PERBANDINGAN METODE PREDIKSI GALAT DALAM PEMODELAN KLASIFIKASI DENGAN ALGORITMA C4.5 UNTUK DATA SEIMBANG

SKRIPSI

Diajukan sebagai salah satu persyaratan untuk memperoleh gelar Sarjana statistika

Oleh
Ichlas Djuazva
Nim. 18337035

PROGRAM STUDI SARJANA STATISTIKA
DEPARTEMEN STATISTIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS NEGERI PADANG

2023

PERSETUJUAN SKRIPSI

PERBANDINGAN METODE PREDIKSI GALAT DALAM PEMODELAN KLASIFIKASI DENGAN ALGORITMA C4.5 UNTUK DATA SEIMBANG

Nama	: Ichlas Djuazva
NIM	: 18337035
Program Studi	: S1 Statistika
Departemen	: Statistika
Fakultas	: Matematika dan Ilmu Pengetahuan Alam

Padang, September 2023

Mengetahui:
Ketua Departemen Statistika

Dodi Vionanda, M.Si., Ph.D
NIP. 197806112005011002

Disetujui Oleh:
Pembimbing

Dodi Vionanda, M.Si., Ph.D
NIP. 197806112005011002

PENGESAHAN LULUS UJIAN SKRIPSI

Nama	: Ichlas Djuazva
NIM	: 18337035
Program Studi	: S1 Statistika
Departemen	: Statistika
Fakultas	: Matematika dan Ilmu Pengetahuan Alam

PERBANDINGAN METODE PREDIKSI GALAT DALAM PEMODELAN KLASIFIKASI DENGAN ALGORITMA C4.5 UNTUK DATA SEIMBANG

Dinyatakan lulus setelah dipertahankan di depan Tim Penguji Skripsi
Departemen Statistika Fakultas Matematika dan Ilmu Pengetahuan Alam
Universitas Negeri Padang

Padang, September 2023

Tim Penguji	Nama	Tanda Tangan
Ketua	: Dodi Vionanda, Ph.D	
Anggota	: Dra. Nonong Amalita, M.Si	
Anggota	: Zilrahmi, S.Pd, M.Si	

SURAT PERNYATAAN TIDAK PLAGIAT

Saya yang bertandatangan di bawah ini:

Nama	: Ichlas Djuazva
NIM	: 18337035
Program Studi	: S1 Statistika
Departemen	: Statistika
Fakultas	: Matematika dan Ilmu Pengetahuan Alam

Dengan ini menyatakan bahwa, skripsi saya dengan judul **"Perbandingan Metode Prediksi Galat dalam Pemodelan Klasifikasi dengan Algoritma C4.5 untuk Data Seimbang"** adalah benar merupakan hasil karya saya dan bukan merupakan plagiat dari karya orang lain atau pengutipan dengan cara-cara yang tidak sesuai dengan etika yang berlaku dalam tradisi keilmuan.

Apabila suatu saat terbukti saya melakukan plagiat maka saya bersedia diproses dan menerima sanksi akademis maupun hukum sesuai dengan hukum dan ketentuan yang berlaku, baik di institusi UNP maupun di masyarakat dan negara.

Demikian pernyataan ini saya buat dengan penuh kesadaran rasa tanggung jawab sebagai anggota masyarakat ilmiah.

Diketahui oleh,
Ketua Departemen Statistika,

Dodi Vionanda, Ph.D
NIP. 197806112005011002

Saya yang menyatakan,

Ichlas Djuazva
NIM. 18337035

Perbandingan Metode Prediksi Galat dalam Pemodelan Klasifikasi dengan Algoritma C4.5 untuk Data Seimbang

Ichlas Djuazva

ABSTRAK

Algoritma C4.5 merupakan salah satu algoritma pohon keputusan yang bertujuan membangun model pohon keputusan yang dapat digunakan untuk klasifikasi data. Algoritma C4.5 merupakan pengembangan dari *Iterative Dichotomiser* 3 dengan peningkatan. Model yang dibentuk dengan algoritma perlu diuji akurasinya untuk melihat kinerja dari modelnya. Akurasi model dapat dilihat dengan melakukan prediksi nilai kesalahan atau prediksi galat.Metode prediksi galatyang digunakan adalah metode *Cross Validation* (CV). CV membagi data menjadi data *training* untuk membentuk model dan data *testing* untuk menguji model. CV terdiri dari beberapa metode yaitu *Leave One Out* (LOO), *Hold Out* (HO), dan *k-folds* CV. Tujuan penelitian ini adalah untuk melihat metode prediksi galat mana yang paling cocok digunakan pada algoritma C4.5.

Penelitian ini menggunakan data bangkitan yang berdistribusi normal dengan tiga kasus data yaitu univariat, bivariat, dan multivariat dengan bebarapa kombinasi perbedaan rataan dan korelasi. Korelasi ditambahkan untuk melihat pengaruh terhadap prediksi galat yang dihasilkan. Pada kasus univariat menggunakan 2 struktur rataan yang berbeda, bivariat menggunakan 4 struktur rataan dengan 3 struktur korelasi yang berbeda, dan multivariat menggunakan 10 struktur rataan berbeda dengan 5 struktur korelasi berbeda.

Pada kasus univariat, bivariat, dan multivariat, metode prediksi galat *k-folds* CV merupakan metode prediksi galat yang paling cocok dalam melakukan prediksi laju galat pada algoritma C4.5.

Kata kunci : C4.5, *cross validation,* HO, *k-folds,*LOO, prediksi galat.

A Comparison of Error Prediction Methods in Classification Modeling with the C4.5 Algorithm for Balanced Data

Ichlas Djuazva

ABSTRACT

Algorithm C4.5 is one of decision tree algorithms designed to build decision tree models for data classification. C4.5 is an improvement over the Iterative Dichotomiser 3 with enhancements. Models created with this algorithm need to be tested for their accuracy to evaluate their performance. Model accuracy can be assessed by predicting error rates or prediction errors. The method used for error prediction is Cross Validation (CV). CV divides data into training data for model building and testing data for model evaluation. CV consists of several methods, namely Leave One Out (LOO), Hold Out (HO), and k-folds CV. The objective of this research is to determine which error prediction method is most suitable for use with the C4.5 algorithm.

This study uses generated data that follows a normal distribution with three data cases: univariate, bivariate, and multivariate, with various combinations of mean differences and correlations. Correlations are added to observe their impact on the resulting error prediction. In the univariate case, two different mean structures are used, in the bivariate case, four different mean structures with three different correlation structures are used, and in the multivariate case, ten different mean structures are used with five different correlation structures.

For the univariate, bivariate, and multivariate cases, k-folds CV error prediction method is found to be the most suitable method for predicting error rates in the C4.5 algorithm.

Keywords: C4.5, cross-validation, HO,k-Folds, LOO (Leave-One-Out), error prediction.

KATA PENGANTAR

Bismillahirrahmanirrahiim, Alhamdulillahirabil'alamiin, puji dan syukurpenulis ucapkan kepada Allah SWT atas segala rahmat dan hidayah-Nya sehingga penulis dapat menyelesaikan skripsiyang berjudul "Perbandingan Metode Prediksi Galat dalam Pemodelan Algoritma C4.5 untuk Data Seimbang".

Skripsi ini disusun untuk memenuhi syarat memperoleh gelar sarjana padaProgram Studi Sarjana Statistika, Departemen Statistika, Fakultas Matematika danIlmu Pengetahuan Alam, Universitas Negeri Padang. Dalam penyusunan skripsiini penulis mendapatkan banyak bantuan, bimbingan, serta dukungan dari berbagaipihak. Oleh sebab itu, dalam kesempatan ini penulis menyampaikan ucapanterimakasih kepada :

1. Bapak Dodi Vionanda, M.Si., Ph.D, selaku Kepala Departemen Statistika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Negeri Padang sekaligus pembimbing akademik dan pembimbing skripsi yang telah memberikan arahan,bimbingan, dukungan dan motivasi dalam perkuliahan sampai proses penyusunan skripsi.
2. Ibu Dra. Nonong Amalita, M.Si, selaku sekretaris Departemen Statistika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Negeri Padang dosen penguji skripsi peneliti.
3. Ibu Zilrahmi, S.Pd, M.Si, selaku dosen penguji yang telah memberikan saran danmasukan positif untuk kesempurnaan skripsi.

4. Bapak/Ibu Dosen serta staf Departemen Statistika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Negeri Padang.
5. Orang tua dan adik penulis yang telah memberikan segala do'a, kasih sayang dan dukungannya.
6. Teman-teman dan rekan-rekan yang telah banyak membantu dan memberikan dukungan kepada penulis.

Penulis menyadari bahwa skripsi ini masih terdapat kekurangan dan kesalahannya didalamnya. Oleh karena itu, penulis meminta maaf atas segala kesalahan yang dibuat dalam penulisan skripsi ini serta menerima kritik dan saran yang bersifat membangun. Semoga penulisan skripsi ini bermanfaat bagi seluruh pihak. Aamiin.

Padang, Agustus 2023

Ichlas Djuazva

DAFTAR ISI

ABSTRAK vi
KATA PENGANTAR vi
DAFTAR ISI v
DAFTAR GAMBAR vi
DAFTAR TABEL vii
DAFTAR LAMPIRAN viii
BAB IPENDAHULUAN 1
A. Latar Belakang 1
B. Batasan Masalah 5
C. Rumusan Masalah 5
D. Tujuan Penelitian 6
E. Manfaat Penelitian 6
BAB II LANDASAN TEORI 7
A. Algoritma C4.5 7
B. Prediksi Galat 13
C. Boxplot 19
BAB III METODOLOGI PENELITIAN 22
A. Jenis Penelitian 22
B. Sumber Data 22
C. Teknik Analasis 26
BAB IVHASIL DAN PEMBAHASAN 28
A. Hasil Penelitian 28
B. Pembahasan 36
BAB V KESIMPULAN DAN SARAN 40
A. Kesimpulan 40
B. Saran 40
DAFTAR PUSTAKA 41

DAFTAR GAMBAR

Gambar **Halaman**

1. Contoh pohon keputusan 8
2. Model pohon keputusan algoritma C4.5. 17
3. Skema LOO. 17
4. Skema *k-folds* CV. 19
5. Komponen boxplot. 21
6. Diagram alir langkah analisis 27
7. Boxplot kasus data univariat. 29
8. Boxplot kasus data bivariat. 30
9. Boxplot untuk kasus data bivariat dengan penambahan korelasi untuk (a) Dua variabel relevant, (b) Variabel relevant dengan irelevantt. 31
10. Boxplot untuk kasus data multivariat. 33
11. Boxplot hasil penambahan korelasi terhadap tiga variabel relevant kasus multivariat. 34
12. Boxplot hasil penambahan korelasi terhadap dua variabel relevant dan satu irrelevant kasus multivariat. 34
13. Boxplot hasil penambahan korelasi terhadap 1 variabel relevant dan 2 variabel irrelevant kasus multivariat. 35

DAFTAR TABEL

Tabel **Halaman**

1. Ketentuan untuk Data Univariat. 23
2. Ketentuan untuk Data Bivariat dan Multivariat. 23
3. Ketentuan Struktur Korelasi Kasus Bivariat. 24
4. Ketentuan Struktur Korelasi Kasus Multivariat. 25

DAFTAR LAMPIRAN

Lampiran **Halaman**

1. Tabel hasil univariat. 23
2. Tabel hasil bivariat. 23
3. Tabel hasil multivariat. 24
4. Gambar boxplot univariat. 25
5. Gambar boxplot bivariat. 46
6. Gambar boxplot multivariat. 49

BAB I
PENDAHULUAN

A. Latar Belakang

Beberapa tahun terakhir *data mining* menjadi teknik yang sangat populer untuk memperoleh informasi pada *database* dari berbagai bidang yang berbeda mengacu pada fleksibelitas kerjanya dalam beragam jenis *database*. *Data mining* merupakan langkah analisis yang dilakukan terhadap sekumpulan data berukuran besar untuk mendapatkan hubungan antar data tersebut kemudian merangkumnya kedalam bentuk yang mudah dipahami dan dapat digunakan. Hubungan dan rangkuman yang dihasilkan dapat berupa model atau pola (Prasetyowati, 2017).

Data mining mempelajari model dari sekumpulan data jumlah besar yang tersedia yang disajikan dalam bentuk model regresi maupun klasifikasi. Pohon keputusan adalah salah satu metode yang paling populer untuk klasifikasi di berbagai aplikasi *data mining* (Witten dan Frank, 2002). Pohon keputusan adalah teknik statistik berbasis pohon di mana setiap jalur yang dimulai dari akar dijelaskan oleh urutan pemisahan data hingga hasil *boolean* pada simpul daun tercapai (Yadav dan Pal, 2012). Tujuan utamanya adalah untuk membangun model *training* yang dapat digunakan untuk memprediksi kelas atau nilai variabel target melalui aturan keputusan yang disimpulkan dari data *training*.

Algoritma pohon keputusan dapat digunakan untuk menyelesaikan masalah regresi dan klasifikasi. Pohon keputusan dapat digunakan untuk klasifikasi dengan

menentukan akar terlebih dahulu dan terus bergerak sampai daun nya ditentukan (Quinlan, 2014). Penelitian ini akan berfokus pada salah satu algoritma pohon keputusan yang cukup banyak digunakan yaitu algoritma C4.5. Algortima C4.5 sendiri merupakan salah satu metode pada pohon keputusan dimana data akan dibentuk menjadi bentuk pohon yang terdiri dari sebuah daun yang mengindikasikan kelas atau node keputusan yang menjelaskan beberapa tes yang akan ditempatkan sebagai nilai atribut tunggal, dengan satu ranting dan sub pohon untuk setiap kemungkinan hasil tes.

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 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, dengan menggunakan proses *binarization*. Pada tahapan *training*, algoritma C4.5 menggunakan strategi *topdown* yang didasarkan pada pendekatan *divide* and *conquer* untuk membentuk pohon keputusan (Liu dan Gegov, 2016). Strategi *top down* akan memilih atribut yang memiliki informasi paling banyak hingga paling sedikit,

dikombinasikan dengan pendekatan *conquer and divide* yang memecah data menjadi beberapa kelompok masalah yang lebih kecil sehingga pohon yang dihasilkan relatif lebih kecil dan sederhana. Tahapan ini memetakan *trainingset* dan dengan informasi *gain ratio* sebagai tolak ukur untuk memisahkan atribut dan menghasilkan *nodes* dari akar hingga daun (Dai dan Ji, 2014). Proses klasifikasi pada Algortima C4.5 akan memprediksi data yang akan diambil sebagai keputusan (Chaturvedi, 2015).

Algoritma C4.5 akan menghasilkan model berupa diagram alir yang berbentuk seperti pohon. Model yang dihasilkan ini perlu diuji kinerjanya. Menurut Dougherty dkk (2010) metode prediksi galat (*error prediction*) adalah metode yang umum digunakan untuk mengevaluasi kinerja model. Evaluasi atau pengujian akurasi model yang diperoleh perlu dilakukan untuk melihat kemampuan model dalam melakukan prediksi data yang tidak digunakan dalam membangun model. Metode prediksi galat juga dapat digunakan untuk membandingkan dua metode maupun memilih model terbaik. Model yang tidak diuji akurasinya dapat mempengaruhi keakuratan keputusan yang akan dibuat berdasarkan model karena tidak diketahui seberapa baik atau buruknya model dalam melakukan klasifikasi. Melalui pengujian akurasi model juga dapat dilihat kecocokan model dalam melakukan analisis terhadap data atau permasalahan yang terjadi dalam model seperti *overfitting*.

Dalam penggunaan nya metode prediksi galat berguna untuk memperkirakan kesalahan yang ada pada suatu metode statistik yang diberikan untuk mengevaluasi kinerja model yang dihasilkan. Performa estimasi yang dihasilkan dari estimasi error atau prediksi galat bergantung pada aturan yang digunakan dalam pembentukan klasifikasi, sehingga setiap metode dapat menghasilkan estimasi yang berbeda.Ada

dua metode prediksi galat yang dapat digunakan dalam menguji kinerja model yaitu *training error rate* dan *testing error rate*.

Menurut Mansour dan McAllester (2002) dalam memprediksi galat *training error rate* menggunakan data *training* atau data yang telah digunakan untuk membentuk model, hal ini akan mengakibatkan nilai prediksi galat yang dihasilkan menjadi rendah bahkan nol karena data yang sama digunakan untuk membentuk model dan menguji model. *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. *Cross validation*(CV) merupakan metode *testing error rate* yang membagi data menjadi *training* dan *testing*, yang terdiri dari tiga metode yaitu *Leave One Out* (LOO), *Hold Out* (HO), dan *k-folds* CV.

Menurut James dkk (2013) LOO memisahkan satu set observasi menjadi dua bagian yaitu satu amatan pada data sebagai data *testing* sementara n-1 amatan data akan menjadi *training*. Metode HO atau juga disebut *test sample estimation* membagi data menjadi dua kelompok yang khas yang disebut dengan data *training set* dan data *test set*, atau *hold out set*. Di mana 2/3 dari data akan digunakan sebagai *data training* dan sisanya 1/3 data akan digunakan sebagai *data testing* (Kohavi, 1995).

Berdasarkan beberapa penjelasan yang diuraikan di atas, dan penelitian terdahulu yang dilakukan kohavi (1995) yang membandingkan metode CV dan *bootsrap* dalam menguji akurasi dan memilih model terbaik pada algoritma C4.5 dan *naïve bayes.* Juga penelitian terbaru mengenai pengaruh pemilihan metode prediksi galat CV terhadap metode *machine learning* oleh Tougui dkk tahun 2021, di mana

penelitian tersebut membandingkan ketiga metode prediksi galat pada *support vector machine* dan *random forest*. Kedua penelitian tersebut menghasilkan metode CV sebagai metode terbaik.

Maka dalam penelitian ini peneliti akan membandingkan ketiga metode prediksi galat yang disebutkan di atas yaitu LOO, HO, dan *k-folds* CV yang akan dilihat metode manakah yang terbaik digunakan untuk metode klasifikasi *decision tree* algoritma C4.5. Berdasarkan penjelasan yang sudah dipaparkan diatas penelitian ini akan meneliti lebih lanjut mengenai perbandingan metode prediksi galat dalam tugas akhir dalam judul "**Perbandingan Metode Prediksi Galat dalam Pemodelan Klasifikasi dengan Algoritma C4.5 untuk Data Seimbang**"

B. Batasan Masalah

Berdasarkan latar belakang yang telah dijelaskan , maka batasan masalah untuk penelitian ini adalah:

1. Dalam penelitian ini akan membahas perbandingan metode prediksi galat LOO, HOdan *k-folds* CV.
2. Data yang digunakan dalam penelitian ini merupakan data yang dibangkitkan dari software R-studio.
3. Data yang dibangkitkan pada penelitian ini berupa data numerik.

C. Rumusan Masalah

Dari latar belakang yang telah disampaikan , rumusan masalah untuk penelitian ini adalah:

Diantara metode prediksi galat (LOO, HO, dan *k-folds* CV). Metode prediksi galat mana yang paling cocok digunakan dengan algoritma C4.5 ?.

D. Tujuan Penelitian

Berdasarkan rumusan masalah yang telah dikemukakan, maka tujuan penelitian ini adalah:

Mengetahui kinerja metode prediksi galat (LOO, HO dan *k-folds* CV) pada algoritma C4.5 kemudian membandingkan kinerja ketiga metode prediksi galat sehingga didapatkan metode prediksi galat yang paling cocok digunakan pada algoritma C4.5 untuk data seimbang.

E. Manfaat Penelitian

Manfaat dari penelitian ini adalah sebagai berikut:

1. Bagi penulis menambah pengetahuan dan pemahaman penulis dalam metode prediksi galat, khususnya metode prediksi galat yang dapat digunakan pada algoritma C4.5.
2. Bagi pembaca, dapat menambah wawasan pembaca dan membantu pembaca dalam memahami perbandingan metode prediksi galat pada algoritma C4.5, serta dapat menjadi referensi dalam penelitian selanjutnya.