Leibniz : Jurnal Matematika
Volume 2, Nomor 2, Halaman 1–10
Juli 2022
e-ISSN: 2775-2356

PERFORMA REGRESI RIDGE DAN REGRESI LASSO PADA DATA DENGAN MULTIKOLINEARITAS

Fitri Rahmawati[1*)], Risky Yoga Suratman[2)]

[1,2)]*Alumni Magister Matematika, Universitas Gadjah Mada*
[*]*email: fitrirahma66.fr@gmail.com*

Abstrak: Analisis regresi klasik dengan metode kuadrat terkecil (*ordinary least square*) memiliki beberapa asumsi. Salah satu asumsinya yaitu tidak terjadi multikolinearitas pada variabel-variabel prediktor. Jika pada data terjadi multikolinearitas, ada beberapa metode lain yang dapat digunakan diantaranya regresi lasso dan regresi ridge. Dua model regresi ini merupakan metode *shrinkage* yang dapat menyusutkan koefisien regresi sehingga variansinya turun. Pada penelitian ini dibandingkan performa dari regresi ridge dan regresi lasso untuk data dengan multikolinearitas. Hasil dari rata-rata kuadrat eror (MSE) menunjukkan bahwa performa regresi ridge lebih baik dibanding regresi lasso. Adapun dalam hal interpretasi model, regresi lasso dinilai lebih unggul. Hal ini karena regresi lasso dapat menyusutkan beberapa koefisien menjadi nol sehingga hanya tersisa 4 dari 9 variabel yang digunakan dalam final model.

Kata Kunci: *regresi ridge, regresi lasso, multikolinearitas*

Abstract: *Classical regression analysis with the OLS (ordinary least square) has several assumptions. One of the assumptions is that there is no multicollinearity in the predictor variables. If multicollinearity occurs in the data, there are several other methods that can be used, including lasso regression and ridge regression. These two regression models are shrinkage methods that can shrink the regression coefficient so that the variance decreases. In this study, the performance of ridge regression and lasso regression was compared for data with multicollinearity. The result of the mean of squared errors (MSE) shows that the performance of the ridge regression is better than the lasso regression. In terms of model interpretation, lasso regression is considered superior. This is because lasso regression can shrink some coefficients to zero so that only 4 of the 9 variables used in the final model.*

Keywords: *ridge regression, lasso regression, multicollinearity*

PENDAHULUAN

Analisis regresi adalah metode analisis data yang digunakan untuk memodelkan pengaruh variabel prediktor terhadap variabel respon. Model regresi linier dengan variabel respon y dan beberapa variabel prediktor $x_i,\ i = 1,2, \dots, k$, persamaannya dapat dinyatakan:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon \quad (1)$$

dimana $\beta_0, \beta_1, \beta_2, \dots, \beta_k$ merupakan koefisien regresi dan ε adalah eror yang tidak dapat dijelaskan oleh variabel-variabel prediktor x_i.

Estimasi parameter yang paling umum digunakan dalam analisis regresi adalah metode kuadrat terkecil (*Ordinary Least Square*). *Ordinary Least Square* (OLS) merupakan salah satu cara untuk mengestimasi parameter β dengan meminimumkan jumlah dari kuadrat eror persamaan regresinya. Rencher and Schaalje (2008) menyatakan estimasi parameter β dari persamaan (1) dengan menggunakan OLS ialah:

$$\hat{\beta} = (X'X)^{-1}X'y. \quad (2)$$

Gujarati and Porter (2008) menyatakan bahwa salah satu asumsi yang harus dipenuhi dalam analisis regresi adalah tidak adanya multikolinearitas. Lebih lanjut, (Rencher and Schaalje 2008) menyatakan tingginya multikolinearitas mengindikasikan variabel-variabel prediktor berkorelasi satu sama lain. Multikolinearitas dapat menyebabkan variansi parameter menjadi lebih besar dan menurunkan akurasi dari estimasi. Untuk mengatasi hal ini, salah satu cara yang dapat digunakan menurut (James et al. 2013) adalah dengan menyusutkan (*shrinkage*) koefisien yang diestimasi. Metode *shrinkage* sering juga disebut metode regularisasi. Pendekatan ini menyusutkan parameter mendekati nol relatif terhadap estimasi kuadrat terkecil. Dua pendekatan *shrinkage* yang sering digunakan adalah Regresi Ridge dan Regresi Lasso.

Regresi lasso pertama kali diperkenalkan oleh Robert Tibshirani pada tahun 1996. (Tibshirani 1996) menyatakan LASSO merupakan kependekan dari *Least Absolute Shrinkage and Selection Operator*, yaitu suatu metode estimasi yang meminimumkan jumlah kuadrat eror yang bergantung pada jumlahan nilai mutlak dari koefisien. Regresi Lasso mampu menghilangkan variabel-variabel pada model regresi dengan menyusutkan koefisiennya menjadi nol. Estimasi parameter regresi lasso yaitu mengestimasi β yang meminimumkan:

$$\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{k} x_{ij}\beta_j\right)^2 + \lambda \sum_{j=1}^{k} |\beta_j|. \quad (3)$$

Regresi ridge memiliki metode estimasi yang mirip dengan regresi lasso. Jika pada regresi lasso estimasi parameter menggunakan jumlah kuadrat eror yang bergantung pada jumlahan nilai mutlak dari koefisien, pada regresi ridge estimasi parameter menggunakan jumlah kuadrat eror yang bergantung pada jumlahan kuadrat koefisien-koefisiennya. Secara matematis estimasi parameter regresi ridge berdasarkan (James et al. 2013) dapat dinyatakan:

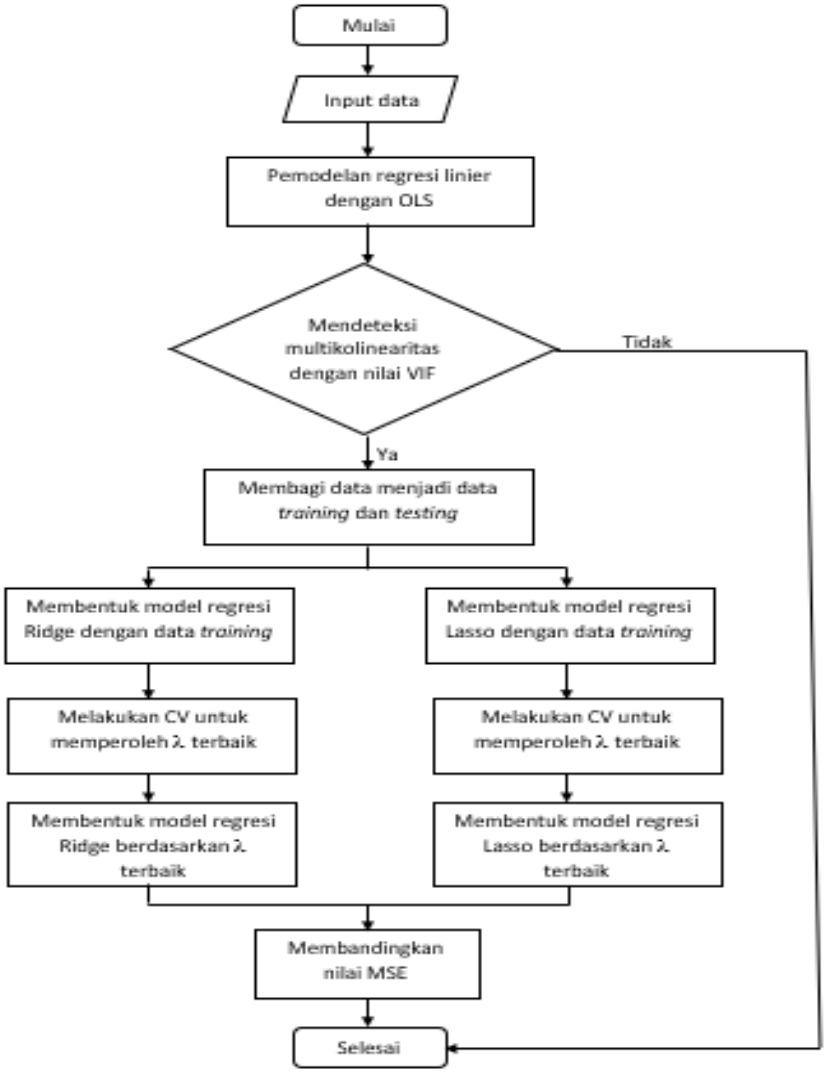
Leibniz : Jurnal Matematika
Volume 2, Nomor 2, Halaman 1–10
Juli 2022
e-ISSN: 2775-2356

$$\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{k} x_{ij}\beta_j\right)^2 + \lambda \sum_{j=1}^{k} \beta_j^2. \quad (4)$$

Persamaan (3) dan (4) menunjukkan bahwa perbedaan estimasi pada regresi ridge dan regresi lasso hanya terletak pada bentuk $|\beta_j|$ dan β_j^2. Regresi ridge memungkinkan koefisien menyusut mendekat nol adapun regresi lasso memungkinkan koefisien regresi menyusut tepat menjadi nol. Pada penelitian ini akan diterapkan regresi ridge dan regresi lasso pada data dengan multikolinearitas. Performa keduanya akan dibandingkan dengan menggunakan nilai MSE (*mean square error*).

METODE PENELITIAN

Langkah analisis data ditampilkan dalam *flowchart* berikut:



Gambar 1. Langkah Analisis Data

Metode penelitian yang dilakukan adalah studi literatur, yaitu mempelajari sumber literatur yang terkait dan penelitian-penelitian yang telah dilakukan sebelumnya. Data yang akan diaplikasikan dalam penelitian ini adalah dataset *College* pada package *ISLR* di software R. Dataset ini merupakan data statistik kampus-kampus di Amerika Serikat. Data yang digunakan berupa pengamatan pada 777 kampus dengan 10 variabel. Variabel data yang digunakan pada penelitian ini dimuat pada Tabel 1 berikut ini.

Tabel 1. Variabel Pada Data

Nama variabel	Keterangan
Apps (y)	Jumlah pendaftar di universitas tersebut
F.Undergraduate (x_1)	Jumlah mahasiswa yang mengambil program sarjana *fulltime*
P.Undergraduate (x_2)	Jumlah mahasiswa yang mengambil program sarjana *parttime*
Room.Board (x_3)	Biaya asrama mahasiswa
Enroll (x_4)	Jumlah mahasiswa baru yang melakukan registrasi
Expend (x_5)	Pengeluaran tiap mahasiswa
Books (x_6)	Perkiraan biaya buku
PhD (x_7)	Presentase pengajar dengan gelar PhD
Grad.Rate (x_8)	Tingkat kelulusan
Accept (x_9)	Total mahasiswa baru yang diterima di universitas

Pengujian multikolinearitas pada data dilakukan dengan mencari nilai VIF atau *variance inflation factor*. Menurut (Supriyadi, Mariani, and Sugiman 2017) variabel dikatakan mengalami multikolinearitas jika nilai VIF nya lebih dari 10. Dalam (Andana, Safitri, and Rusgiyono 2017), nilai VIF dapat dicari dengan formula:

$$VIF_j = \frac{1}{(1 - R_j^2)}, \text{dengan } j = 1,2, \dots, k. \quad (5)$$

R_j^2 adalah koefisien determinasi yang diperoleh dengan meregresikan variabel prediktor x_j dengan variabel-variabel prediktor lainnya.

Pemodelan regresi ridge dan regresi lasso pada data dilakukan dengan menggunakan package *glm.net* pada software R. Sebelumnya, data dibagi menjadi dua yaitu data *training* dan data *testing*. Menurut Fallo (2021) Data *training* digunakan untuk membentuk model klasifikasi sedangkan data *testing* digunakan untuk mengukur ketepatan keberhasilan model. Perhitungan akurasi model menggunakan nilai MSE (*mean square error*). Pham (2019) menyatakan MSE dapat digunakan untuk menghitung deviasi dari hasil prediksi dengan nilai sebenarnya. Formula dari MSE yang digunakan yaitu:

Leibniz : Jurnal Matematika
Volume 2, Nomor 2, Halaman 1–10
Juli 2022
e-ISSN: 2775-2356

$$MSE = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}. \quad (6)$$

Pencarian λ terbaik pada persamaan (3) dan (4) menggunakan metode *cross validation.* (Efron and Tibshirani 1993) menyatakan bahwa salah satu metode dari *cross validation* adalah dengan menggunakan *k-fold.* Dalam (James et al. 2013), *k-fold cross validation* dilakukan dengan membagi data ke dalam *k-fold*. Selanjutnya *fold* pertama digunakan sebagai data untuk validasi sedangkan k-1 *fold* sisanya digunakan untuk membentuk model. Hal ini diulang hingga *fold* terakhir sebagai data validasi. Formula dari *k-fold cross validation* adalah:

$$CV_k = \frac{1}{k}\sum_{i=1}^{k} MSE_i. \quad (7)$$

HASIL DAN PEMBAHASAN

Hasil Penelitian

Pada penelitian ini ingin dibentuk model regresi dari faktor-faktor yang mempengaruhi jumlah pendaftar universitas di Amerika Serikat. Data yang terdiri dari 777 pengamatan dengan 1 variabel respon dan 9 variabel lain sebagai variabel prediktor diterapkan pada regresi linier klasik menggunakan software R. Pemodelan regresi linier dengan OLS menghasilkan estimasi parameter yang termuat pada Tabel 2 berikut ini.

Tabel 2. Estimasi Parameter Regresi Linier

Variabel	Estimasi Parameter
Intercept	-145,8
x_1	0,1194
x_2	0,01221
x_3	-0,01025
x_4	-0,7115
x_5	0,1022
x_6	0,2699
x_7	-2,535
x_8	9,626
x_9	1,498

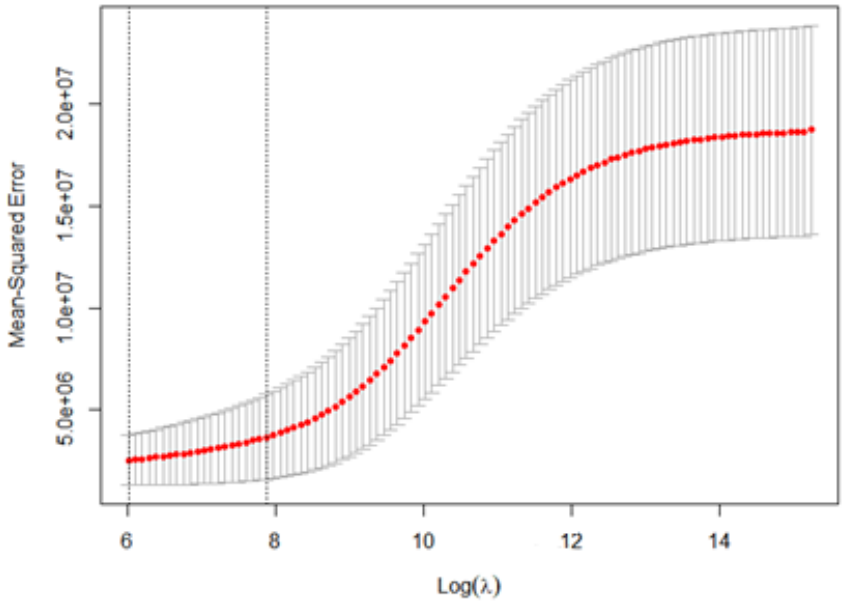
Selanjutnya untuk mendeteksi multikolinearitas, dicari nilai VIF yang ditampilkan pada Tabel 3 berikut.

Tabel 3. Nilai VIF Tiap Variabel Prediktor

Variabel	Nilai VIF
x_1	16,795379
x_2	1,640588
x_3	1,616053
x_4	20,862959
x_5	1,582015
x_6	1,046724
x_7	1,501072
x_8	1,456727
x_9	6,627111

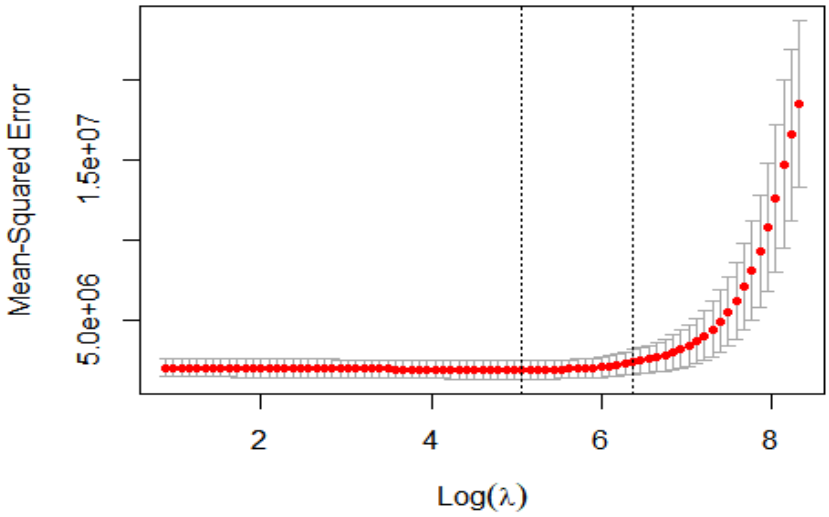
Berdasarkan Tabel 3, terdapat dua variabel yang memiliki nilai VIF > 10, yaitu x_1 dan x_4. Berdasarkan hal tersebut terdeteksi terjadinya multikolinearitas dari variabel-variabel prediktor yang menyimpang dari asumsi regresi klasik. Selanjutnya dilakukan pemodelan regresi lasso dan regresi ridge pada data.

Data dibagi menjadi data *training* dan data *testing*, sebanyak 388 pengamatan sebagai data *training* dan 389 sebagai data *testing*. Data *training* diaplikasikan pada regresi ridge dan regresi lasso dengan membangkitkan λ acak sebanyak 100. Selanjutnya menggunakan *k-fold cross validation* diperoleh nilai MSE pada tiap λ sebagai berikut:



Gambar 2. *Cross Validation* pada Regresi Ridge

Leibniz : Jurnal Matematika
Volume 2, Nomor 2, Halaman 1–10
Juli 2022
e-ISSN: 2775-2356



Gambar 3. *Cross Validation* pada Regresi Lasso

Diperoleh nilai λ terbaik yang meminimumkan MSE pada regresi ridge yaitu λ=410,7007. Adapun pada regresi lasso diperoleh nilai λ terbaik yang meminimumkan MSE adalah λ=158,2647.

Koefisien regresi pada regresi ridge dan regresi lasso dengan masing-masing λ terbaik hasil *k-fold cross validation* ditampilkan dalam Tabel 4 berikut.

Tabel 4. Koefisien Regresi Ridge dan Regresi Lasso pada Variabel

Nama variabel	Koefisien Regresi Ridge	Koefisien Regresi Lasso
Intercept	-2421,382	-831,1356
x_1	0,0710799	0
x_2	0,005016038	0
x_3	0,1580417	0,003635027
x_4	0,4945965	0
x_5	0,07803019	0,06797697
x_6	0,366622	0
x_7	2,79822	0
x_8	11,91173	4,502778
x_9	1,070781	1,427879

Berdasarkan Tabel 2 dan Tabel 4, untuk variabel yang mengalami multikolinearitas yaitu x_1 dan x_4 terjadi penyusutan koefisien regresi pada model regresi ridge dan regresi lasso. Koefisien regresi dengan OLS pada x_1 sebesar 0,1194, adapun pada

regresi ridge menyusut menjadi 0,0710799 dan menjadi 0 pada regresi lasso. Selanjutnya untuk variabel x_4 koefisien awalnya pada regresi dengan OLS sebesar -0,7115, adapun pada regresi ridge menyusut menjadi 0,4945965 dan menjadi 0 pada regresi lasso.

Regresi ridge dalam Kusuma and Wulansari (2020) mampu menstabilkan koefisien regresi yang disebabkan adanya multikolinearitas. Meskipun estimasi parameter yang dihasilkan bersifat bias, akan tetapi estimasi parameter regresi ridge mampu mendekati nilai parameter yang sebenarnya. Semakin besar nilai λ pada persamaan (4) maka koefisien regresi yang dihasilkan akan semakin mendekati nol yang berakibat berkurangnya sensitivitas model terhadap variabel independen. Pada estimasi parameter dengan regresi lasso, koefisien regresi variabel dengan multikolinearitas yakni x_1 dan x_4 disusutkan tepat menjadi nol. Artinya variabel x_1 dan x_4 tidak digunakan lagi dalam model regresi lasso, dengan demikian masalah multikolinearitas pada variabel-variabel independen teratasi.

Persamaan regresi ridge yang dihasilkan:

$$y = -2421{,}382 + 0{,}0710799x_1 + 0{,}005016038\,x_2 + 0{,}1580417x_3 + 0{,}4945965x_4 + 0{,}07803019x_5 + 0{,}366622x_6 + 2{,}79822x_7 + 11{,}91173x_8 + 1{,}070781x_9 \quad (8)$$

Adapun persamaan regresi lasso adalah:

$$y = -831{,}1356 + 0{,}003635027x_3 + 0{,}06797697x_5 + 4{,}502778x_8 + 1{,}427879x_9 \quad (9)$$

Akurasi kedua model dihitung menggunakan MSE pada persamaan (6). Dengan menggunakan data *testing* sejumlah 389 pengamatan, data variabel-variabel prediktor dimasukkan ke dalam persamaan (8) untuk regresi ridge dan (9) untuk regresi lasso. MSE yang diperoleh dimuat pada Tabel 5 berikut.

Tabel 5. Nilai MSE tiap Model Regresi

Model	Nilai MSE
Regresi Ridge	1107166
Regresi Lasso	1138842

Pembahasan

Regresi ridge pada persamaan (8) menghasilkan koefisien-koefisien regresi yang relatif menyusut mendekati nol. Adapun regresi lasso pada persamaan (9) menghasilkan beberapa koefisien regresi yang menyusut tepat menjadi nol, diantaranya pada variabel x_1, x_2, x_4, x_6, dan x_7. Dengan demikian variabel prediktor yang berpengaruh pada variabel *y* (jumlah pendaftar di universitas) di regresi lasso hanyalah

variabel: biaya asrama mahasiswa, pengeluaran tiap mahasiswa, tingkat kelulusan, serta total mahasiswa baru yang diterima di universitas.

Hal ini sejalan dengan yang dikemukakan (Kusuma and Wulansari 2020) bahwa koefisien pada regresi ridge akan disusutkan menuju nol, adapun pada regresi lasso beberapa koefisien regresi dapat disusutkan tepat menjadi nol. Penyusutan koefisien tepat menjadi nol ini yang membuat regressi lasso lebih unggul dalam hal seleksi variabel. Berkurangnya beberapa variabel berdapak pada interpretasi model yang lebih efisien.

Nilai MSE dari kedua model dapat dilihat pada Tabel 5. Berdasarkan hasil komputasi, regresi ridge menpunyai MSE 1107166 adapun pada regresi lasso sebesar 1138842. Dapat disimpulkan bahwa regresi ridge mempunya MSE yang lebih kecil dari regresi lasso. Akibatnya performa regresi ridge lebih bagus dalam kaitannya dengan besaran eror. Regresi lasso mengasumsikan beberapa koefisien bernilai nol, sehingga menurut (James et al. 2013) hal ini berdampak pada MSE nya yang sedikit lebih besar dibanding regresi ridge.

Secara garis besar James et al. (2013) menyatakan regresi lasso mempunyai performa yang bagus pada data dengan sedikit variabel prediktor yang koefisiennya cukup besar, adapun variabel prediktor sisanya mempunyai koefisien yang kecil mendekati nol. Sebaliknya, regresi ridge akan memberikan performa yang lebih bagus pada data dengan variabel prediktor yang cukup banyak dengan koefisien yang relatif sama besar.

SIMPULAN

Berdasarkan hasil dan pembahasan yang telah dipaparkan, dapat disimpulkan regresi ridge lebih unggul dari regresi lasso dalam hal ukuran MSE. Regresi ridge menpunyai MSE yang lebih kecil yaitu 1107166 dibandingkan pada regresi lasso sebesar 1138842. Akan tetapi dalam hal seleksi variabel dan interpretasi model, regresi lasso dinilai lebih unggul. Hal ini karena dapat direduksinya variabel $x_{1,}$ $x_{2,}$ $x_{4,}$ $x_{6,}$ dan x_7 sehingga interpretasi model lebih efisien dengan hanya menggunakan 4 variabel prediktor yaitu $x_{3,}$ $x_{5,}$ $x_{8,}$ dan $x_{9.}$

DAFTAR PUSTAKA

Andana, Aulia Putri, Diah Safitri, and Agus Rusgiyono. 2017. "*Model Regresi Menggunakan Least Absolute Shrinkage and Selection Operator (Lasso) Pada Data Banyaknya Gizi* Buruk Kabupaten/Kota Di Jawa Tengah." *Jurnal Gaussian* 6(1):21–30.

Efron, Bradley., and Robert Tibshirani. 1993. *An Introduction to the Bootstrap*. 1st ed.

New York: Chapman & Hall.

Fallo, S. I. (2021). *Support Vector Machine, Naive Bayes Classifier, dan Regresi Logistik Ordinal dalam Prediksi Cuaca* (Doctoral dissertation, Universitas Gadjah Mada).

Gujarati, Damodar N., and Dawn C. Porter. 2008. *Basic Econometrics*. Boston: McGraw-Hill Irwin.

James, G., D. Witten, T. Hastie, and R. Tibshirani. 2013. *An Introduction to Statistical Learning - with Applications in R*. New York: Springer.

Kusuma, Guntur Wahyu, and Ika Yuni Wulansari. 2020. "*Analisis Kemiskinan Dan Kerentanan Kemiskinan Dengan Regresi Ridge, Lasso, Dan Elastic-Net Di Provinsi Jawa Tengah Tahun 2017*." *Seminar Nasional Official Statistics* 2019(1):503–13. doi: 10.34123/semnasoffstat.v2019i1.189.

Pham, Hoang. 2019. "*A New Criterion for Model Selection*." *Mathematics* 7(12):1–12. doi: 10.3390/MATH7121215.

Rencher, Alvin C., and G. Bruce Schaalje. 2008. *Linear Models in Statistics*. Hoboken, New Jersey: John Wiley and Sons, Inc.

Supriyadi, E., S. Mariani, and Sugiman. 2017. "*Perbandingan Metode Partial Least Square (PLS) Dan Principal Component Regression (PCR) Untuk Mengatasi Multikolineritas Pada Model Regresi Linear Berganda*." *UNNES Journal of Mathematics* 6(2):117–28.

Tibshirani, Robert. 1996. "*Regression Shrinkage and Selection Via the Lasso*." *Journal of the Royal Statistical Society: Series B (Methodological)* 58(1):267–88. doi: 10.1111/j.2517-6161.1996.tb02080.x.