PERBANDINGAN METODE PREDIKSI GALAT DALAM PEMODELAN REGRESI LOGISTIK BINER UNTUK DATA TIDAK SEIMBANG (*IMBALANCED*)

SKRIPSI

Diajukan sebagai salah satu persyaratan untuk memperoleh gelar sarjana statistika

Oleh

BAHRI ANNUR SINAGA
NIM : 18337016/2018

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

2023

PERSETUJUAN SKRIPSI

PERBANDINGAN METODE PREDIKSI GALAT DALAM PEMODELAN REGRESI LOGISTIK BINER UNTUK DATA TIDAK SEIMBANG (*IMBALANCED*)

Nama	: Bahri Annur Sinaga
NIM	: 18337016
Program Studi	: S1 Statistika
Departemen	: Statistika
Fakultas	: Matematika dan Ilmu Pengetahuan Alam

Padang, 24 Agustus 2023

Mengetahui:
Ketua Departemen Statistika

Dodi Vionanda, M.Si., Ph.D
NIP. 197906112005011002

Disetujui Oleh:
Pembimbing

Dodi Vionanda, M.Si., Ph.D
NIP. 197906112005011002

PENGESAHAN LULUS UJIAN SKRIPSI

Nama	: Bahri Annur Sinaga
NIM	: 18337016
Program Studi	: S1 Statistika
Departemen	: Statistika
Fakultas	: Matematika dan Ilmu Pengetahuan Alam

PERBANDINGAN METODE PREDIKSI GALAT DALAM PEMODELAN REGRESI LOGISTIK BINER UNTUK DATA TIDAK SEIMBANG (*IMBALANCED*)

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

Padang, 24 Agustus 2023

Tim Penguji	Nama	Tanda Tangan
Ketua	: Dodi Vionanda, Ph.D	
Anggota	: Dr. Dony Permana, M.Si	
Anggota	: Admi Salma, S.Pd, M.Si	

SURAT PERNYATAAN TIDAK PLAGIAT

Saya yang bertandatangan di bawah ini:

Nama	: Bahri Annur Sinaga
NIM	: 18337016
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 Regresi Logistik Biner untuk Data Tidak Seimbang (*Imbalanced*)"** 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. 197906112005011002

Saya yang menyatakan,

Bahri Annur Sinaga
NIM. 18337016

Perbandingan Metode Prediksi Galat Dalam Pemodelan Regresi Logistik Biner Untuk Data Tidak Seimbang (*Imbalanced*)

Bahri Annur Sinaga

ABSTRAK

Regresi logistik biner merupakan analisis regresi yang digunakan dalam pemodelan klasifikasi. Kinerja regresi logistik biner bisa dilihat dari tingkat akurasi model yang terbentuk. Akurasi dapat diukur dengan memprediksi galat. Salah satu metode prediksi galat yang sering digunakan adalah *cross validation*. Pada kenyataannya data riil yang ditemukan seringkali tidak seimbang. Data tidak seimbang merupakan data yang memiliki jumlah amatan kelas yang berbeda jauh. Dalam regresi logistik data tidak seimbang memberikan pengaruh terhadap hasil prediksi. Ketika data semakin tidak seimbang hasil prediksi akan mendekati jumlah kelas mayoritas. Penelitian ini berfokus melihat bagaimana perbandingan kinerja metode prediksi galat dalam pemodelan regresi logistik biner dengan data tidak seimbang.

Terdapat tiga algoritma pada *cross validation*, yaitu *leave one out, hold out*, dan *k-fold cross validation. Leave one out* adalah metode yang membagi data berdasarkan jumlah amatan sehingga setiap amatan berkesempatan menjadi data training namun memerlukan waktu yang cukup lama dalam proses analisis ketika jumlah amatan besar. *Hold out* merupakan algoritma yang paling sederhana yang hanya membagi data menjadi dua bagian secara acak sehingga ada kemungkinan data yang penting tidak menjadi data training. *K-fold cv* adalah algoritma yang membagi data menjadi beberapa kelompok, namun *k-fold cv* tidak cocok digunakan pada data yang memiliki jumlah amatan yang kecil. Penelitian ini menggunakan data simulasi yang dibangkitkan dengan pengaturan yang berbeda-beda.

Hasil yang diperoleh adalah algoritma *k-fold cv* merupakan algoritma prediksi galat yang paling cocok diterapakan pada regresi logistik biner dengan data tidak seimbang. Semakin tidak seimbang data hasil *error rate* mendekati kelas minoritas sehingga nilainya kecil.

Kata kunci : Data Tidak Seimbang, *Hold Out*, *K-fold Cross Validation*, *Leave One Out*, Regresi Logistik Biner.

Comparison of Error Prediction Methods in Binary Logistic Regression Modeling for Imbalanced Data

Bahri Annur Sinaga

ABSTRACT

Binary logistic regression is a regression analysis used in classification modeling. The performance of binary logistic regression can be seen from the accuracy of the model formed. Accuracy can be measured by predicting errors. One error prediction method that is often used is cross validation. In reality, real data found is often imbalanced. Imbalanced data is data that has a significantly different number of class observations. In logistic regression, imbalanced data affects the prediction results. When the data is increasingly imbalanced the prediction results will approach the number of majority classes. This research focuses on seeing how the performance of error prediction methods in binary logistic regression modeling with imbalanced data compares.

There are three algorithms in cross validation: leave one out, hold out, and k-fold cross validation. Leave one out is a method that divides the data based on the number of observations so that each observation has the opportunity to become training data but requires a long time in the analysis process when the number of observations is large. Hold out is the simplest algorithm that only divides the data into two parts randomly so there is a possibility that important data does not become training data. K-fold cv is an algorithm that divides data into several groups, but k-fold cv is not suitable for data that has a small number of observations. This study uses simulated data generated with different settings.

The results obtained are k-fold cv algorithm is the most suitable error prediction algorithm applied to binary logistic regression with imbalanced data. The more imbalanced the data, the result of the error rate approaches the minority class so that the value is small.

Keywords: Binary Logistc Regression, Hold Out, Imbalanced, K-fold Cross Validation, Leave One Out.

KATA PENGANTAR

Bismillahirrahmanirrahiim, Alhamdulillahirrabil'alamiin, segala puji dan syukur penulis ucapkan kepada Allah SWT yang telah memberikan rahmat dan hidayah-Nya sehingga penulis dapat menyelasikan skripsi penulis yang berjudul "Perbandingan Metode Prediksi Galat Dalam Pemodelan Regresi Logistik Biner Untuk Data Tidak Seimbang (*Imbalanced*)".

Skripsi ini disusun untuk memenuhi syarat memperoleh gelar sarjana pada Program Studi Sarjana Statistika, Departemen Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Padang. Dalam menyusun skripsi ini penulis mendapatkan arahan dan bimbingan dari berbagai pihak. Untuk itu dalam kesempatan ini penulis menyampaikan ucapan terimakasih kepada.

1. Bapak Dodi Vionanda, M.Si., Ph.D, selaku pembimbing akademik dan pembimbing skripis serta ketua Departemen Statistika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Negeri Padang 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.
3. Bapak Dr. Dony Permana, M.Si., selaku Dosen Penguji yang telah memberikan saran dan masukan positif untuk kesempurnaan Skripsi.
4. Ibu Admi Salma S.Pd, M.Si., selaku Dosen Penguji yang telah memberikan saran dan masukan positif untuk kesempurnaan Skripsi.

5. Bapak dan Ibu Dosen serta Staf Departemen Statistika FMIPA UNP yang telah membantu dan berbagi ilmu kepada penulis selama menimba ilmu di Program Studi Statistika.
6. Orang tua penulis yang telah memberikan doa-doa terbaik dan memberikan dukungan materil dan nonmateril.
7. Serta rekan-rekan Program Studi S1 Statistika yang telah berjuang bersama selama perkuliahan.

Penulis menyadari masih terdapat kekurangan dalam Skripsi ini, oleh karena itu penulis sangat mengharapkan kritik dan saran yang membangun agar berguna untuk perbaikan berikutnya. Akhir kata penulis ucapkan terima kasih.

Padang, 24 Agustus 2023

Bahri Annur Sinaga

DAFTAR ISI

DAFTAR GAMBAR

Gambar **Halaman**

DAFTAR TABEL

Tabel **Halaman**

DAFTAR LAMPIRAN

Lampiran **Halaman**

BAB I
PENDAHULUAN

A. Latar Belakang

Penggunaan metode analisis data kategorik semakin meningkat dalam bidang biomedis maupun dalam bidang ilmu pengetahuan sosial. Hal ini menunjukkan bahwa adanya pengembangan metode dalam menganalisis data kategorik. Saat ini banyak ditemukan metode statistika yang menganalisis data kategorik salah satu metodenya adalah regresi logistik (Agresti,2002:1). Regresi logistik merupakan analasis regresi yang digunakan dalam pemodelan klasifikasi. Regresi logistik pada umumnya memiliki tugas yang sama dengan analisi regresi lainnya yaitu melihat hubungan atau pengaruh antara variabel terikat dengan satu atau lebih variabel bebas. Namun yang membedakan regresi logistik dengan analisis regresi lainnya adalah variabel terikatnya bersifat kategorik ((Hosmer & S.Lemeshow, 2013:1).

Data mining biasanya bertugas mempelajari model dari data yang tersedia baik model regresi ataupun model klasifikasi. Masalah yang biasa dihadapi dalam hal mengevaluasi model adalah sebuah model mampu memprediksi data latih secara baik namun gagal dalam memprediksi data masa depan yang tidak terlihat (Refaeilzadeh, dkk, 2016). Regresi logistik pada analisisnya akan membentuk sebuah model. Model yang dibentuk perlu dinilai akurasinya. Akurasi model dapat dinilai dengan melihat laju galatnya.

Laju galat dilihat dari perbandingan banyaknya galat dengan keseluruhan data yang digunakan dalam analisis. Perhitungan laju galat ini dilakukan dengan

mempertimbangkan bahwa suatu metode mungkin menjadi yang paling baik ketika digunakan dalam memprediksi suatu gugus data, tetapi metode tersebut belum tentu dapat memprediksi gugus data yang berbeda dengan baik. Oleh karena itu, penting untuk memprediksi laju galat gugus data tersebut sehingga mendapatkan metode yang layak. Metode yang dapat digunakan dalam memprediksi laju galat salah satunya adalah *cross validation* (CV). Metode ini membagi data menjadi perangkat latihan (*training*) dan perangkat tes (*testing*) (Molinaro dkk, 2005).

CV melakukan pembagian data berulang kali menjadi dua bagian, dimana bagian pertama digunakan untuk melatih model dan bagian kedua digunakan untuk meguji model. CV mengasumsikan bahwa data *training* dan data *testing* bersifat independen (saling bebas). CV memiliki beberapa algoritma pembelajaran dan yang paling umum digunakan adalah *leave one out* (LOO), *hold out* dan *k-fold cross validation* (James, dkk, 2013:176).

Leave one out (LOO) adalah algoritma CV yang membagi data berdasarkan jumlah pengamatan, sehingga seluruh amatan berkesempatan menjadi data *training* dan data *testing* (James, dkk, 2013:179). LOO tidak membagi data secara acak karena pembagian datanya hanya meninggalkan satu pengamatan untuk data *testing* sehingga tidak perlu dilakukan pengacakan (Wong, 2015). Refaeilzadeh, dkk (2016) dalam artikelnya mengatakan bahwa LOO menghasilkan estimasi akurasi hampir tidak bias, tetapi memiliki variansi tinggi yang mengarah kepada penelitian yang tidak dapat diandalkan. LOO umumnya digunakan pada penelitian jumlah amatan kecil, namun ketika jumlah amatan besar maka akan membutuhkan cukup banyak waktu untuk melakukan analisis.

Hold out merupakan algoritma prediksi laju galat yang paling sederhana. *Hold out* hanya membagi data menjadi dua bagian sehingga setiap amatan tidak memiliki kesempatan yang sama untuk menjadi data *training* maupun data *testing*. Hal ini mengakibatkan ada kemungkinan data yang penting pada data *training* tetapi masuk kedalam data *testing*, sehingga mempengaruhi kinerja algoritma (James, dkk, 2013:176).

Menurut Refaeilzadeh, dkk (2016), *k-fold cross validation* adalah algoritma prediksi laju galat yang membagi data kedalam beberapa kelompok. Dengan mengelompokkan data kedalam beberapa kelompok akan memudahkan dalam proses perhitungannya. *K-fold* lebih baik dari segi komputasi dibanding dengan LOO. Selain itu, variansi data yang dihasilkan pada *k-fold* cenderung lebih rendah. Kelemahan dari *k-fold* adalah data dengan ukuran sampel yang kecil tidak cocok digunakan (James, dkk, 2013:181).

Data riil yang dihasilkan pada penelitian kebanyakan merupakan data tidak seimbang. Data tidak seimbang (*imbalanced*) adalah data yang memiliki jumlah kelas amatan yang berbeda. Data tidak seimbang memberi tantangan tersendiri dalam pengklasifikasian. Data tidak seimbang jika diklasifikasikan dengan benar akan memberikan nilai yang lebih besar (Maalouf & Trafalis, 2011). Ketidakseimbangan data berdampak pada hasil prediksi yang tidak stabil. Hasil prediksi akan cenderung mengarah pada kelas mayoritas dan mengabaikan kelas minoritas (Sari, 2019). Selain ketidakseimbangan data, korelasi juga memberikan dampak kepada hasil prediksi. Pada regresi logistik variabel bebas tidak boleh memiliki korelasi yang tinggi karena akan mengakibatkan model yang menjadi bias dan selang kepercayaan yang dihasilkan sangat lebar (Courvoisier, dkk, 2011).

Penelitian ini bertujuan untuk membandingkan algoritma prediksi galat sehingga mendapatkan algoritma yang cocok diterapkan pada analisis regresi logistik biner untuk data tidak seimbang. Penelitian ini menngunakan alat perbandingan berupa boxplot. Kinerja dari metode prediksi laju galat dapat dilihat dari nilai variasi terkecil pada *output boxplot* masing-masing metode prediksi laju galat. Boxplot dapat digunakan untuk penyebaran/keragaman data pengamatan serta sebagai deteksi awal ada atau tidaknya data pencilan.

Berdasarkan hasil uraian di atas, untuk membandingkan kinerja prediksi galat, peneliti ingin melakukan penelitian yang diberi judul "**Perbandingan Metode Prediksi Galat dalam Pemodelan Regresi Logistik Biner untuk Data Tidak Seimbang (*Imbalanced*).**"

B. Batasan Masalah

Berdasarkan latar belakang di atas, batasan masalah dalam penelitian ini adalah sebagai berikut.

1. Penelitian ini menggunakan data simulasi yang dibangkitan pada *RStudio* untuk melakukan analisis.
2. Data yang digunakan pada penelitian ini adalah data tidak seimbang (*imbalanced*).
3. Algoritma prediksi galat yang digunakan adalah LOO, *hold out* dan *k-fold cross validation*.

C. Rumusan Masalah

Berdasarkan latar belakang di atas, maka masalah yang dapat dirumuskan adalah sebagai berikut.

1. Bagaimana kinerja prediksi galat LOO*, hold out* dan *k-fold cross validation* terhadap data tidak seimbang?
2. Apa algoritma prediksi galat yang cocok diterapkan pada model regresi logistik biner dengan data tidak seimbang?

D. Tujuan Penelitian

Berdasarkan rumusan masalah, maka tujuan penelitian ini adalah sebagai berikut.

1. Mengidentifikasi kinerja masing-masing algoritma prediksi galat pada data tidak seimbang.
2. Membandingkan dan mengetahui algoritma prediksi galat yang paling cocok diterapkan pada model regresi logistik biner dengan data tidak seimbang.

E. Manfaat Penelitian

Manfaat yang dapat diharapkan dari penelitian ini adalah sebagai berikut.

1. Bagi penulis, diharapkan dapat menambah ilmu tentang prediksi galat terutama dalam membandingkan metode prediksi galat dengan data tidak seimbang.
2. Bagi pembaca atau peneliti selanjutnya, diharapkan dapat menjadi salah satu pedoman dalam melakukan penelitian sejenis sehingga menambah kajian mengenai metode terkait.