MIMS Majalah Ilmiah Matematika dan Statistika 24 (1), 2024, 12 – 26
eISSN 2722-9886 pISSN 1411-6669
https://jurnal.unej.ac.id/index.php/MIMS/index

Regularisasi model pembelajaran mesin dengan regresi terpenalti pada data yang mengandung multikolinearitas (Studi kasus prediksi Indeks Pembangunan Manusia di 34 provinsi di Indonesia)

(Regularization of machine learning models with penalized regression on multicollinear data (Case study of human development index prediction in 34 provinces in Indonesia))

Nur Khamidah*, Kusman Sadik, Agus M Soleh, Gerry Alfa Dito
IPB Universiy, Bogor, Indonesia

*korespondensi: nur.khamidah@apps.ipb.ac.id

Received: 20-06-2023, accepted: 20-02-2024

Abstract

This research intends to model high-dimensional data that contains multicollinearity in four machine-learning algorithms: Random Forest, K-Nearest Neighbor, XGBoost, and Regression Tree. Previously, regularization was carried out with penalized ridge regression, least absolute shrinkage and selection operator (LASSO) regression, and Elastic Net regression. A total of 100 predictor variables and 1 response variable which are the Development Index 2022 data of 34 provinces in Indonesia from BPS were used and standardized. The simulation is also applied to highly correlated data on two distributions, uniform and normal with parameter values taken from existing empirical data. The results showed that the ridge regularization method is the best for producing accurate and stable predictions. Furthermore, there was no difference in the root mean square error (RMSE) results between the data with standardization and without standardization, wherein all the data analyzed it was found that the kNN model was better than other models on simulation data, and the Random Forest and XGBoost models were better than other models on empirical data. In addition, the Regression Tree model is not recommended according to the results of this study.

Keywords: regularization, multicollinearity, ridge, LASSO, elastic net
MSC2020: 62J07

1. Pendahuluan

Regresi merupakan sebuah metode yang memungkinkan dapat diketahui pengaruh peubah prediktor terhadap peubah respon yang bertipe numerik ataupun kategorik[1]. Dalam konteks pembelajaran mesin (*machine learning*), analisis regresi termasuk dalam algoritma pembelajaran terbimbing (*supervised learning*) yang membutuhkan respon numerik kontinu yang telah diketahui [2]. Beberapa contoh model regresi yang dapat dibangun dengan algoritma pembelajaran mesin antara lain regresi gulud (*ridge*), *least*

absolute shrinkage and selection operator (LASSO), Elastic Net, Support Vector Regression, regresi hutan acak (*random forest*), dan lain-lain.

Permasalahan terjadi ketika data yang digunakan dalam analisis regresi tidak cukup besar. Walaupun di era big data seperti saat ini, masih terdapat beberapa penelitian di bidang pengobatan, penelitian yang menyasar pada subjek langka atau yang sulit diamati, penelitian yang subjek bersifat agregat menjadi sulit untuk diperoleh data yang berukuran besar [3]. Direkomendasikan data yang berukuran lebih besar [4], namun beberapa peneliti menentukan batas minimum ukuran data, di antaranya N = 5 pada setiap kelompok yang diteliti [5], N = 10-20 pada setiap peubah prediktor [6], bahkan [7] merekomendasikan N sebesar 50 kali lipat dari peubah prediktor yang ada.

Era *big data* saat ini, data yang muncul memiliki tipe atau karakteristik yang umum di mana satu amatan dalam data pada satu waktu dapat mengandung fitur yang berukuran sangat banyak, yang selanjutnya dikenal dengan tipe data berdimensi tinggi [8]. Hal ini mengakibatkan kinerja komputasi menjadi buruk akibat waktu running yang lama, beresiko terjadi *overfit*, pendugaan parameter sulit dilakukan terutama pada model yang berbasis OLS, karena data berdimensi tinggi rentan terjadi multikolinearitas yang mengakibatkan matriks X bersifat singular yang berakibat buruk pada pendugaan parameter model [9]. Masalah-masalah ini dapat diantisipasi dengan menggunakan beberapa cara, antara lain seleksi fitur [10], peningkatan ukuran data, reduksi dimensi [11] [12], dan regularisasi [13].

Regularisasi adalah bagaimana mengontrol kompleksitas model dengan menambahkan penalti pada fungsi kerugian pada pelatihan model yang dilakukan untuk mengatasi permasalahan data berdimensi tinggi [13]. Hal ini dimaksudkan untuk mengantisipasi permasalahan *overfit* [14] dan mengurangi dampak multikolinearitas pada data berdimensi tinggi [10]. Beberapa model regularisasi yang dapat digunakan antara lain *ridge*/gulud (regularisasi L2) [15], LASSO (regularisasi L1) [16], dan kombinasi keduanya dengan Elastic Net [17]. Model-model regularisasi mengendalikan koefisien model dan melakukan penyusutan pada koefisien-koefisien yang tidak diperlukan dalam model.

Beberapa penelitian sebelumnya melakukan kombinasi dari model pembelajaran mesin dan metode regularisasi dalam meningkatkan performa model. Penelitian [19] melakukan kombinasi Random Forest dan Elastic Net menghasilkan reduksi *root mean square error* (RMSE) sebesar 16% dari pohon semula. Penelitian [20] melakukan kombinasi model yang sama dengan teknik *ensemble* menghasilkan penurunan MSE dibandingkan menggunakan salah satu model. Penelitian [21] melakukan *stacking* beberapa model regresi pembelajaran mesin antara lain LASSO, ridge, OLS, *Adaboost*, dan lain-lain dengan Random Forest *Regressor* untuk mengurangi galat, dan dihasilkan penyusutan RMSE model *stacking* yang lebih kecil mencapai 67% dari RMSE awal.

Dalam penelitian ini, akan dilakukan pemodelan data berdimensi tinggi dan mengandung multikolinearitas pada empat algoritma pembelajaran mesin yang meliputi Random Forest, K-Nearest Neighbor, XGBoost, dan Regression Tree yang dengan menggunakan data Indeks Pembangunan Manusia (IPM). Indeks Pembangunan Manusia (IPM) merupakan suatu indikator yang digunakan untuk mengukur kemajuan suatu wilayah berdasarkan tiga aspek penting, antara lain harapan hidup ketika penduduk dilahirkan, akses terhadap pendidikan yang berkualitas, dan standar hidup yang layak yang diukur melalui pendapatan kasar penduduk [18]. Mengingat indeks ini mencakup indikator kelayakan hidup penduduk di berbagai bidang, keberadaan indikator ini baik secara langsung maupun tidak langsung dipengaruhi oleh berbagai indikator lain. Sebelumnya, dilakukan regularisasi dengan regresi terpenalti *ridge*, LASSO, dan Elastic Net. Hal ini dilakukan untuk menguji apakah metode regularisasi dengan model regresi terpenalti yang dilakukan mampu memberikan kontribusi terhadap performa model pembelajaran mesin yang dibentuk.

2. Metodologi

Penelitian ini mencoba melakukan perbandingan performa model regresi berbasis pembelajaran mesin yang sebelumnya diterapkan penalti menggunakan model regularisasi ridge, LASSO, dan Elastic Net pada dua jenis data: data yang dilakukan standardisasi dan tanpa standardisasi, yang mana keduanya memiliki peubah prediktor berdimensi tinggi dan saling berkorelasi. Keseluruhan proses penelitian dilakukan dengan menggunakan software RStudio versi 2023.03.0+386 dengan beberapa package antara lain *caret, simstudy, MASS, glmnet,* dan *tidymodels*.

2.1 Data

Penelitian ini menggunakan dua jenis data, yaitu data simulasi dan data empiris. Sebagaimana terlihat pada Gambar 1, data empiris yang digunakan merupakan data Indeks Pembangunan Manusia 2022 dari 34 provinsi di Indonesia [22] yang berisi 100 peubah prediktor dari berbagai bidang yang diambil dari situs BPS dan publikasi BPS yang secara detail tersaji di bagian Lampiran.

	Y <dbl>	X1 <dbl>	X2 <int>	X3 <int>	X4 <int>	X5 <dbl>	X6 <dbl>	X7 <dbl>	X8 <int>
1	72.80	26063.50	45648	111920	1372	0.291	77.55	12014.96	4361460
2	72.71	37943.83	88348	126732	5650	0.326	82.02	13859.09	15929695
3	73.26	32377.51	46336	94090	2543	0.292	68.68	21514.66	4479643
4	73.52	80057.79	44784	44976	819	0.323	83.64	9127.69	7019927
5	72.14	44536.39	26370	28754	630	0.335	80.36	5760.10	12799855
6	70.90	39676.95	66728	77494	1621	0.330	77.29	15458.87	14183140
7	72.16	24230.02	14964	20636	220	0.315	79.81	3329.27	5042041
8	70.45	28064.39	66816	82805	1219	0.313	83.89	21176.20	14144972
9	72.24	38674.15	9458	18578	574	0.255	92.24	3395.04	1128115
10	76.46	87238.26	12552	17878	147	0.325	91.62	1390.39	1244159

1–10 of 34 rows | 1–10 of 101 columns Previous 1 2 3 4 Next

Gambar 1. Data empiris yang diteliti

Data simulasi yang dibangkitkan memiliki karakteristik peubah yang disesuaikan dengan data empiris, dengan rincian sebagai berikut:

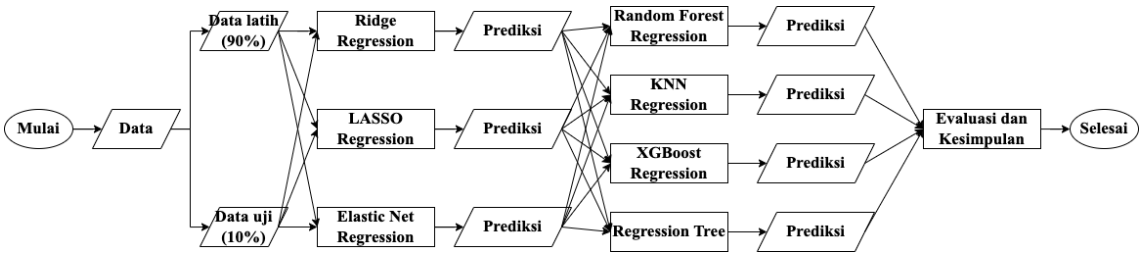
1. Data simulasi yang dibangkitkan menggunakan sebaran seragam berisi 100 peubah prediktor dan 50 amatan dan peubah tersebut saling berkorelasi tinggi satu sama lain. Parameter sebaran yang digunakan pada masing-masing peubah (nilai terkecil dan nilai maksimum) diambil dari parameter sebaran masing-masing peubah prediktor pada data empiris.

2. Data simulasi yang dibangkitkan menggunakan sebaran Normal berisi 100 peubah prediktor dan 50 amatan dan peubah tersebut saling berkorelasi tinggi satu sama lain. Parameter sebaran yang digunakan pada masing-masing peubah (rata-rata dan ragam) diambil dari parameter sebaran masing-masing peubah prediktor pada data empiris.

Perbedaan sebaran yang dibangkitkan pada kedua data dimaksudkan untuk mengetahui apakah terdapat pengaruh standardisasi pada data terhadap performa model stacking yang dibangun.

2.2 Metode Analisis

Secara rinci penelitian ini dilakukan dengan alur pada Gambar 2 sebagai berikut:



Gambar 2. Alur penelitian

Setelah dilakukan pembangkitan data, analisis diawali dengan menguji multikolinearitas untuk memastikan terjadinya multikolinearitas pada data yang diteliti agar metode regularisasi dapat dilakukan, uji adanya multikolinearitas dilakukan dengan uji Bartlett pada hipotesis sebagai berikut [23]:

H_0: semua korelasi dalam populasi bernilai 0 (tidak terjadi multikolinearitas).
H_1: terdapat korelasi dalam populasi bernilai $\neq 0$ (terjadi multikolinearitas).

Penelitian ini mengkombinasikan model regresi terpenalti ridge, LASSO, dan Elastic Net untuk menghasilkan prediksi yang bebas dari multikolinear karena beberapa koefisiennya disusutkan. Selanjutnya, hasil prediksi akan digunakan sebagai peubah respon yang baru untuk dilakukan pembentukan model pembelajaran mesin. Dalam membentuk model, terdapat beberapa hyperparameter yang perlu dilakukan tuning atau optimasi yang

menghasilkan model dengan performa terbaik. Beberapa hyperparameter yang dilakukan tuning pada penelitian ini antara lain dirincikan pada Tabel 1.

Tabel 1. Deskripsi *hyperparameter* yang di-*tuning*

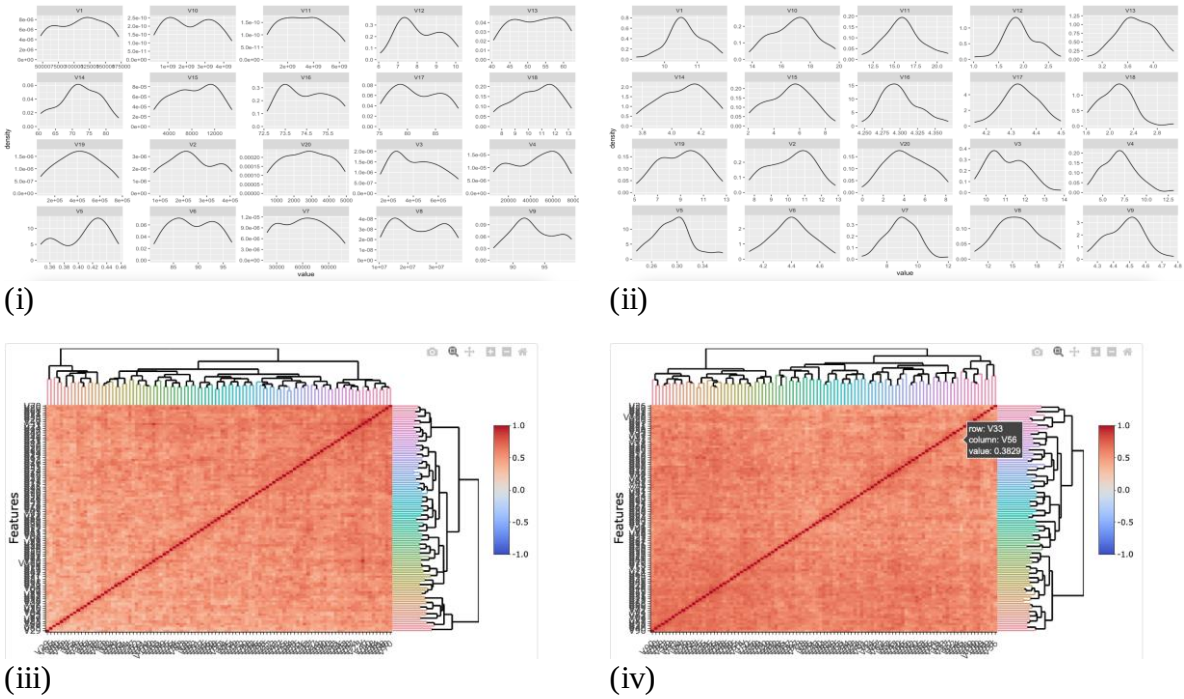
Model	**Hyperparameter**	**Definisi**
Ridge, LASSO	lambda	Konstanta yang mewakili besaran penalti L2 (*ridge*) dan L1 (LASSO) yang diterapkan pada model, ditetapkan dalam penelitian ini antara 10^{-5} hingga 10^{5} sebanyak 500 nilai.
	alpha	Koefisien optimasi penalti yang digunakan untuk regularisasi. Pada *ridge*, alpha = 0, dan pada LASSO, alpha = 1.
Elastic Net	lambda	Konstanta yang mewakili besaran penalti L2 (*ridge*) dan L1 (LASSO) yang diterapkan pada model, ditetapkan dalam penelitian ini antara 10^{-5} hingga 10^{5} sebanyak 500 nilai.
	alpha	Koefisien optimasi penalti yang digunakan untuk regularisasi. Nilai alpha yang di-*tuning* dalam penelitian ini adalah dari 0.1, 0.2, …, 0.9.
Random Forest	mtry	Banyak prediktor yang secara acak dilakukan *sampling* pada setiap pembangunan model pohon, dalam hal ini nilainya antara 4 hingga 9.
	trees	Banyak pohon yang dimuat dalam setiap *ensemble*, ditentukan acak oleh *package*.
	min_n	Minimum banyak amatan pada setiap *node* agar bisa dilakukan pemisahan, ditentukan acak oleh *package*.
kNN	k	Banyak tetangga terdekat yang disertakan dalam proses pemodelan, dalam hal ini dipilih antara 1, 4, 8, 10, 12, 16, dan 20.
XGBoost	eta	Parameter *learning rate* yang mengontrol seberapa banyak informasi pada pohon baru yang diambil untuk *boosting*, dalam hal ini ditentukan 0.001, 0.01, 0.1, 0.5, dan 0.75.
	nrounds	Banyak iterasi yang dilakukan, dipilih antara 5, 10, 25, 50, dan 75.
	max_depth	Maksimum kedalaman pohon yang terbentuk, ditentukan antara 1, 3, 5, 7, dan 10.
Regression Tree	maxdepth	Maksimum kedalaman pohon yang terbentuk, ditentukan antara 2 hingga 10.

Selanjutnya, setelah terbentuk model yang optimal, dilakukan prediksi dan evaluasi prediksi menggunakan RMSE. RMSE dirumuskan dengan:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} e_i^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}\,.$$

3. Hasil dan Pembahasan

Data berkorelasi tinggi dibangkitkan mengikuti dua aturan, yaitu data terstandardisasi dan tidak terstandardisasi. Sebaran prediktor dan korelasi dari dua data simulasi yang dibangkitkan dapat dilihat pada Gambar 3.



Gambar 3. Sebaran peubah prediktor dan korelasi data hasil bangkitan, di mana: (i) sebaran sebagian prediktor pada data simulasi yang dibangkitkan dengan sebaran seragam, (ii) sebaran sebagian prediktor pada data simulasi yang dibangkitkan dengan sebaran normal, (iii) plot korelasi data simulasi data tidak distandardisasi, dan (iv) plot korelasi data simulasi data standardisasi

Untuk memastikan adanya multikolinearitas, dilakukan uji Bartlett dengan tingkat signifikansi sebesar 5% yang kemudian diperoleh hasil uji pada masing-masing data.

Tabel 2. Hasil uji Bartlett

Data	**df**	***p-value***	**Keputusan**
Data empiris tanpa standardisasi	99	2×10^{-6}	Tolak H_0
Data empiris dengan standardisasi	99	2×10^{-6}	Tolak H_0
Data simulasi seragam	99	2×10^{-6}	Tolak H_0
Data simulasi normal	99	2×10^{-6}	Tolak H_0

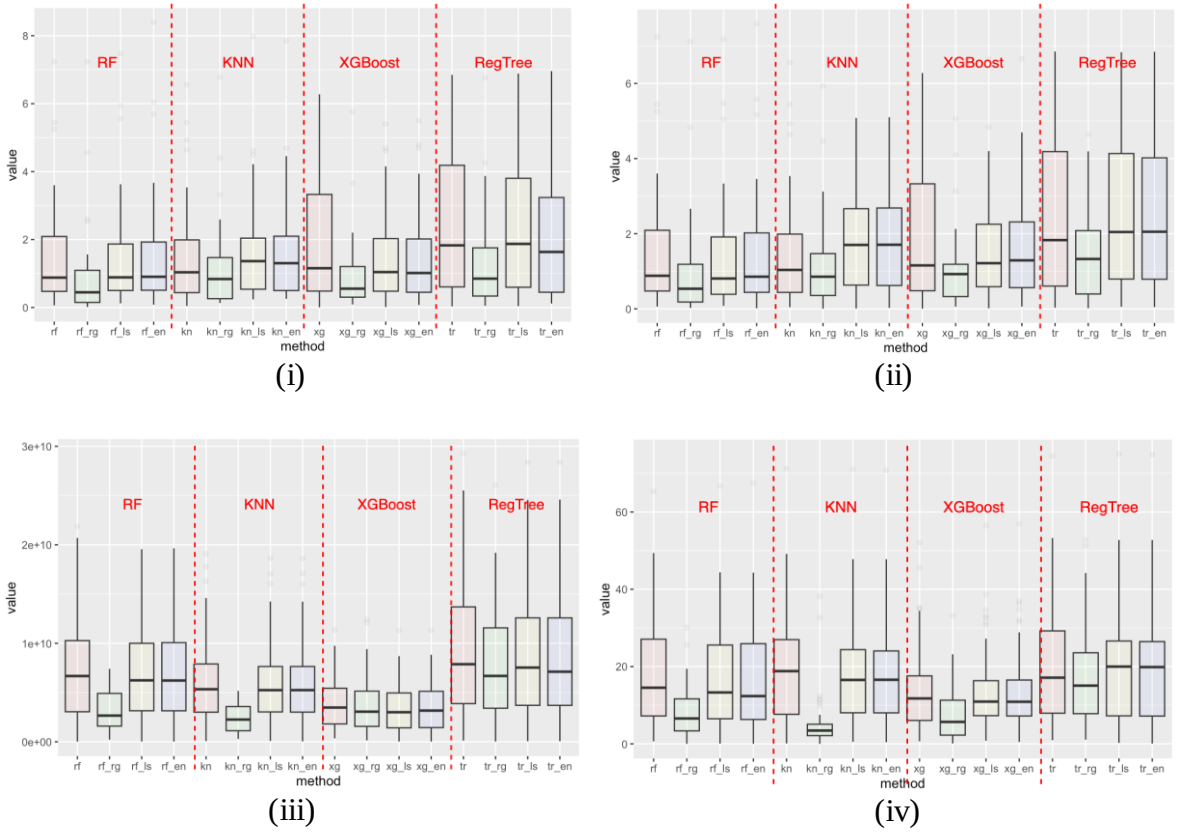
Dapat diketahui dari Tabel 2, bahwa keseluruhan data yang diteliti mengandung multikolinearitas. Kemudian dari hasil *tuning* yang dilakukan, diperoleh nilai-nilai hyperparameter pada Tabel 3.

Tabel 3. Hasil *tuning hyperparameter* masing-masing model

Metode regularisasi	**Model regresi**	**Hyperparameter**	**Data**			
			i	**ii**	**iii**	**iv**
Ridge	-	lambda	34.125	32.585	0.0001	237.00
LASSO	-	lambda	0.0973	0.0386	0.0001	0.2448
Elastic Net	-	lambda	0.1692	0.0405	0.0000	0.3381
		alpha	0.5	0.9	0.9	0.7
Ridge	Random Forest	mtry	4	4	4	2
		trees	937	937	796	11
		min_n	7	7	4	9
	kNN	k	8	4	4	4
	XGBoost	eta	0.1	0.5	0.1	0.1
		nrounds	75	50	75	75
		max_depth	5	1	3	3
	Regression Tree	maxdepth	2	2	3	2
LASSO	Random Forest	mtry	4	6	9	4
		trees	1385	1896	524	189
		min_n	9	4	19	9
	kNN	k	1	1	4	10
	XGBoost	eta	0.1	0.5	0.1	0.1
		nrounds	75	75	75	75
		max_depth	1	1	1	1
	Regression Tree	maxdepth	2	2	3	2
Elastic Net	Random Forest	mtry	8	9	9	4
		trees	107	1854	524	187
		min_n	21	12	19	9
	kNN	k	1	1	4	10
	XGBoost	eta	0.1	0.1	0.5	0.1
		nrounds	75	50	75	75
		max_depth	1	1	1	1
	Regression Tree	maxdepth	2	2	3	2

3.1 Hasil

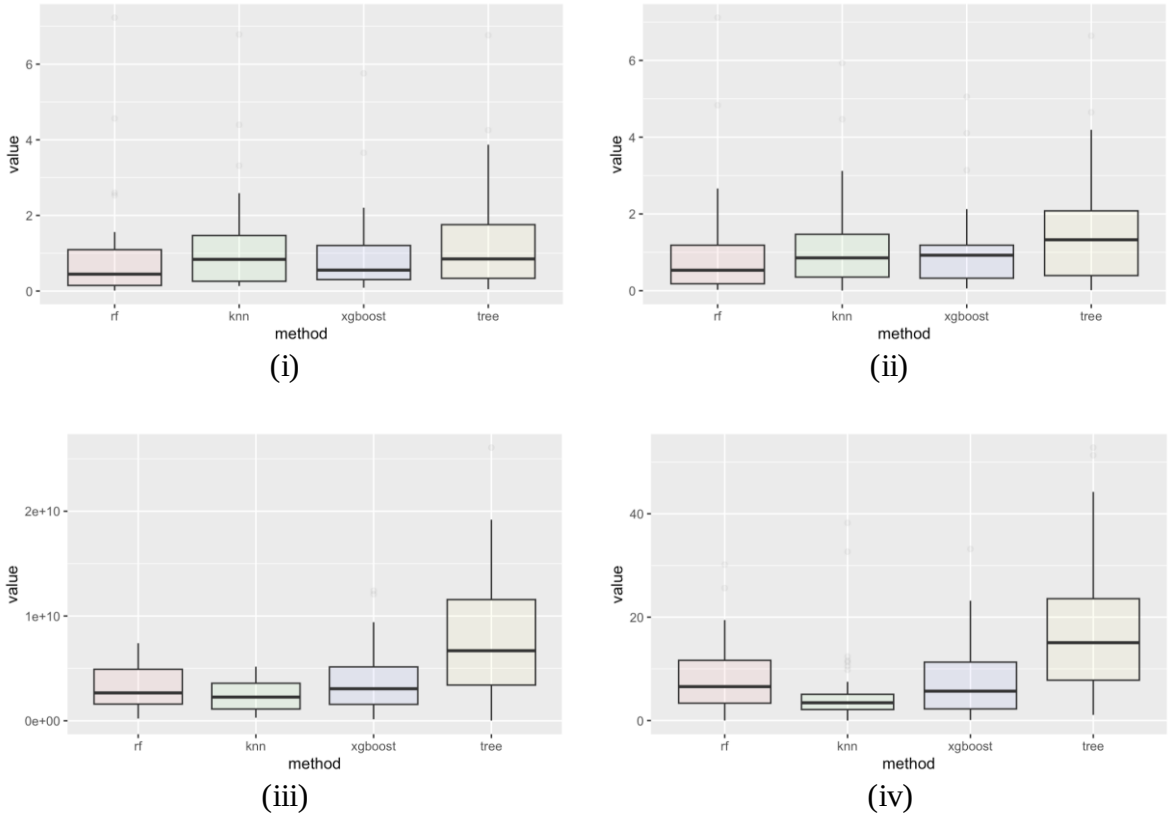
Adapun dari hasil tuning yang dilakukan, dilakukan pemodelan Random Forest, kNN, XGBoost, dan Regression Tree dengan tiga teknik regularisasi (*ridge,* LASSO, dan Elastic Net) menggunakan 1000 kali pengulangan sehingga diperoleh sebaran nilai-nilai RMSE masing-masing model pada Gambar 4.



Gambar 4. Hasil skor RMSE pada pemodelan regresi pada 4 data yang berbeda: (i) data empiris tanpa standardisasi, (ii) data empiris dengan standardisasi, (iii) data simulasi berkorelasi bersebaran seragam, dan (iv) data simulasi berkorelasi bersebaran normal

Diagram kotak garis (*boxplot*) yang tampak pada Gambar 4 secara urut menampakkan metode regularisasi model yang digunakan, antara lain secara berurutan dari kiri: merah (tanpa regularisasi), hijau (dengan *ridge*), kuning (dengan LASSO), dan biru (dengan Elastic Net). Dari plot sebaran RMSE yang ditampilkan pada keempat gambar di atas, terlihat bahwa tidak ada perbedaan hasil antara data yang dilakukan standardisasi dan yang tidak. Kemudian, pada masing-masing data, tampak jelas bahwa model regresi dengan Random Forest, kNN, XGBoost, dan Regression Tree menghasilkan performa yang paling baik ketika dilakukan regularisasi dengan *ridge*. Hal ini terlihat dari *boxplot* yang lebih kecil dan lebih rendah daripada *boxplot* model lainnya pada keseluruhan data yang digunakan.

Hasil pemodelan dengan empat model pembelajaran mesin dengan regularisasi *ridge* terlihat pada gambar berikut.



Gambar 5. Hasil pemodelan pembelajaran mesin dengan regularisasi *ridge* pada 4 data berbeda: (i) data empiris tanpa standardisasi, (ii) data empiris dengan standardisasi, (iii) data simulasi berkorelasi bersebaran seragam, dan (iv) data simulasi berkorelasi bersebaran normal

Dari Gambar 5 atas terlihat bahwa pada empiris (i dan ii), model XGBoost dan Random Forest menghasilkan RMSE yang paling kecil namun tidak berbeda jauh dengan model kNN. Sedangkan pada data simulasi berkorelasi, diperoleh bahwa model kNN merupakan model yang paling akurat dan stabil dibandingkan dengan metode lain, serta model Random Forest dan XGBoost menghasilkan hasil RMSE yang mirip. Sementara model Regression Tree tidak disarankan berdasarkan hasil penelitian ini karena menghasilkan nilai RMSE paling besar dan yang paling tidak stabil.

3.2 Pembahasan

Hasil penelitian menunjukkan bahwa regularisasi model pembelajaran mesin dengan ridge memberikan hasil yang paling akurat dan paling stabil, terlihat dari *boxplot* sebaran RMSE model regularisasi *ridge* berbentuk paling pendek dan terletak paling rendah daripada *boxplot* model lainnya. Ridge memiliki karakter yang berbeda dengan LASSO, di mana pada metode *ridge*, pendugaan koefisien-koefisien model disusutkan hingga ke nilai yang sangat dekat dengan nol, sementara LASSO menyusutkan koefisien pendugaan hingga nol. Pada konteks menemukan model yang lebih *interpretable*, penggunaan LASSO lebih disarankan karena kompleksitas model dapat diminimalisir. Dalam konteks mencari model yang tinggi akurasi prediksi, *ridge* lebih unggul karena prediksi mungkin

membutuhkan kontribusi seluruh peubah sehingga model masih memuat seluruh peubah prediktor tanpa ada yang dibuang. Penelitian [17] juga menyatakan bahwa jika dua variabel prediktor adalah kolinear, penduga LASSO tidak menyusutkan kedua koefisien. Dengan demikian, LASSO tidak memiliki efek pengelompokan yang diinginkan, di mana dua variabel prediktor berkorelasi tinggi harus menarik koefisien pendugaan yang serupa (dan koefisien identik dalam nilai absolut jika keduanya berkorelasi sempurna).

Hasil analisis pada data yang berbeda menunjukkan bahwa Regression Tree memberikan hasil yang tidak lebih baik dibandingkan Random Forest, kNN, dan XGBoost. Regression Tree cenderung *overfit* pada data pelatihan, yang mana bekerja dengan baik pada data pelatihan tetapi buruk pada data pengujian. Sebaliknya, Random Forest, kNN, dan XGBoost menggunakan teknik *ensemble* untuk mengurangi *overfitting* dan meningkatkan performa generalisasi. Hasil ini sejalan dengan penelitian [24] yang menunjukkan bahwa Decision Tree menghasilkan akurasi lebih rendah dari model-model pembelajaran mesin yang lain (Regresi Logistik, kNN, SVM, Decision Tree, Random Forest, dan XGBoost).

4. Kesimpulan

Penelitian pemodelan data berdimensi tinggi dan mengandung multikolinearitas pada empat algoritma pembelajaran mesin yang meliputi Random Forest, K-Nearest Neighbor, XGBoost, dan Regression Tree. Sebelumnya, dilakukan regularisasi dengan regresi terpenalti *ridge*, LASSO, dan Elastic Net. Tidak ada perbedaan berarti antara data yang dilakukan standardisasi atau tidak, dan data bangkitan yang memiliki sebaran seragam atau normal. Namun, pada data empiris baik distandardisasi atau tidak, diperoleh bahwa model Regression Tree memberikan performa yang tidak lebih baik daripada model lain. Pada penelitian selanjutnya, perlu dilakukan penanganan data berkorelasi lainnya untuk diterapkan pada model pembelajaran mesin guna meningkatkan performanya utamanya pada data berkorelasi tinggi. Selain itu, dapat pula dibandingkan dengan metrik lainnya seperti MAE, MAPE, dan lain-lain untuk mendapatkan perbandingan model secara lebih komprehensif.

Daftar Pustaka

[1] P. C. Sen, M. Hajra, dan M. Ghosh, “Supervised classification algorithms in machine learning: A survey and review,” dalam *Advances in Intelligent Systems and Computing*, Springer Verlag, vo;. 937, pp. 99–111, 2020. [CrossRef]

[2] Jason Bell, “What is machine learning?,” dalam *Machine Learning: Hands-On for Developers and Technical Professionals*, 2 ed.John Wiley & Sons, Inc., pp. 1–14, 2020.

[3] D. G. Jenkins dan P. F. Quintana-Ascencio, "A solution to minimum sample size for regressions," *PLoS One*, vol. 15, no. 2, 2020. [CrossRef]

[4] J. P. A. Ioannidis, "Why most published research findings are false," dalam *Getting to Good: Research Integrity in the Biomedical Sciences*, Springer International Publishing, pp. 2–8, 2018. [CrossRef]

[5] M. J. Curtis *dkk.*, "Experimental design and analysis and their reporting: new guidance for publication in BJP," *Br J Pharmacol*, pp. 3461–3471, 2015. [CrossRef]

[6] H. J. Gotelli dan A. M. Ellison, "A primer of ecological statistics," *Biometrics*, vol. 62, no. 1, pp. 308–308, 2006. [CrossRef]

[7] E. J. Pedhazur dan L. P. Schmelkin, *Measurement, Design, and Analysis*. Psychology Press, 2013. [CrossRef]

[8] C. Giraud, *Introduction to High-Dimensional Statistics*, 2 ed. New York: Chapman and Hall/CRC, 2021. [CrossRef]

[9] M. Brimacombe, "High-dimensional data and linear models: A review," *Open Access Med Stat*, pp. 17, 2014. [CrossRef]

[10] J. Y. Le Chan *dkk.*, "Mitigating the multicollinearity problem and its machine learning approach: A review," *Mathematics*, vol. 10, no. 8, pp. 1–17, 2022. [CrossRef]

[11] S. H. Bae, J. Y. Choi, J. Qiu, dan G. C. Fox, "Dimension reduction and visualization of large high-dimensional data via interpolation," dalam *HPDC 2010 - Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing*, 2010, pp. 203–214. [CrossRef]

[12] D. Engel, L. Hüttenberger, dan B. Hamann, "A survey of dimension reduction methods for high-dimensional data analysis and visualization," dalam *OpenAccess Series in Informatics*, pp. 135–149, 2012. [CrossRef]

[13] T. Sirimongkolkasem dan R. Drikvandi, "On regularisation methods for analysis of high dimensional data," *Annals of Data Science*, vol. 6, no. 4, pp. 737–763, 2019. [CrossRef]

[14] J. Kolluri, V. K. Kotte, M. S. B. Phridviraj, dan S. Razia, "Reducing overfitting problem in machine learning using novel L1/4 regularization method," dalam *Proceedings of the 4th International Conference on Trends in Electronics and Informatics, ICOEI 2020*, Institute of Electrical and Electronics Engineers Inc., Jun 2020, pp. 934–938, 2020. [CrossRef]

[15] A. E. Hoerl dan R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970. [CrossRef]

[16] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *J. R. Statist. Soc. B*, vol. 58, no. 1, pp. 267–288, 1996. [CrossRef]

[17] H. Zou dan T. Hastie, "Regularization and variable selection via the elastic net," *J. R. Statist. Soc. B*, vol. 67, no. 2, pp. 301–320, 2005. [CrossRef]

[18] United Nations Development Programme, *Human development report 1995*. Oxford University Press for the United Nations Development Programme (UNDP), 1995.

[19] Z. Farhadi, H. Bevrani, dan M.-R. Feizi-Derakhshi, "Improving random forest algorithm by selecting appropriate penalized method," *Commun Stat Simul Comput*, pp. 1–16, 2022. [CrossRef]

[20] A. Tiffin, "Seeing in the dark: A machine-learning approach to nowcasting in Lebanon IMF working paper middle east and Central Asia department seeing in the dark: A machine-learning approach to nowcasting in Lebanon," 2016. [CrossRef]

[21] E. G. Dada, D. O. Oyewola, S. B. Joseph, O. Emebo, dan O. O. Oluwagbemi, "Ensemble machine learning for monkeypox transmission time series forecasting," *Applied Sciences (Switzerland)*, vol. 12, no. 23, 2022. [CrossRef]

[22] Badan Pusat Statistik, "Indeks Pembangunan Manusia 2022," 2022. [WebLink]

[23] Y. Haitovsky, "Multicollinearity in regression analysis: Comment," *Rev Econ Stat*, vol. 51, no. 4, pp. 486–489, 1969. [CrossRef]

[24] A. Yilmaz, A. A. Demircali, S. Kocaman, dan H. Uvet, "Comparison of deep learning and traditional machine learning techniques for classification of pap smear images," *arXiv preprint arXiv:2009.06366.*, 2020. [GreenVersion]

Lampiran

Variabel	Makna	Satuan
Y	Indeks Pembangunan Manusia	-
X1	PDRB berdasarkan Harga Konstan 2010 Perkapita (2022)	ribu Rupiah
X2	Jumlah Pernikahan (2019)	Kasus
X3	Jumlah Perusahaan Mikro (2021)	Unit
X4	Jumlah Perusahaan Kecil (2021)	Unit
X5	Rasio Gini (2022)	-
X6	Presentase Rumah Tangga yang Memiliki Akses Sanitasi Layak (2021)	Persen
X7	Produksi Daging Sapi (2022)	Ton
X8	Produksi Ayam Buras (2022)	Ton
X9	Presentase Rumah Tangga yang Memiliki Akses Air Minum Layak (2021)	Persen
X10	Nilai Pembelian (2018)	juta Rupiah
X11	Nilai Penjualan (2018)	juta Rupiah
X12	Unmet Need Pelayanan Kesehatan (2022)	Persen
X13	Presentase Tenaga Kerja Formal (2022)	Persen
X14	Proporsi Lapangan Kerja Informal (2022)	Persen
X15	Kepadatan Penduduk (2022)	jiwa/km2
X16	Indeks Kebahagiaan (2021)	Persen
X17	Indeks Demokrasi (2020)	-
X18	Tingkat Setengah Pengangguran (2022)	Persen
X19	Jumlah Tenaga Kerja (2021)	Orang
X20	Jumlah Desa Tertinggal (2021)	Desa
X21	Persentase Penduduk yang Mempunyai Keluhan Kesehatan dan Berobat Jalan Selama Sebulan Terakhir (2021)	Persen
X22	Jumlah Koperasi Aktif (2021)	Unit
X23	Jumlah Penduduk Miskin (2022)	ribu
X24	Presentase Penduduk Miskin (2022)	Persen
X25	Volume Usaha Koperasi (2021)	juta Rupiah
X26	Pengeluaran untuk Tenaga Kerja (2021)	Rupiah
X27	Kapasitas Terpasang Pembangkit Listrik (2021)	mega Watt
X28	Proporsi Individu Pengguna Internet (2019)	Persen
X29	Jumlah Desa Penerima Sinyal 4G (2021)	Desa
X30	Listrik yang Didistribusikan (2021)	GWh
X31	Prevalensi Tekanan Darah Tinggi (2018)	Persen
X32	Indeks Khusus Penanganan Stunting (2019)	Persen
X33	Nilai Tambah (Harga Pasar) Mikro (2021)	juta Rupiah
X34	Nilai Tambah (Harga Pasar) Kecil (2021)	juta Rupiah
X35	Upah Minimum Provinsi (2020)	Rupiah

Variabel	Makna	Satuan
X36	Jumlah Desa/Kelurahan yang Memiliki Fasilitas Sekolah - SD (2021)	Desa
X37	Jumlah Desa/Kelurahan yang Memiliki Fasilitas Sekolah - SMP (2021)	Desa
X38	Jumlah Desa/Kelurahan yang Memiliki Fasilitas Sekolah - SMA (2021)	Desa
X39	Jumlah Desa/Kelurahan yang Memiliki Fasilitas Sekolah - SMK (2021)	Desa
X40	Jumlah Desa/Kelurahan yang Memiliki Fasilitas Sekolah - Uni (2021)	Desa
X41	Jumlah Desa/Kelurahan yang Memiliki Rumah Sakit (2021)	Desa
X42	Jumlah Desa/Kelurahan yang Memiliki Apotek (2021)	Desa
X43	Presentase Penyelesaian Pendidikan - SD (2022)	Persen
X44	Presentase Penyelesaian Pendidikan - SMP (2022)	Persen
X45	Presentase Penyelesaian Pendidikan - SMA (2022)	Persen
X46	Jumlah Rumah Tangga Penerima Manfaat Bantuan Sosial Pangan (2022)	Pasang
X47	Anggaran Penerima Manfaat Bantuan Sosial Pangan (2022)	ribu Rupiah
X48	Prevalensi Gizi Buruk Balita (0-59 Bulan) (2018)	Persen
X49	Presentase Penduduk Pengguna Komputer (2021)	Persen
X50	Jumlah Penduduk (2022)	ribu
X51	Laju Pertumbuhan Penduduk Per Tahun (2022)	-
X52	Presentase Penduduk (2022)	Persen
X53	Rata-Rata Lama Sekolah Penduduk Umur ≥ 15 Tahun (2022)	Tahun
X54	Angka Partisipasi Kasar - PT (2022)	Persen
X55	Presentase Penduduk Penerima Jaminan Kesehatan Pemerintah (2022)	Persen
X56	Presentase Anak Usia 10-17 Tahun yang Bekerja (2022)	Persen
X57	Rata-rata Konsumsi Protein Per Hari (2016)	gram
X58	Jumlah Kejahatan yang Dilaporkan (2021)	Kasus
X59	Risiko Penduduk Terjadi Tindak Pidana per 100.000 Penduduk (2021)	Penduduk
X60	Persentase Penyelesaian Tindak Pidana (2021)	Persen
X61	Presentase Rumah Tangga dengan Fasilitas Buang Air Sendiri (2021)	Persen
X62	Tingkat Pengangguran Terbuka (2022)	Persen
X63	Tingkat Partisipasi Angkatan Kerja (2022)	Persen
X64	Tingkat Kelahiran Perempuan 15-19 Tahun (2017)	Persen
X65	Presentase Penduduk yang Memiliki dan Menguasai Ponsel (2021)	Persen
X66	Presentase Perempuan Berada di Level Manajerial (2022)	Persen
X67	Presentase Perokok 15 Tahun Ke Atas (2022)	Persen
X68	Angka Melek Huruf Penduduk 15 Tahun Ke Atas (2022)	Persen
X69	Angka Kelahiran Kasar (2020)	Persen
X70	Rata-rata Pendapatan Bersih Pekerja Bebas (2021)	ribu Rupiah
X71	Rata-rata Pendapatan Bersih Pekerja Mandiri (2021)	ribu Rupiah

Variabel	Makna	Satuan
X72	Angka Kematian Bayi (2020)	Persen
X73	Proporsi Penduduk Dengan Asupan Kalori Minimum Di Bawah 1400 Kkal/Kapita/Hari (2021)	Persen
X74	Proporsi Penduduk yang Berada di bawah 50 persen Median Pendapatan (2022)	Persen
X75	Tingkat Partisipasi Pembelajaran Terorganisir (1 Tahun Sebelum SD) (2022)	
X76	Rata-rata Banyak Anggota Keluarga (2016)	Orang
X77	Indeks Triwulanan Balas Jasa Pekerja Tetap dan Upah Pekerja Harian Konstruksi	-
X78	Persentase Perempuan Pernah Kawin Berusia15-49 Tahun yang Proses Melahirkan Terakhirnya di Fasilitas Kesehatan (2022)	Persen
X79	Proporsi Rumah Tangga dengan Hunian Milik Sendiri (2021)	Persen
X80	Persentase Rumah Tangga yang Masih Mempraktikkan Buang Air Besar Sembarangan (BABS) di Tempat Terbuka (2022)	Persen
X81	Persentase Populasi Penduduk > 10 Tahun yang Belum Pernah Sekolah	Persen
X82	Proporsi Rumah Tangga Yang Memiliki Fasilitas Cuci Tangan Dengan Sabun Dan Air (2022)	Persen
X83	Angka Partisipasi Kasar - PAUD (2022)	-
X84	Persentase Pengetahuan Dan Pemahaman Pasangan Usia Subur (PUS) Tentang Metode Kontrasepsi (2017)	Persen
X85	Banyak Perguruan Tinggi Negeri dan Swasta (2022)	Unit
X86	Banyak Tenaga Pendidik Perguruan Tinggi Negeri dan Swasta (2022)	Orang
X87	Banyak Mahasiswa Perguruan Tinggi Negeri dan Swasta (2022)	Orang
X88	Proporsi Remaja Dan Dewasa Usia 15-24 Tahun Dengan Keterampilan Teknologi Informasi Dan Komputer (2022)	Persen
X89	Jumlah Kabupaten/Kota Endemis Filariasis yang Mencapai Eliminasi (2018)	Kab/Kota
X90	Angka Pemakaian Kontrasepsi (CPR) Semua Cara Pada Pasangan Usia Subur Usia 15-49 Tahun Yang Pernah Kawin (40% Bawah) (2019)	Persen
X91	Indeks Pemberdayaan Gender (2022)	-
X92	Sumbangan Pendapatan Perempuan (2022)	Persen
X93	Indeks Pembangunan Gender (2022)	-
X94	Rata-rata Lama Sekolah (2022)	Tahun
X95	Umur Harapan Hidup Saat Lahir (2022)	Tahun
X96	Harapan Lama Sekolah (2022)	Tahun
X97	Indeks Keparahan Kemiskinan (2022)	-
X98	Indeks Kedalaman Kemiskinan (2022)	-
X99	Presentase Anak 12-23 Bulan yang Menerima Imunisasi Dasar Lengkap (2022)	Persen
X100	Dependency Ratio 2035	-