PERBANDINGAN KINERJA ALGORITMA *CLASSIFICATION AND REGRESSION TREE* DAN REGRESI LOGISTIK MENGGUNAKAN UJI *F COMBINED 5×2CV*

SKRIPSI

Diajukan sebagai salah satu persyaratan untuk memperoleh gelar Sarjana Statistika

Oleh:
FAYZA ANNISA FEBRIANTI
NIM. 19337028/2019

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

2023

SURAT PERNYATAAN TIDAK PLAGIAT

Saya yang bertandatangan di bawah ini:

Nama	: Fayza Annisa Febrianti
NIM	: 19337028
Program Studi	: S1 Statistika
Departemen	: Statistika
Fakultas	: Matematika dan Ilmu Pengetahuan Alam

Dengan ini menyatakan bahwa, skripsi saya dengan judul **"Perbandingan Kinerja Algoritma *Classification and Regression Tree* dan Regresi Logistik Menggunakan Uji *F Combined 5×2cv*"** 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,

Fayza Annisa Febrianti
NIM. 19337028

PERSETUJUAN SKRIPSI

PERBANDINGAN KINERJA ALGORITMA *CLASSIFICATION AND REGRESSION TREE* DAN REGRESI LOGISTIK MENGGUNAKAN UJI *F COMBINED 5×2CV*

Nama	: Fayza Annisa Febrianti
NIM	: 19337028
Program Studi	: S1 Statistika
Departemen	: Statistika
Fakultas	: Matematika dan Ilmu Pengetahuan Alam

Padang, 25 Agustus 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	: Fayza Annisa Febrianti
NIM	: 18337028
Program Studi	: S1 Statistika
Departemen	: Statistika
Fakultas	: Matematika dan Ilmu Pengetahuan Alam

PERBANDINGAN KINERJA ALGORITMA *CLASSIFICATION AND REGRESSION TREE* DAN REGRESI LOGISTIK MENGGUNAKAN UJI *F COMBINED 5×2CV*

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

Padang, 25 Agustus 2023

Tim Penguji	Nama	Tanda Tangan
Ketua	: Dodi Vionanda, M.Si., Ph.D	
Anggota	: Dr. Yenni Kurniawati, M.Si.	
Anggota	: Fadhilah Fitri, S.Si., M.Stat	

Perbandingan Kinerja Algoritma *Classification and Regression Tree* dan Regresi Logistik Menggunakan Uji *F Combined 5×2cv*

Fayza Annisa Febrianti

ABSTRAK

Klasifikasi adalah metode untuk memperkirakan kelas suatu objek berdasarkan karakteristiknya. Beberapa algoritma pembelajaran dapat diterapkan dalam klasifikasi, seperti *Classification and Regression Tree* (CART) dan regresi logistik. Menemukan algoritma pembelajaran terbaik untuk diterapkan guna mendapatkan hasil klasifikasi terbaik merupakan hal yang penting. Dalam membandingkan dua algoritma, perbandingan langsung mungkin dapat dilakukan ketika perbedaannya sangat jelas. Namun, hal ini dapat menghasilkan kekeliruan dan kesimpulan yang tidak memadai. Statistik uji diperlukan untuk mengetahui apakah perbedaan yang terdapat dalam perbandingan tersebut terjadi secara nyata atau acak. Uji *paired t 5×2cv* dan *F combined 5×2cv* merupakan uji untuk mengetahui apakah tingkat kesalahan kedua algoritma sama atau tidak. Pada penelitian ini dilakukan perbandingan kinerja algoritma CART dan regresi logistik dengan menggunakan perbandingan langsung dan statistik uji untuk mengetahui apakah kedua algoritma memiliki tingkat kesalahan yang sama atau tidak.

Penelitian ini merupakan penelitian simulasi dengan menggunakan data bangkitan. Data dibangkitkan secara univariat dan bivariat dengan ketentuan tertentu. Penelitian dimulai dengan membangkitkan data, kemudian dilanjutkan dengan membandingkan kedua algoritma berdasarkan metode yang telah ditentukan.

Hasil penelitian pada data univariat maupun bivariat menunjukkan bahwa perbandingan langsung menggunakan *k-fold cross validation* menghasilkan kesimpulan yang tidak memadai karena perbedaan tingkat kesalahan antara CART dan regresi logistik kecil. Hasil uji *paired t 5×2cv* dengan menggunakan $p_i^{(j)}$ yang berbeda memberikan hasil yang beragam. Perubahan $p_i^{(j)}$ seharusnya tidak mempengaruhi hasil uji. Hasil uji *F combined 5×2cv* secara keseluruhan menunjukkan gagal tolak hipotesis yang berarti kedua algoritma memiliki tingkat kesalahan yang sama. Hal ini berarti bahwa CART dan regresi logistik memiliki kinerja yang sama pada penelitian ini.

Kata Kunci: CART, Regresi Logistik, *K-fold Cross Validation*, Uji *paired t 5×2cv*, uji *F combined 5×2cv*.

Algorithms Comparison of Classification and Regression Tree and Logistic Regression Using Combined 5×2cv F Test

Fayza Annisa Febrianti

ABSTRACT

Classification is a method to estimate the class of an object based on its characteristics. Several learning algorithms can be applied in classification, such as Classification and Regression Tree (CART) and logistic regression. Finding the best learning algorithm to apply to get the best classification results is important. In comparing two algorithms, the direct comparison may be possible when the differences are clear. However, this can be misleading and lead to inadequate conclusions. Statistical test are needed to determine whether the differences are real or random. The 5×2cv paired t test and the combined 5×2cv F test are to determine whether the error rates of the two algorithms are the same or not. In this study, the performance of the CART and logistic regression algorithms is compared using direct comparison and statistical test to determine whether the two algorithms have the same error rate or not.

This research is a simulation research using generated data. The data is generated in univariate and bivariate on several conditions. The research starts by generating data, then proceeds to compare the two algorithms based on predetermined methods.

The results of this research on univariate and bivariate data show that direct comparison using k-fold cross validation produces inadequate conclusions because the difference in error rates between CART and logistic regression is small. The results of the 5×2cv paired t test using different $p_i^{(j)}$ produces various results. The change in $p_i^{(j)}$should not affect the test results. The overall results of the combined 5×2cv F test show that the tests fail to reject the hypothesis which means the two algorithms have the same error rate. This indicates that CART and logistic regression perform identically in this case.

Keywords: CART, Logistic Regression, K-fold Cross Validation, 5×2cv Paired t-test, Combined 5×2cv F Test.

KATA PENGANTAR

Puji dan syukur kehadirat Allah SWT atas segala berkat, rahmat dan karunia-Nya sehingga skripsi yang berjudul Perbandingan Kinerja Algoritma *Classification and Regression Tree* dan Regresi Logistik Menggunakan Uji *F Combined 5×2cv* ini dapat terselesaikan. Penulis mengucapkan terima kasih kepada pihak-pihak yang telah membantu selama pengerjaan skripsi ini baik secara langsung maupun tidak langsung, terutama kepada:

1. Bapak Dodi Vionanda, S. Si, M. Si., Ph.D selaku dosen pembimbing skripsi yang telah membimbing, memberikan saran dan arahan dalam menyelesaikan skripsi ini.
2. Ibu Dr. Yenni Kurniawati, M.Si dan Ibu Fadhilah Fitri S.Si, M.Stat selaku dosen penguji skripsi atas waktu, bimbingan, serta kritik dan saran yang telah diberikan demi perbaikan skripsi ini.
3. Bapak Dr. Dony Permana, M.Si selaku dosen pembimbing akademik yang telah memberikan arahan, bimbingan, dan motivasi dalam proses perkuliahan.
4. Bapak dan Ibu dosen yang telah mendidik dan mengajarkan ilmunya kepada penulis selama menempuh pendidikan di Departemen Statistika, Fakultas Matematika dan Ilmu Pengetahuan, Universitas Negeri Padang.
5. Orangtua serta seluruh keluarga atas segala do'a, bantuan, dan motivasi agar segera terselesaikannya skripsi ini.
6. Seluruh rekan-rekan yang telah memberikan bantuan dan dukungan.

Semoga semua kebaikan dan ketulusan dibalas oleh Allah SWT sebagai amal ibadah. Penulis menyadari bahwa penulisan skripsi ini masih terdapat kekurangan,

dan kesalahan di dalamnya. Oleh karena itu, penulis meminta maaf atas segala kesalahan yang dibuat dalam penulisan skripsi ini dan menerima kritik saran yang bersifat membangun. Semoga penulisan skripsi ini dapat bermanfaat bagi seluruh pihak yang memerlukannya.

Padang, Agustus 2023
Penulis

Fayza Annisa Febrianti

DAFTAR ISI

DAFTAR GAMBAR

Gambar **Halaman**

DAFTAR LAMPIRAN

Lampiran **Halaman**

BAB I

PENDAHULUAN

A. Latar Belakang

Klasifikasi merupakan salah satu metode analisis data untuk mengetahui atau memperkirakan kelas dari suatu objek berdasarkan karakteristik yang ada. Klasifikasi mengelompokkan data sesuai dengan karakteristiknya secara sistematis ke dalam kelas yang telah ditentukan. Untuk mengetahui atau memperkirakan kelas dari suatu objek diperlukan sebuah model atau pengklasifikasi. Pengklasifikasi adalah sebuah fungsi yang bertugas untuk memetakan atau menentukan kelas dari sebuah input yang diberikan (Dietterich, 1998). Pembentukan pengklasifikasi memerlukan sebuah algoritma pembelajaran yang nantinya membangun pengklasifikasi dari satu set data berlabel (Kohavi, 1995).

Terdapat berbagai algoritma pembelajaran yang dapat diterapkan dalam klasifikasi, seperti *Classification and Regression Tree* (CART) dan regresi logistik. CART adalah salah satu metode dari pohon keputusan yang dikembangkan oleh Leo Breiman, Jerome H. Friedman, Richard A. Olsen dan Charles J. Stone pada tahun 1980-an. Prinsip dari metode CART adalah memilah seluruh amatan menjadi dua gugus amatan dan memilah kembali gugus amatan berikutnya, hingga diperoleh jumlah amatan minimum atau homogen pada tiap-tiap gugus amatan berikutnya (Timofeev, 2004).

Regresi logistik adalah bagian dari analisis regresi yang variabel dependennya bersifat kategorik (Rogel-Salazar, 2017: 211). Regresi logistik diterapkan untuk memodelkan variabel dependen yang bersifat kategori berdasarkan satu atau lebih variabel independen. Regresi ini digunakan untuk

melihat hubungan antara variabel dependen yang bersifat kategorik dengan variabel independen yang bersifat kategorik maupun kontinu.

Dalam beberapa kasus, tujuan utama dari klasifikasi adalah untuk menemukan algoritma pembelajaran terbaik yang dapat diterapkan untuk menghasilkan pengklasifikasi atau model terbaik. Pada sistem surat elektronik untuk mendeteksi dan menyaring apakah pesan yang masuk adalah sampah atau tidak, digunakan sebuah algoritma pembelajaran dalam programnya. Algoritma pembelajaran ini yang nantinya akan menganalisis akumulasi contoh pesan sampah atau tidak dan memperbarui aturan penyaringannya secara berkala. Sehingga, pada kasus ini sangat penting untuk memilih algoritma pembelajaran terbaik yang cocok untuk digunakan dalam program (Dietterich, 1998).

Pemilihan algoritma terbaik dapat dilihat melalui akurasi. Akurasi berkaitan dengan performa kinerja seberapa bagus kemampuan model yang dihasilkan dalam memprediksi di masa depan. Galat dapat mengukur kinerja model dengan menghitung segala bentuk tingkat kesalahan prediksi pada model. Pada klasifikasi, galat dapat dilihat dengan melihat apakah hasil prediksi sama dengan nilai yang sebenarnya (Wood, 2017).

Data berukuran besar memberikan hasil prediksi galat yang dapat diandalkan dibandingkan data berukuran kecil. Pada data berukuran besar, data dapat disisihkan sebagai data uji untuk mengevaluasi kinerja algoritma. Sehingga didapatkan dua set data yang nantinya dapat digunakan untuk membentuk model dan mengevaluasi kinerja algoritma. Penggunaan data yang sama untuk membentuk model dan mengevaluasi kinerja algoritma akan menghasilkan bias yang disebabkan oleh *overfitting* (Raschka, 2018).

Pada umumnya ukuran data yang tersedia cukup terbatas, sehingga semua data yang ada harus digunakan dalam algoritma pembelajaran guna membentuk model. Hal ini mengakibatkan tidak memungkinkannya untuk menyisihkan data uji. Untuk mengatasi permasalahan ini dapat dilakukan resampling sebagai solusi untuk membagi data menjadi data latih dan data uji. Salah satu contoh resampling yang dapat dilakukan adalah *cross validation*. *Cross validation* membagi data menjadi dua bagian (data latih dan data uji) secara berulang, sehingga setiap data memiliki kesempatan yang sama (Raschka, 2018).

Salah satu teknik *cross validation* adalah *k-fold cross validation*. *K-fold cross validation* membagi data hampir sama rata secara acak kedalam *k* kelompok. Pada setiap perulangan (*fold*) sebanyak *k,* akan digunakan satu kelompok sebagai data uji dan pada setiap perulangannya digunakan kelompok yang berbeda untuk data uji. Estimasi tingkat kesalahan atau prediksi galat dapat dihitung pada setiap perulangan, sehingga prediksi galat dari metode *k-fold cross validation* dapat dihitung dengan merata-ratakan nilai prediksi galat yang diperoleh dari setiap perulangan.

Dalam membandingkan dua algoritma pembelajaran, perbandingan secara langsung dengan membandingkan dan melihat nilai akurasi yang lebih besar atau nilai prediksi galat yang lebih kecil mungkin dapat dilakukan ketika perbedaannya sangat jelas. Namun pada kebanyakan kasus, perbandingan langsung ini mungkin dapat menyesatkan, sehingga tidak cukup untuk ditarik kesimpulan. Untuk mengatasi hal ini, maka diperlukan sebuah statistik uji untuk mengetahui apakah perbedaan yang terdapat dalam perbandingan tersebut terjadi secara nyata atau acak (Stapor, 2017).

Statistik uji yang baik adalah statistik uji yang mampu mengontrol segala sumber keragaman yang ada. Sumber keragaman ini dapat berasal dari pemilihan data latih dan data ujinya. Perbedaan pemilihan data latih dan data uji memungkinkan perbedaan hasil. Sumber keragaman dari data latih dapat diatasi dengan menjalankan algoritma pembelajaran secara berulang dan mengukur variasi akurasi dari model yang dihasilkan. Sumber keragaman yang berasal dari pemilihan data uji, dapat diatasi dengan mempertimbangkan ukuran data uji dan konsekuensi bahwa perubahan data uji memungkinkan hasil prediksi galat yang berbeda (Dietterich, 1998).

Untuk mengatasi masalah tersebut, pada tahun 1998 Dietterich mengenalkan uji *paired t 5×2cv* sebagai salah satu uji untuk mengetahui apakah kedua algoritma memiliki tingkat kesalahan yang sama atau tidak. Pada uji ini dilakukan 5 replikasi dari *2-fold cross validation*. Pada setiap replikasi, data secara acak dibagi menjadi dua bagian sama banyak.

Menurut Alpaydin (1999), penggunaan uji *paired t 5×2cv* memungkinkan adanya hasil yang berbeda. Pergantian urutan replikasi atau kelompok yang akan digunakan pada numerator $p_i^{(j)}$memberikan hasil yang berbeda-beda. Uji ini terkadang menerima dan menolak hipotesis yang ada. Padahal seharusnya pergantian urutan replikasi atau kelompok yang akan digunakan pada numerator $p_i^{(j)}$ tidak mempengaruhi hasil yang ada atau dengan kata lain memberikan hasil uji yang sama.

Alpaydin memperkenalkan sebuah uji yang lebih kuat untuk mengatasi kekurangan yang ada pada uji *paired t 5×2cv*, yaitu dengan mengkombinasikan seluruh hasil galat yang didapatkan pada setiap replikasi atau kelompok. Uji ini

adalah uji *F Combined 5×2cv*. Berbeda dengan uji *paired t 5×2cv*, uji ini menggunakan uji *F* untuk mengkombinasikan sepuluh hasil galat yang didapatkan pada setiap replikasi atau kelompok.

Berdasarkan uraian tersebut, maka akan dilakukan penelitian untuk membandingan algoritma CART dan regresi logistik menggunakan uji *F Combined 5×2cv*. Judul yang akan diangkat pada penelitian ini adalah "**Perbandingan Kinerja Algoritma *Classification and Regression Tree* dan Regresi Logistik menggunakan Uji *F Combined 5×2cv***.

B. Batasan Masalah

Berdasarkan latar belakang yang telah dikemukakan, maka batasan masalah dari penelitian ini adalah sebagai berikut.

1. Pembahasan dalam penelitian ini mengenai perbandingan kinerja dari algoritma CART dan regresi logistik dalam membangun sebuah model yang dilihat sama atau tidaknya tingkat kesalahan (*error rate*).
2. Perbandingan langsung dalam penelitian ini menggunakan *10-fold cross validation* dengan melihat selisih tingkat kesalahan dari kedua algoritma.
3. Penelitian ini menggunakan data bangkitan yang dibangkitkan dari populasi berbeda secara univariat dan bivariat.

C. Rumusan Masalah

Berdasarkan latar belakang yang telah dikemukakan, maka rumusan masalah dari penelitian ini adalah sebagai berikut.

1. Bagaimana perbandingan langsung kinerja algoritma CART dan regresi logistik dengan menggunakan metode *k-fold cross validation*?

2. Bagaimana perbandingan kinerja algoritma CART dan regresi logistik dengan menggunakan uji *paired t 5×2cv*?
3. Bagaimana perbandingan kinerja algoritma CART dan regresi logistik dengan menggunakan uji *F Combined 5×2cv*?

D. Tujuan Penelitian

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

1. Untuk mengetahui perbandingan langsung kinerja algoritma CART dan regresi logistik dengan menggunakan metode *k-fold cross validation*.
2. Untuk mengetahui perbandingan kinerja algoritma CART dan regresi logistik dengan menggunakan uji *paired t 5×2cv.*
3. Untuk mengetahui perbandingan kinerja algoritma CART dan regresi logistik dengan menggunakan uji *F Combined 5×2cv.*

E. Manfaat Penelitian

Manfaat yang diharapkan dari hasil penelitian ini adalah sebagai berikut.

1. Bagi penulis, dapat menjadi pengalaman dalam menganalisis data serta menambah ilmu dan pemahaman mengenai uji *F Combined 5×2cv*.
2. Bagi pembaca atau mahasiswa lainnya, penelitian ini diharapkan dapat menjadi pedoman atau referensi sebagai alat pertimbangan dalam melakukan penelitian sejenis serta menambah wawasan tentang kajian mengenai metode terkait.