

OPTIMASI PARAMETER TUNING PADA MODEL REGRESI LOGISTIK LASSO

Syarifah Fadhilah

(Matematika, Fakultas Ilmu Pengetahuan Alam, Universitas Sumatera Utara, Medan, Indonesia)

E-mail: syarifahfadhilah13@students.usu.ac.id*

Abstrak

Regresi logistik LASSO merupakan metode klasifikasi biner yang efektif dalam mengatasi multikolinieritas dan menyederhanakan model melalui seleksi variabel otomatis. Penelitian ini bertujuan untuk mengevaluasi pengaruh kombinasi tiga metode validasi silang (KFold, Stratified KFold, dan Repeated KFold) dan tiga rentang nilai parameter tuning terhadap performa dan kesederhanaan model regresi logistik LASSO. Evaluasi dilakukan pada data sekunder dari Kaggle serta tiga dataset bangkitan dengan ukuran dan kompleksitas berbeda. Model dievaluasi menggunakan metrik log loss, F1 score, apparent error rate (APER), dan jumlah variabel terseleksi. Hasil menunjukkan bahwa nilai tuning optimal bergantung pada karakteristik data dan metode validasi silang yang digunakan. Stratified KFold memberikan performa terbaik pada data tidak seimbang, sedangkan KFold lebih efisien untuk dataset besar. LASSO terbukti mampu menyederhanakan model tanpa kehilangan akurasi secara signifikan.

Kata Kunci: regresi logistik, LASSO, parameter tuning, cross-validation

Abstract

LASSO logistic regression is an effective binary classification method to handle multicollinearity and simplify models through automatic variable selection. This study aims to evaluate the impact of combining three cross-validation methods (KFold, Stratified KFold, and Repeated KFold) and three ranges of tuning parameter values on the performance and simplicity of LASSO logistic regression models. The evaluation was conducted on a secondary dataset from Kaggle and three synthetic datasets with varying sizes and complexities. Models were assessed using log loss, F1 score, apparent error rate (APER), and the number of selected variables. The results show that the optimal tuning value depends on the characteristics of the data and the cross-validation method used. Stratified KFold performed best on imbalanced data, while KFold was more efficient for large datasets. LASSO effectively reduced model complexity without significantly sacrificing accuracy.

Keywords: logistic regression, LASSO, parameter tuning, cross-validation

PENDAHULUAN (GUNAKAN STYLE SECTION)

Regresi logistik merupakan salah satu metode statistik yang banyak digunakan dalam klasifikasi biner karena kemampuannya mengestimasi probabilitas dan menginterpretasi pengaruh variabel prediktor. Namun, performa regresi logistik dapat menurun ketika data mengandung multikolinieritas atau jumlah prediktor yang besar. Untuk mengatasi permasalahan tersebut, regresi logistik dengan penalti LASSO (Least Absolute Shrinkage and Selection Operator) menjadi solusi alternatif yang efektif. LASSO tidak hanya

membantu mengatasi multikolinieritas tetapi juga melakukan seleksi variabel secara otomatis dengan menyusutkan sebagian koefisien menjadi nol.

Kendati demikian, performa model regresi logistik LASSO sangat ditentukan oleh nilai parameter tuning (λ) yang digunakan. Nilai λ yang terlalu kecil dapat menyebabkan overfitting karena model menjadi terlalu kompleks, sedangkan nilai yang terlalu besar dapat menyebabkan underfitting akibat banyaknya variabel penting yang dieliminasi. Oleh karena itu, penentuan nilai λ yang optimal menjadi tahap penting dalam membangun model yang akurat dan efisien.

Proses pemilihan nilai tuning parameter λ umumnya dilakukan menggunakan pendekatan validasi silang (cross-validation), seperti KFold, Stratified KFold, dan Repeated KFold. Cross-validation membantu mengevaluasi performa model terhadap berbagai nilai λ dan memilih parameter yang mampu menjaga keseimbangan antara generalisasi model dan kesederhanaan struktur. Namun, efektivitas masing-masing metode cross-validation serta rentang nilai λ dapat berbeda tergantung pada karakteristik dataset yang digunakan.

Penelitian ini bertujuan untuk:

1. Menganalisis pengaruh kombinasi metode validasi silang dan rentang nilai parameter tuning terhadap performa dan kesederhanaan model regresi logistik LASSO;
2. Menentukan nilai parameter tuning λ yang memberikan performa terbaik dan struktur model paling efisien;
3. Membandingkan model regresi logistik LASSO hasil tuning dengan regresi logistik biasa tanpa penalti.

Penelitian ini menggunakan data sekunder dari platform Kaggle serta tiga dataset bangkitan yang memiliki karakteristik berbeda dari segi ukuran sampel dan tingkat multikolinieritas. Evaluasi performa model dilakukan dengan mengamati metrik log loss, F1 score, APER (Apparent Error Rate), dan jumlah variabel terseleksi.

Dengan melakukan eksplorasi sistematis terhadap parameter tuning dan metode cross-validation, penelitian ini diharapkan mampu memberikan kontribusi dalam pengembangan model prediktif yang lebih akurat dan efisien, serta dapat menjadi referensi dalam penerapan LASSO pada berbagai kasus klasifikasi biner lainnya.

KAJIAN TEORI

Regresi Logistik Biner

Regresi logistik biner merupakan model regresi yang digunakan ketika variabel respons bersifat dikotomi (bernilai 0 atau 1). Model ini digunakan

untuk memprediksi probabilitas suatu kejadian berdasarkan satu atau lebih variabel prediktor. Bentuk umum dari model regresi logistik biner adalah:

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p}}$$

Dimana:

$\pi(x)$: probabilitas bahwa kejadian "sukses" yaitu $Y = 1$ terjadi
 $x_1, x_2, \dots, x_p$: variabel independen
 $\beta_0, \beta_1, \dots, \beta_p$: koefisien regresi

Multikolinieritas

Multikolinieritas adalah kondisi ketika dua atau lebih variabel prediktor dalam model regresi memiliki hubungan linear yang tinggi. Hal ini dapat menyebabkan ketidakstabilan dalam estimasi parameter dan kesalahan interpretasi. Salah satu cara untuk mendeteksi multikolinieritas adalah dengan melihat nilai Variance Inflation Factor (VIF). Untuk mengatasinya, digunakan metode regularisasi seperti LASSO.

Regresi Logistik LASSO

LASSO (Least Absolute Shrinkage and Selection Operator) adalah teknik regularisasi yang menggunakan penalti L1 untuk menyusutkan koefisien regresi. Dalam konteks regresi logistik, fungsi objektif yang diminimasi adalah:

$$\hat{\beta}^{lasso} = \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{2} \sum_{i=1}^N \left(y_i - \beta_0 - \sum_{j=1}^p x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^p |\beta_j| \right\}$$

dengan λ sebagai parameter tuning yang mengatur kekuatan penalti. LASSO mampu mengatasi multikolinieritas dan melakukan seleksi variabel secara otomatis dengan membuat beberapa koefisien menjadi nol.

Cross-Validation(CV)

Cross-validation adalah teknik statistik untuk mengevaluasi kinerja model dengan membagi data menjadi subset latih dan uji secara berulang.

- *KFold CV* membagi data ke dalam kkk bagian dan melakukan pelatihan dan pengujian sebanyak kkk kali.
- *Stratified KFold* memastikan proporsi kelas pada setiap fold tetap seimbang.
- *Repeated KFold* mengulang proses KFold beberapa kali dengan pembagian data yang berbeda.

Evaluasi Model

Beberapa metrik evaluasi model yang digunakan dalam penelitian ini adalah:

- *Log Loss*: mengukur ketepatan prediksi probabilitas model; semakin kecil nilainya, semakin baik model.
- *F1 Score*: rata-rata harmonik dari presisi dan recall, cocok untuk data tidak seimbang.
- *Apparent Error Rate (APER)*: proporsi klasifikasi yang salah dari seluruh data uji.
- *Jumlah Variabel Terseleksi*: mencerminkan kesederhanaan model; lebih sedikit variabel menandakan model yang lebih efisien.

METODE (GUNAKAN STYLE SECTION)

Penelitian ini merupakan penelitian kuantitatif komputasional yang bertujuan untuk mengevaluasi kombinasi metode validasi silang dan rentang nilai parameter tuning pada model regresi logistik LASSO. Rancangan penelitian melibatkan eksperimen sistematis menggunakan data sekunder dan data bangkitan yang dianalisis dengan pendekatan statistik prediktif.

Data yang digunakan terdiri atas satu dataset sekunder dan tiga dataset bangkitan. Dataset sekunder diambil dari platform Kaggle (Saiful Islam, 2022), yang merupakan data klasifikasi biner dengan jumlah 300 observasi dan 11 variabel prediktor. Tiga dataset bangkitan dibuat menggunakan distribusi normal multivariat untuk mencerminkan struktur korelasi antar variabel, dengan rincian:

- Dataset I: 150 observasi, 11 variabel (9 ordinal dan 2 kontinu), tanpa noise.
- Dataset II: 550 observasi, 11 variabel utama + 4 variabel noise, multikolinieritas sedang.
- Dataset III: 1500 observasi, 11 variabel utama + 9 variabel noise, multikolinieritas tinggi.

Data dibagi menjadi dua bagian, yaitu data pelatihan (80%) dan data pengujian (20%), untuk menguji generalisasi model. Teknik pengumpulan data dilakukan secara dokumentasi (untuk data sekunder) dan simulasi komputer (untuk data bangkitan). Seluruh proses dilakukan dengan bahasa pemrograman Python menggunakan pustaka *numpy*, *pandas*, dan *scikit-learn*.

Model yang dibangun adalah regresi logistik dengan penalti LASSO, yang diimplementasikan melalui fungsi *LogisticRegressionCV* dari pustaka *scikit-learn*. Parameter tuning yang diuji adalah nilai $C=1/\lambda$, dengan tiga rentang nilai C yang disusun dalam skala logaritmik. Tiga metode validasi silang digunakan untuk mengevaluasi model, yaitu:

- KFold (5 fold),
- Stratified KFold (5 fold, menjaga proporsi kelas),
- Repeated KFold (5 fold, 3 kali pengulangan).

Evaluasi model dilakukan berdasarkan empat indikator, yaitu log loss, F1 score, apparent error rate (APER), dan jumlah variabel terseleksi (jumlah koefisien regresi non-nol). Log loss digunakan untuk menilai akurasi prediksi probabilistik, F1 score untuk data tidak seimbang, APER untuk kesalahan klasifikasi, dan jumlah variabel terseleksi untuk mengevaluasi kesederhanaan model.

HASIL DAN PEMBAHASAN

Penelitian ini bertujuan untuk mengevaluasi pengaruh kombinasi metode validasi silang dan rentang nilai parameter tuning terhadap performa model regresi logistik LASSO. Hasil penelitian ini disajikan berdasarkan empat dataset berbeda, yaitu dataset asli dan tiga dataset bangkitan yang

dirancang memiliki kompleksitas dan karakteristik multikolinieritas yang bervariasi.

a) Dataset Asli

Pada dataset asli, kombinasi nilai $C=0,1$ $C = 0,1$ dengan metode Stratified KFold menghasilkan performa terbaik dengan nilai log loss yang rendah, F1 score tertinggi (0,91), serta mempertahankan 9 dari 11 variabel prediktor. Hal ini menunjukkan bahwa metode Stratified KFold efektif menangani distribusi kelas yang tidak seimbang. Model penalti LASSO juga terbukti berhasil mengurangi multikolinieritas dengan menyisihkan variabel yang tidak signifikan.

b) Dataset Bangkitan I

Dataset ini memiliki ukuran kecil (150 observasi) dan tanpa variabel noise. Hasil menunjukkan bahwa model cenderung mengalami overfitting saat menggunakan nilai CCC besar, dengan perbedaan signifikan antara performa pelatihan dan pengujian. Nilai CCC kecil menghasilkan model yang lebih stabil tetapi dengan F1 score yang lebih rendah. Stratified KFold dan Repeated KFold memberikan hasil yang lebih konsisten dibandingkan KFold, menunjukkan pentingnya menjaga proporsi kelas saat jumlah data terbatas.

c) Dataset Bangkitan II

Pada dataset ini ditambahkan 4 variabel noise, dengan jumlah sampel menengah (550). Hasil evaluasi menunjukkan bahwa nilai CCC menengah (0,1-0,3) memberikan keseimbangan antara akurasi dan jumlah variabel terseleksi. Model dengan C optimal dari rentang kedua cenderung menyisihkan variabel noise, menunjukkan bahwa LASSO bekerja efektif dalam menyaring fitur yang tidak relevan. KFold dan Stratified KFold memiliki performa yang hampir seimbang, namun Stratified sedikit lebih unggul dalam menjaga kestabilan F1 score pada data pengujian.

d) Dataset Bangkitan III

Dengan jumlah observasi besar (1500) dan 9 variabel noise, semua metode validasi silang

memberikan model yang relatif stabil. Nilai $C=0,3162$ $C = 0,3162$ menghasilkan log loss terkecil dan F1 score tertinggi (0,97) baik pada pelatihan maupun pengujian, namun menyebabkan model mempertahankan lebih banyak variabel. Hal ini menunjukkan bahwa pada data besar, model dengan kompleksitas tinggi tetap mampu menghasilkan prediksi yang baik. KFold terbukti lebih efisien secara komputasi dibanding metode lain, tanpa penurunan performa yang signifikan.

PENUTUP

SIMPULAN

Penelitian ini menunjukkan bahwa pemilihan parameter tuning pada model regresi logistik LASSO sangat memengaruhi performa dan kesederhanaan model. Kombinasi antara metode validasi silang dan rentang nilai parameter tuning $C=1/\lambda$ $C = 1/\lambda$ memberikan dampak yang berbeda-beda tergantung pada karakteristik dataset. Pada dataset dengan distribusi kelas tidak seimbang dan jumlah sampel kecil, Stratified KFold terbukti lebih stabil. Sementara itu, KFold lebih cocok digunakan pada dataset berukuran besar karena efisiensinya.

Model regresi logistik LASSO mampu melakukan seleksi variabel secara otomatis dengan menyusutkan koefisien variabel yang tidak signifikan ke nol, sehingga menghasilkan model yang lebih sederhana dan mudah diinterpretasikan dibandingkan dengan regresi logistik biasa. Nilai parameter tuning optimal yang paling sering muncul berada pada rentang menengah ($C=0,1$ $C = 0,1$), yang memberikan keseimbangan antara akurasi dan kompleksitas model.

SARAN

Penelitian ini dapat diperluas lebih lanjut dengan mengevaluasi pengaruh pemilihan parameter tuning pada jenis penalti lain seperti Ridge atau ElasticNet. Selain itu, penerapan model LASSO yang telah dioptimasi juga perlu diuji pada kasus nyata dalam bidang lain seperti kesehatan, pemasaran, atau pendidikan untuk mengukur generalisasi model di luar data simulasi. Penelitian di masa depan juga dapat mempertimbangkan

pendekatan automasi pemilihan tuning parameter berbasis optimasi bayesian atau grid search adaptif.

DAFTAR PUSTAKA

Aliu, A. (2024). *MATH LOCUS: Jurnal Riset dan Inovasi Pendidikan Matematika Pemodelan Tingkat Kecanduan Games Online Menggunakan Regresi Logistik Ordinal* Nurhidayah 1a) , Indahwati 2b) , Anwar Fitrianto 3c) , Erfiani 4d) *Modeling The Level of Online Games Addiction Using Ordinal Logistic Regression*. 5(1), 41–48.

Amalia, N. (2024). Optimasi Algoritma Random Forest dengan Hyperparameter Tuning Menggunakan GridSearchCV untuk Prediksi Nasabah Churn pada Industri Perbankan. *Jurnal Manajemen Informatika*, 16(2).

Anil, A. K. P., & Singh, U. K. (2023). An Optimal Solution to the Overfitting and Underfitting Problem of Healthcare Machine Learning Models. *Journal of Systems Engineering and Information Technology (JOSEIT)*, 2(2), 77–84. <https://doi.org/10.29207/joseit.v2i2.5460>

Arlot, S., & Celisse, A. (2010a). A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4, 40–79. <https://doi.org/10.1214/09-SS054>

Arlot, S., & Celisse, A. (2010b). A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4, 40–79. <https://doi.org/10.1214/09-SS054>

Bain, C., Shi, D., Banad, Y., Ethridge, L., Norris, J., & Loeffelman, J. (2025). A Tutorial on Supervised Machine Learning Variable Selection Methods in Classification for the Social and Health Sciences in R. *Journal of Behavioral Data Science*, 5(1), 1–45. <https://doi.org/10.35566/jbds/bain>

Berrar, D. (2018). Cross-validation. In *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics* (Vols. 1–3, pp. 542–545). Elsevier. <https://doi.org/10.1016/B978-0-12-809633-8.20349-X>

Chen, Y., & Yang, Y. (2021). The One Standard Error Rule for Model Selection: Does It Work? *Stats*, 4(4), 868–892. <https://doi.org/10.3390/stats4040051>

Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow* SECOND

EDITION Concepts, Tools, and Techniques to Build Intelligent Systems (R. Roumeliotis & N. Tache, Eds.; second). O'Reilly Media, inc. <http://oreilly.com>

Habibi, L. N., Matsui, T., & Tanaka, T. S. T. (2024). Critical evaluation of the effects of a cross-validation strategy and machine learning optimization on the prediction accuracy and transferability of a soybean yield prediction model using UAV-based remote sensing. *Journal of Agriculture and Food Research*, 16. <https://doi.org/10.1016/j.jafr.2024.101096>

Hastie, T., Martin, R. T., Hastie, W., Tibshirani, •, & Wainwright, •. (n.d.). *Statistical Learning with Sparsity The Lasso and Generalizations Statistical Learning with Sparsity Monographs on Statistics and Applied Probability* 143.

Holter, J. C., & Stallrich, J. W. (2023). Tuning parameter selection for penalized estimation via R2. *Computational Statistics and Data Analysis*, 183. <https://doi.org/10.1016/j.csda.2023.107729>

Iqbal, M., Mahmud Nawawi, H., Rangga Ramadhan Saelan, M., Sony Maulana, M., & Mustopa, A. (2023). OPTIMASI HYPERPARAMETER MULTILAYER PERCEPTRON UNTUK PREDIKSI DAYA BELI MOBIL. *Jurnal Manajemen Informatika & Sistem Informasi (MISI)*, 6(1). <https://doi.org/10.36595/misi.v5i2>

Restu Ningsih, D., Keumala Intan, P., Yuliati, D., Prodi Matematika, M., Sains dan Teknologi, F., Sunan Ampel, U., & Prodi Matematika, D. (2023). Pemodelan Tindak Pidana Kriminalitas di Kota Tangerang Menggunakan Metode Regresi Lasso. *Estimasi: Journal of Statistics and Its Application*, 4(1), 2721–379. <https://doi.org/10.20956/ejsa.vi.24853>

Rusman, J., Haryati, B. Z., & Michael, A. (2023). Optimisasi Hiperparameter Tuning pada Metode Support Vector Machine untuk Klasifikasi Tingkat Kematangan Buah Kopi. *Jurnal Komputer Dan Informatika*, 11(2), 195–202. <https://doi.org/10.35508/jicon.v11i2.12571>

Sahar Fadhil AL-KHATEEB. (2019). TUNING PARAMATER SELECTION IN PENALIZED LOGISTIC REGRESSION WITH APPLICATION IN CANCER. *Istanbul Commerce University Journal of Science*, 11–22. <http://dergipark.gov.tr/ticaretfbid>

Wijanarto, W., & Puspitasari, R. (2019). Optimasi Algoritma Klasifikasi Biner dengan Tuning Parameter pada Penyakit Diabetes Mellitus. *Eksplora Informatika*, 9(1), 50–59.
<https://doi.org/10.30864/eksplora.v9i1.257>

Zuliana, S. U., & Ula, A. (2024). Optimisasi Analisis Regresi LASSO Pendugaan Penyakit IMS dengan Algoritma LARS Menggunakan GCV. In *BIAStatistics: Journal Of Statistics Theory and Applications* (Vol. 18, Issue 2).