai data *training*. Hal ini menyebabkan model dibangun menggunakan lebih sedikit data dibandingkan LOO dan *k-fold cv*.

Penelitian ini hanya menerapkan 3 metode prediksi galat yaitu LOOCV, HO, dan k-fold cross validation yang diterapkan pada satu algoritma pohon keputusan CHAID diharapkan peneltian berikutnya dapat membandingkan metode prediksi galat lain yang diterapkan pada algortima pohon keputusan lain. Diharapkan juga penelitian berikutnya dapat menggunakan data asli yang dapat diimplementasi untuk mengatasi permasalahan asli yang terjadi.

DAFTAR PUSTAKA

Agresti, A. (2013). Categorical Data Analysis (Third). John Wiley & Sons. Inc

Ali, A., Shamsuddin, S.M., & Ralescu, A. L. (2013). Classification with class imbalance problem: A review Classification Int. J. Advance Soft Compu. Appl, 5(3).

Bagozzi, R. P. (1994). Advanced Methods of Marketing Research. Blackwell Publisher Ltd, Oxford.

Baron, S., & Phillips, D. (1994). Attitude Survey Data Reduction Using CHAID: An Example in Shopping Centre Market Research. Journal of Marketing Management, 10(1–3), 75–88.

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 .

DuToit, S. H., Steyn, A. G. W., & Stumpf, R. H. (2012). *Graphical exploratory data analysis*. Springer Science & Business Media.

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

Eherler, D., Lehmann, T. (2001). *Responder Profiling with CHAID and Deependency Analysis*.

Gallagher, Cecily A, Howard M onroe, and Joyce L Fish. (2000). "An Iterative Approach to Classification Analysis". *Journal of Applied Statistiks*, 238-230.

Haughton, D., & Oulabi, S. (1997). Direct marketing modeling with CART and CHAID. *Journal of Direct Marketing*, *11*(4), 42-52.

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

Jhonson. R.A., Wichern. D.W. (2007). *Applied Multivariate Statistical Analysis Sixth Edition*. New Jersey:Prentice Hall Interational. Inc.

Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *29*(2), 119-127.

Kohavi, Ron. (1995) .A Study of Prediksi galat and Bootstrap of Accuracy Estimation and Model Selection, *International Joint Conference on Artifical Intelegence.*

DOI: https://doi.org/10.24036/ujsds/vol1-iss5/116

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

Larose, D. T. (2005). *Discovering Knowledge in Data: An Introduction to Data Mining*. John Wiley & Sons. Inc.

Purwati, R., Erawati. 2020. Pengantar Metode Numerik. Jawa Timur: Klik Media.

Prasetyo, E. (2012). Data Mining: Konsep dan Aplikasi menggunakan MATLAB (1st ed.). C.V Andi Offset.

Reddy, U. S., & Somasundaram, A. (2016). Data Imbalance: Effects and Solutions for Classification of Large and Highly Imbalanced Data. Proc. of 1st International Conference on Research in Engineering, Computers and Technology. https://www.researchgate.net/publication/320895020

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

Sachs, L. (2012). *Applied statistics: a handbook of techniques*. Springer Science & Business Media.

Serrano-Cinca, C., & Gutiérrez-Nieto, B. (2016). The Use of Profit Scoring as An Alternative to Credit Scoring Systems in Peer-To-Peer (P2P) Lending. Decision Support Systems, 89, 113–122

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.

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

Wong, T.-T. (2015). Performance Evaluation of Classification Algorithms by K-Fold and Lesve-One-Out Cross Validation. *Pattern Recorgnition*, 1-8

DOI: https://doi.org/10.24036/ujsds/vol1-iss5/116