



Moses Sinanta P.W.J.¹

PREDIKSI HARGA MOBIL MENGGUNAKAN LINEAR REGRESSION, RIDGE REGRESSION DAN LASSO REGRESSION

Abstrak

Prediksi harga mobil merupakan aspek penting dalam industri otomotif, membantu produsen, dealer dan konsumen dalam menentukan nilai pasar kendaraan secara objektif. Berbagai faktor seperti merek, tahun produksi, jenis bahan bakar, dan lain-lain dapat memengaruhi harga mobil, sehingga diperlukan model prediktif yang akurat. Penelitian ini bertujuan untuk menganalisis performa Linear Regression, Ridge Regression dan Lasso Regression dalam memprediksi harga mobil. Data yang digunakan diproses melalui One Hot Encoding dan Standard Scaler, kemudian dibagi menjadi data training dan testing dengan rasio 80-20. Hasil penelitian menunjukkan bahwa Lasso Regression dengan $\alpha = 0.2$ memberikan performa terbaik dengan MSE sebesar 4188.815 dan R^2 score sebesar 99.954%. Penelitian ini membuktikan bahwa model regresi yang digunakan mampu menunjukkan performa prediksi harga mobil dengan akurasi tinggi, sehingga dapat digunakan sebagai dasar dalam analisis harga kendaraan.

Kata Kunci: Prediksi Harga Mobil, Linear Regression, Ridge Regression, Lasso Regression.

Abstract

Car price prediction is an important aspect in the automotive industry, helping manufacturers, dealers and consumers in objectively determining the market value of vehicles. Various factors such as brand, year of production, type of fuel, etc. Can affect the price of a car, so an accurate predictive model is needed. This study aims to analyze the performance of Linear Regression, Ridge Regression and Lasso Regression in predicting car prices. The data used is processed through One Hot Encoding and Standard Scaler, then divided into training and testing data with a ratio of 80-20. The results showed that Lasso Regression with $\alpha = 0.2$ gave the best performance with an MSE of 4188.815 and an R^2 score of 99.954%. This study proves that the regression model used is able to show the performance of car price prediction with high accuracy, so it can be used as a basis for vehicle price analysis.

Keywords: Car price prediction, Linear Regression, Ridge Regression, Lasso Regression.

PENDAHULUAN

Prediksi harga mobil memainkan peran penting dalam industri otomotif, baik bagi pihak produsen, dealer maupun konsumen (Dewi et al., 2024). Bagi pihak produsen, prediksi harga dapat membantu mengidentifikasi tren pasar. Bagi dealer, prediksi harga dapat membantu menetapkan harga yang kompetitif, sedangkan bagi konsumen, prediksi harga dapat membantu menentukan apakah suatu harga wajar atau tidak.

Harga mobil sendiri dipengaruhi oleh berbagai faktor, termasuk merek, model, tahun produksi, ukuran mesin, jenis bahan bakar dan lain-lain. Beberapa faktor ini memiliki hubungan langsung dengan harga, sebagai contoh tahun produksi yang lebih baru cenderung meningkatkan nilai jual (Wilianto et al., 2024). Oleh karena itu, suatu pemodelan yang tepat diperlukan untuk mengidentifikasi hubungan antar variabel tersebut dengan harga mobil.

Regresi merupakan salah satu metode statistik yang dapat digunakan untuk memprediksi harga mobil dengan mengidentifikasi hubungan antara fitur-fitur kendaraan dan nilai jualnya (Aji Saputra et al., 2024; Kusuma & Hidayat, 2024). Dalam konteks ini, regresi memungkinkan pembuatan model matematis yang dapat memperkirakan harga berdasarkan fitur input. Dengan

¹ Informatika, Universitas Sebelas Maret
 email: mosessinanta@student.uns.ac.id

menggunakan data yang tepat, model regresi dapat dilatih untuk menghasilkan estimasi harga yang lebih akurat dibandingkan pendekatan tradisional.

Linear Regression (regresi linear) merupakan pendekatan dasar yang kerap kali digunakan dalam kasus prediksi (James et al., 2023). Namun model ini memiliki keterbatasan ketika terdapat multikolinearitas (multicollinearity) atau fitur yang tidak relevan. Oleh karena itu, dikembangkan metode regresi lain seperti Ridge Regression (Arashi et al., 2021) dan Lasso Regression (Lee et al., 2022), yang dapat membantu meningkatkan akurasi model dengan mengontrol kompleksitas.

Penelitian ini bertujuan untuk membandingkan performa Linear Regression, Ridge Regression dan Lasso Regression dalam memprediksi harga mobil. Dengan mengevaluasi keakuratan model menggunakan metrik evaluasi seperti MSE dan R^2 score, penelitian ini akan mengidentifikasi model mana yang memberikan hasil paling optimal. Selain itu, penelitian ini juga akan menganalisis bagaimana regularisasi dalam Ridge Regression dan Lasso Regression memengaruhi pemilihan fitur dalam prediksi harga mobil.

Struktur paper ini dapat dijelaskan sebagai berikut. Bagian Dasar Teori menjelaskan konsep regresi dan metrik evaluasi yang digunakan dalam penelitian ini, kemudian dilanjutkan dengan bagian Metodologi yang menjelaskan proses pengolahan data, pemilihan model dan parameter serta evaluasi performa model. Bagian Hasil dan Evaluasi menjelaskan perbandingan model berdasarkan metrik evaluasi yang digunakan, kemudian dilanjutkan dengan bagian Kesimpulan yang merangkum temuan utama dan memberikan saran untuk penelitian lebih lanjut.

Linear Regression

Linear Regression merupakan metode statistik yang digunakan untuk memodelkan hubungan antara satu atau lebih variabel independen dengan variabel dependen [3]. Model ini bekerja dengan mencari garis terbaik yang dapat mewakili pola hubungan antara input dan output berdasarkan data. Dalam kasus prediksi, Linear Regression digunakan untuk memperkirakan suatu nilai, sebagai contoh harga, berdasarkan berbagai variabel.

Persamaan dasar Linear Regression dapat dituliskan sebagai:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n + \epsilon$$

dimana y merupakan variabel dependen, β_0 adalah intercept (nilai variabel dependen ketika semua variabel independen bernilai 0), β_i adalah koefisien regresi untuk setiap variabel independen x_i dan ϵ merupakan nilai error (kesalahan) dalam prediksi.

Agar model menghasilkan prediksi yang optimal, cost function digunakan untuk mengukur seberapa baik model memprediksi data. Dalam Linear Regression, cost function yang umum digunakan adalah Mean Squared Error (MSE), yang menghitung rata-rata kuadrat selisih antara nilai aktual dan nilai prediksi. Hal ini berarti semakin kecil nilai MSE, semakin baik model dalam menyesuaikan data.

Persamaan dasar cost function $J(\beta)$ dalam Linear Regression dapat dituliskan sebagai:

$$J(\beta) = \frac{1}{m} \sum_{i=1}^m (y_i - \hat{y}_i)^2$$

dimana m merupakan jumlah data dan \hat{y} merupakan nilai prediksi.

Ridge Regression

Ridge Regression merupakan pengembangan dari Linear Regression yang menambahkan regularisasi untuk mengatasi masalah multikolinearitas (multicollinearity) (Dar et al., 2023) dan overfitting (Tsigler & Bartlett, 2023). Multikolinearitas merupakan kasus ketika dua atau lebih variabel independen dalam model regresi sangat berkorelasi satu sama lain, yang menyebabkan model kesulitan dalam menentukan pengaruh masing-masing variabel (Tsagris & Pandis, 2021). Overfitting merupakan kasus ketika model terlalu menyesuaikan diri dengan data training, sehingga mengurangi kemampuan untuk menggeneralisasi data testing atau data baru (Montesinos López et al., 2022).

Regularisasi yang digunakan oleh Ridge Regression adalah L2 Regularization. Metode ini menambahkan penalti terhadap nilai besar dari koefisien regresi β sehingga model tidak terlalu dipengaruhi oleh fitur yang memiliki hubungan kuat dengan variabel target. Dengan demikian,

Ridge Regression membantu dalam membuat model yang lebih stabil dan mampu melakukan generalisasi yang lebih baik.

Persamaan cost function $J(\beta)$ dalam Ridge Regression dapat dituliskan sebagai:

$$J(\beta) = \frac{1}{m} \sum_{i=1}^m (y_i - \hat{y}_i)^2 + \alpha \sum_{j=1}^n \beta_j^2$$

dimana α merupakan parameter regularisasi yang mengontrol besar penalti.

Lasso Regression

LASSO (Least Absolute Shrinkage and Selection Operator) atau Lasso Regression merupakan pengembangan lain dari Linear Regression yang menggunakan regularisasi tidak hanya untuk mengatasi overfitting, namun juga dapat melakukan seleksi fitur (feature selection) (Lee et al., 2022; Li et al., 2021). Feature selection merupakan proses pemilihan fitur yang paling relevan untuk digunakan dalam model machine learning untuk meningkatkan akurasi dan mengurangi kompleksitas. Metode ini membantu menghilangkan fitur yang tidak penting atau redundant, sehingga mencegah overfitting dan meningkatkan generalisasi model terhadap data baru.

Regularisasi yang digunakan oleh Lasso Regression adalah L1 Regularization. Metode ini menambahkan penalti terhadap jumlah absolut koefisien regresi β , hingga dapat mengecilkan beberapa koefisien hingga bernilai 0. Hal ini memastikan hanya fitur yang paling penting yang dipertahankan dalam model, yang membuat Lasso Regression lebih efektif dalam menangani dataset dengan fitur banyak fitur yang tidak semuanya relevan.

Persamaan cost function $J(\beta)$ dalam Lasso Regression dapat dituliskan sebagai:

$$J(\beta) = \frac{1}{m} \sum_{i=1}^m (y_i - \hat{y}_i)^2 + \alpha \sum_{j=1}^n |\beta_j|$$

One Hot Encoding

One Hot Encoding merupakan sebuah teknik transformasi data kategori menjadi bentuk numerik sehingga data dapat digunakan dalam model machine learning (Yu et al., 2022). Metode ini bekerja dengan mengonversi setiap kategori unik dalam suatu fitur menjadi vektor biner. Hal ini akan menghasilkan kolom terpisah yang merepresentasikan setiap kategori. Sebagai contoh, jika terdapat data kategori seperti Tabel 1 dibawah:

Tabel 1. Contoh tabel data kategori.

Day	Outlook
D1	Sunny
D3	Overcast
D5	Rain

maka One Hot Encoding akan merubah data tersebut menjadi kolom terpisah berjumlah sesuai dengan jumlah kategori seperti Tabel 2 dibawah:

Tabel 2. Contoh tabel data kategori (Tabel 1) setelah dilakukan One Hot Encoding.

Day	Sunny	Overcast	Rain
D1	1	0	0
D3	0	1	0
D5	0	0	1

Standard Scaler

Standard Scaler merupakan sebuah teknik normalisasi yang digunakan untuk menskalakan fitur agar memiliki nilai rata-rata 0 dan standar deviasi 1. Proses ini penting untuk algoritma machine learning yang sensitif terhadap skala data, karena perbedaan skala antar fitur dapat menyebabkan model lebih condong pada fitur dengan nilai besar (Ahsan et al., 2021). Standard Scaler memastikan bahwa setiap fitur memiliki kontribusi yang seimbang dalam model.

Persamaan dasar Standard Scaler untuk data x dapat dituliskan sebagai:

$$x' = \frac{x - \mu_x}{\sigma_x}$$

dimana x' merupakan fitur yang dinormalisasi, μ_x merupakan rata-rata dari x dan σ_x merupakan standar deviasi dari x .

Mean Squared Error

Mean Squared Error (MSE) merupakan metrik evaluasi yang digunakan untuk mengukur seberapa besar rata-rata kesalahan kuadrat antara nilai sebenarnya dan nilai prediksi (Hodson et al., 2021). MSE memberikan gambaran seberapa berbeda prediksi model dari nilai yang sebenarnya. Nilai MSE yang lebih kecil menunjukkan bahwa model menunjukkan hasil yang lebih akurat.

Persamaan dasar untuk MSE dapat dituliskan sebagai:

$$MSE = \frac{1}{m} \sum_{i=1}^m (y_i - \hat{y}_i)^2$$

Dalam Linear Regression, nilai dari cost function bernilai sama dengan MSE.

R^2 score

R^2 score atau coefficient of determination merupakan metrik evaluasi yang digunakan untuk mengukur seberapa baik model menjelaskan variabilitas dalam data (Chicco et al., 2021). Nilai R^2 score berkisar $[0,1]$, dimana nilai 0 menunjukkan bahwa model tidak lebih baik dibandingkan dengan rata-rata dari nilai target. Sebaliknya, nilai 1 menunjukkan bahwa model memiliki performa sempurna dalam menjelaskan hubungan antara variabel independen dan variabel dependen.

Persamaan dasar untuk R^2 score dapat dituliskan sebagai berikut:

$$R^2 \text{ score} = 1 - \frac{\sum_{i=1}^m (y_i - \hat{y}_i)^2}{\sum_{i=1}^m (y_i - \mu_y)^2}$$

dimana μ_y merupakan rata-rata dari y . R^2 score juga dapat didefinisikan sebagai jumlah bagi Residual Sum of Squares (RSS) dengan Total Sum of Squares (TSS).

METODE

Dataset yang digunakan dalam penelitian ini merupakan dataset “Car Price Dataset” (Tech, 2025). Dataset ini mencakup 10,000 entri informasi mobil dan estimasi harga mereka. Informasi yang disediakan mencakup brand, model, tahun produksi, ukuran mesin, jenis bahan bakar, jenis transmisi, jarak tempuh, jumlah pintu dan jumlah pemilik sebelumnya.

Sebelum model diterapkan, dilakukan beberapa tahap pre-processing untuk memastikan data dalam format yang diinginkan. One Hot Encoding digunakan untuk mengonversi fitur kategori seperti brand, model, jenis bahan bakar, dan lain-lain. Kemudian, Standard Scaler juga diterapkan pada fitur numerik seperti tahun produksi, ukuran mesin, jarak tempuh, dan lain-lain.

Setelah pre-processing, dataset kemudian dibagi menjadi data training dan data testing menggunakan rasio 80-20. Hal ini berarti sebanyak 80% data digunakan sebagai data training dan sebanyak 20% data digunakan sebagai data testing. Pembagian ini memastikan bahwa model yang dihasilkan dapat melakukan generalisasi dengan baik terhadap data baru.

Tiga jenis model regresi yang digunakan dalam penelitian ini merupakan Linear Regression, Ridge Regression dan Lasso Regression. Untuk Ridge Regression dan Lasso Regression, parameter α divariasikan dalam rentang $[0,0,1,0]$ dengan step sebesar 0.1. Hal ini diterapkan untuk mengamati dampak regularisasi terhadap performa model.

Evaluasi model kemudian dilakukan dengan menggunakan dua metrik evaluasi, yaitu MSE dan R^2 score. MSE digunakan untuk mengukur rata-rata kesalahan kuadrat antara nilai sebenarnya dengan nilai prediksi. R^2 score digunakan untuk menilai seberapa baik model menjelaskan variasi dalam data.

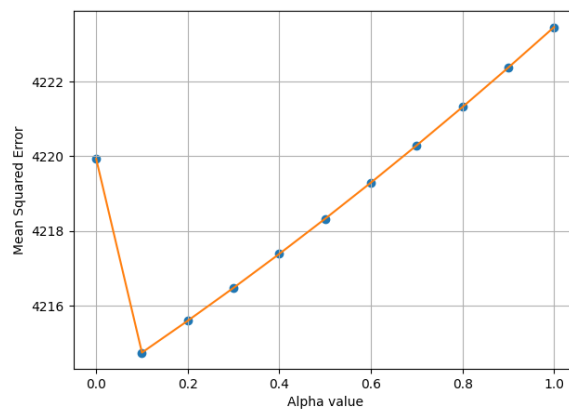
HASIL DAN PEMBAHASAN

Hasil perbandingan Linear Regression, Ridge Regression dan Lasso Regression ditampilkan dalam Gambar 1 dan Tabel 3.

Model Linear Regression menunjukkan performa yang baik dengan nilai MSE sebesar 4213.872 dan R^2 score sebesar 99.954%. Untuk memudahkan pembacaan dalam paper ini, nilai R^2 score akan dituliskan dalam bentuk persentase. Nilai R^2 score yang hampir mendekati 1 atau 100% menunjukkan bahwa model dapat memberikan prediksi yang sangat akurat terhadap harga mobil berdasarkan fitur-fitur yang tersedia.

Model Ridge Regression dengan nilai $\alpha = 1.0$ menunjukkan performa yang hampir serupa dengan Linear Regression. Model Ridge Regression menunjukkan performa dengan nilai MSE sebesar 4223.459 dan R^2 score sebesar 99.954%. Hal ini menunjukkan bahwa Ridge Regression memiliki performa yang sedikit lebih rendah dibandingkan Linear Regression dengan nilai $\alpha = 1.0$.

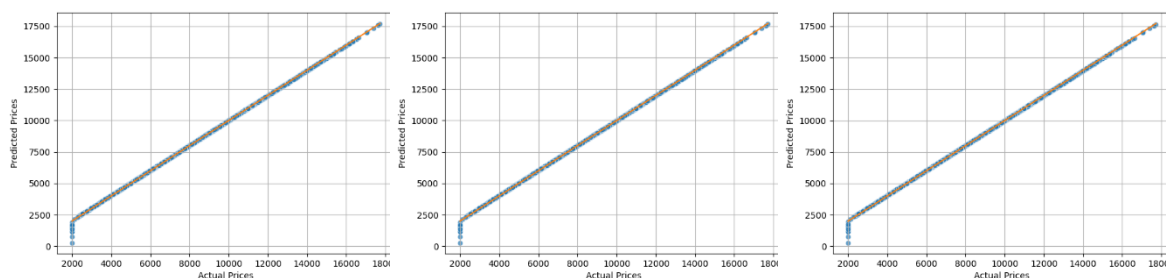
Pengujian dengan berbagai nilai α menunjukkan bahwa Ridge Regression menunjukkan performa terbaik pada nilai $\alpha = 0.1$. Model ini menunjukkan performa dengan nilai MSE sebesar 4214.745 dan R^2 score sebesar 99.954%. Pengaruh nilai α terhadap MSE untuk Ridge Regression ditunjukkan dalam Gambar 1.



Gambar 1. Pengaruh nilai α terhadap MSE untuk Ridge Regression

Tabel 3. Hasil prediksi dari Linear Regression, Ridge Regression dan Lasso Regression dengan parameter awal ($\alpha = 1.0$) dan parameter terbaik.

Model	Parameter	MSE	R^2 score (%)
Linear Regression	-	4213.872	99.954%
Ridge Regression	$\alpha = 1.0$	4223.459	99.954
	$\alpha = 0.1$	4214.745	99.954
Lasso Regression	$\alpha = 1.0$	4234.883	99.953
	$\alpha = 0.2$	4188.815	99.954

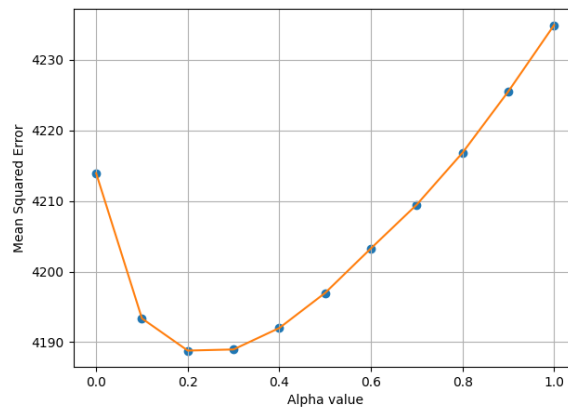


Gambar 2 Perbandingan regresi dari Linear Regression (kiri), Ridge Regression (tengah) dan Lasso Regression (kanan). Ketiga model menghasilkan prediksi yang hampir serupa.

Model Lasso Regression dengan nilai $\alpha = 1.0$ menunjukkan performa yang hampir serupa dengan Linear Regression. Model Lasso Regression menunjukkan performa dengan nilai

MSE sebesar 4234.883 dan R^2 score sebesar 99.953%. Hal ini menunjukkan bahwa Lasso Regression memiliki performa yang sedikit lebih rendah dibandingkan Linear Regression dengan nilai $\alpha = 1.0$.

Pengujian dengan berbagai nilai α menunjukkan bahwa Lasso Regression menunjukkan performa terbaik pada nilai $\alpha = 0.2$. Model ini menunjukkan performa dengan nilai MSE sebesar 4188.815 dan R^2 score sebesar 99.954%. Pengaruh nilai α terhadap MSE untuk Lasso Regression ditunjukkan dalam Gambar 3.



Gambar 3. Pengaruh nilai α terhadap MSE untuk Lasso Regression

SIMPULAN

Hasil penelitian ini menunjukkan bahwa Linear Regression, Ridge Regression dan Lasso Regression mampu memberikan prediksi harga mobil dengan tingkat akurasi yang cukup tinggi, dengan rata-rata model menunjukkan R^2 score sebesar 99.954%. Ridge Regression dengan $\alpha = 0.1$ dan Lasso Regression dengan $\alpha = 0.2$ memberikan hasil terbaik dalam kategori masing-masing, dengan MSE yang lebih rendah dibandingkan dengan nilai α lainnya. Secara keseluruhan, model Lasso Regression dengan nilai $\alpha = 0.2$ memberikan performa terbaik dibandingkan dengan model-model lainnya.

Untuk penelitian masa depan, dataset dapat diperluas dengan menambahkan variabel lain yang berpotensi memengaruhi harga mobil. Hal ini mencakup kondisi kendaraan, lokasi penjualan, teknologi terbaru yang diterapkan pada mobil, dan lain-lain. Selain itu, penggunaan dataset yang lebih besar dan beragam dapat membantu dalam meningkatkan generalisasi model terhadap berbagai kondisi pasar. Kemudian eksperimen menggunakan model lain, seperti Polynomial Regression atau metode machine learning lainnya, juga dapat menjadi arahan penelitian lebih lanjut untuk meningkatkan akurasi prediksi harga mobil.

DAFTAR PUSTAKA

- Ahsan, M., Mahmud, M., Saha, P., Gupta, K., & Siddique, Z. (2021). Effect of Data Scaling Methods on Machine Learning Algorithms and Model Performance. *Technologies*, 9(3), 52. <https://doi.org/10.3390/technologies9030052>
- Aji Saputra, M., Martanto, M., & Hayati, U. (2024). Estimasi Harga Mobil Bekas Toyota Yaris Menggunakan Algoritma Regresi Linier. *JATI (Jurnal Mahasiswa Teknik Informatika)*, 8(2), 1696–1701. <https://doi.org/10.36040/jati.v8i2.8344>
- Arashi, M., Roozbeh, M., Hamzah, N. A., & Gasparini, M. (2021). Ridge regression and its applications in genetic studies. *PLOS ONE*, 16(4), e0245376. <https://doi.org/10.1371/journal.pone.0245376>
- Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science*, 7, e623. <https://doi.org/10.7717/peerj-cs.623>
- Dar, I. S., Chand, S., Shabbir, M., & Kibria, B. M. G. (2023). Condition-index based new ridge regression estimator for linear regression model with multicollinearity. *Kuwait Journal of*

- Science, 50(2), 91–96. <https://doi.org/10.1016/j.kjs.2023.02.013>
- Dewi, B. E. S., Haikal, S., Sulistyowati, H. S., Fitriani, R., & Kuswando, D. P. (2024). Penerapan Machine Learning Menggunakan Algoritma Random Forest untuk Prediksi Harga Mobil Bekas. *Jurnal Teknologi Informasi Dan Digital*, 2(1), 20–31.
- Hodson, T. O., Over, T. M., & Foks, S. S. (2021). Mean Squared Error, Deconstructed. *Journal of Advances in Modeling Earth Systems*, 13(12). <https://doi.org/10.1029/2021MS002681>
- James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). Linear Regression (pp. 69–134). https://doi.org/10.1007/978-3-031-38747-0_3
- Kusuma, M. D. H., & Hidayat, S. (2024). Penerapan Model Regresi Linier dalam Prediksi Harga Mobil Bekas di India dan Visualisasi dengan Menggunakan Power BI. *Jurnal Indonesia: Manajemen Informatika Dan Komunikasi*, 5(2), 1097–1110. <https://doi.org/10.35870/jimik.v5i2.629>
- Lee, J. H., Shi, Z., & Gao, Z. (2022). On LASSO for predictive regression. *Journal of Econometrics*, 229(2), 322–349. <https://doi.org/10.1016/j.jeconom.2021.02.002>
- Li, F., Lai, L., & Cui, S. (2021). On the Adversarial Robustness of LASSO Based Feature Selection. *IEEE Transactions on Signal Processing*, 69, 5555–5567. <https://doi.org/10.1109/TSP.2021.3115943>
- Montesinos López, O. A., Montesinos López, A., & Crossa, J. (2022). Overfitting, Model Tuning, and Evaluation of Prediction Performance. In *Multivariate Statistical Machine Learning Methods for Genomic Prediction* (pp. 109–139). Springer International Publishing. https://doi.org/10.1007/978-3-030-89010-0_4
- Tech. (2025). Car Price Dataset. www.kaggle.com
- Tsagris, M., & Pandis, N. (2021). Multicollinearity. *American Journal of Orthodontics and Dentofacial Orthopedics*, 159(5), 695–696. <https://doi.org/10.1016/j.ajodo.2021.02.005>
- Tsigler, A., & Bartlett, P. L. (2023). Benign overfitting in ridge regression. *Journal of Machine Learning Research*, 24(123), 1–76.
- Wilianto, W., Yuliana, Y., Suwandhi, A., Jimmy, J., & Putra, J. (2024). Penerapan AI dalam Menentukan Harga Mobil Bekas Berdasarkan Tahun Perakitan. *Jurnal Minfo Polgan*, 13(1), 550–560. <https://doi.org/10.33395/jmp.v13i1.13728>
- Yu, L., Zhou, R., Chen, R., & Lai, K. K. (2022). Missing Data Preprocessing in Credit Classification: One-Hot Encoding or Imputation? *Emerging Markets Finance and Trade*, 58(2), 472–482. <https://doi.org/10.1080/1540496X.2020.1825935>