

PREDIKSI PENJUALAN ONLINE MENGGUNAKAN *RIDGE REGRESSION* BERBASIS *DATA HISTORIS*

Ayu Amelia Pertiwi^{1*}, Sarah Putri Syaifullah Nasution², Arnita³, Fanny Ramadhani⁴

Universitas Negeri Medan, Medan, Sumatera Utara, Indonesia¹

Email*: amelia020y@gmail.com

Universitas Negeri Medan, Medan, Sumatera Utara, Indonesia²

Email: srhptr1907@gmail.com

Universitas Negeri Medan, Medan, Sumatera Utara, Indonesia³

Email: arnita@unimed.ac.id

Universitas Negeri Medan, Medan, Sumatera Utara, Indonesia⁴

Email: fannyr@unimed.ac.id

(*) *Corresponding Author*

ABSTRAK

Perkembangan teknologi digital telah meningkatkan pembelian online, mengubah pola konsumsi, dan menciptakan kebutuhan akan prediksi penjualan yang akurat untuk pengambilan keputusan strategis. Namun, akurasi prediksi sering terhambat oleh masalah *multikolinearitas* antar fitur numerik dan risiko *overfitting* pada model regresi linier konvensional. Penelitian ini bertujuan mengatasi masalah tersebut dengan menerapkan *Ridge Regression* berbasis *regularisasi L2* untuk memprediksi total penjualan menggunakan data historis e-commerce. Metode yang digunakan meliputi pengumpulan data transaksi (541.909 records), preprocessing (penghapusan outlier dan missing values), serta pemodelan dengan *scikit-learn* yang dievaluasi menggunakan *MAE* (73.25), *RMSE* (129.80), dan koefisien determinasi (R^2). Hasil penelitian menunjukkan model mampu memprediksi total penjualan dengan akurasi memadai meskipun terdapat deviasi pada transaksi bernilai ekstrem. Secara praktis, implementasi *Ridge Regression* terbukti efektif mengurangi *overfitting* dan meningkatkan stabilitas prediksi, sehingga dapat menjadi acuan dalam perencanaan inventori dan strategi pemasaran. Temuan ini memperkuat perlunya pendekatan *machine learning* adaptif dalam analisis data penjualan di era digital.

Kata Kunci: *e-commerce*, *multikolinearitas*, prediksi penjualan, *regularisasi L2*, *Ridge Regression*

ABSTRACT

The development of digital technology has increased online purchases, changed consumption patterns, and created a need for accurate sales predictions for strategic decision-making. However, prediction accuracy is often hindered by multicollinearity issues among numeric features and the risk of overfitting in conventional linear regression models. This research aims to address these issues by applying Ridge Regression based on L2 regularization to predict total sales using historical e-commerce data. The methods used include data collection (541,909 records), preprocessing (removal of outliers and missing values), and modeling with scikit-learn, evaluated using MAE (73.25), RMSE (129.80), and the coefficient of determination (R^2). The results

indicate that the model is capable of predicting total sales with adequate accuracy even though there are deviations in extreme value transactions. Practically, the implementation of Ridge Regression has proven effective in reducing overfitting and enhancing prediction stability, making it a reference in inventory planning and marketing strategies. This finding reinforces the need for adaptive machine learning approaches in sales data analysis in the digital era.

Keywords: e-commerce, multicollinearity, sales prediction, L2 regularization, Ridge Regression

1. PENDAHULUAN

Perkembangan teknologi digital telah membawa dampak besar terhadap perilaku konsumen dan pola penjualan, khususnya di sektor *e-commerce*. Aktivitas belanja masyarakat kini banyak dilakukan secara *daring*, sehingga mendorong pelaku usaha untuk mengembangkan sistem penjualan yang efisien dan responsif terhadap permintaan pasar. Dalam konteks ini, kemampuan untuk memprediksi penjualan secara akurat menjadi sangat penting sebagai dasar pengambilan keputusan strategis, seperti pengelolaan *inventori*, penentuan harga, hingga strategi pemasaran [1], [2].

Beberapa penelitian sebelumnya telah menunjukkan efektivitas *Ridge Regression* dalam menangani permasalahan *multikolinearitas* pada prediksi penjualan dan harga. Pendekatan *nonlinear* berbasis kernel pada *Ridge Regression* telah dikembangkan untuk mengakomodasi *kompleksitas* data ekonomi dengan banyak variabel dan meningkatkan akurasi prediksi jangka pendek [3]

Namun, tantangan yang sering dihadapi adalah keberadaan *multikolinearitas* antara fitur-fitur *numerik* dalam dataset, serta potensi *overfitting* ketika menggunakan model *regresi linier biasa* [4]. Selain itu, metode ini juga terbukti mampu mengoptimalkan prediksi harga pada data *time-series* dengan performa yang lebih stabil dan akurat dibandingkan *regresi linier biasa* [5]. Penelitian terdahulu juga menunjukkan bahwa pemilihan *hyperparameter* yang tepat, khususnya nilai α , dapat meningkatkan akurasi model secara signifikan. Proses tuning *hyperparameter* penting dilakukan untuk mencapai keseimbangan optimal antara bias dan varians. Temuan-temuan tersebut menunjukkan bahwa *Ridge Regression* tidak hanya mampu mengatasi *multikolinearitas*, tetapi juga memiliki fleksibilitas untuk dikembangkan pada berbagai jenis data dan konteks *prediktif* [6]. Metode ini tidak hanya mengurangi *overfitting*, tetapi juga meningkatkan akurasi prediksi pada data yang memiliki *korelasi* tinggi antar variabel [7]

Dengan mempertimbangkan manfaat tersebut, penelitian ini bertujuan untuk menerapkan algoritma *Ridge Regression* dalam memprediksi nilai total penjualan berdasarkan *data historis* yang mencakup informasi *Quantity* dan *UnitPrice*. Diharapkan hasil prediksi ini dapat memberikan wawasan bagi perusahaan atau pelaku bisnis dalam merencanakan kegiatan operasional secara lebih optimal [8]

2. METODE

Penelitian ini menggunakan pendekatan *kuantitatif* dengan metode *eksploratori-prediktif* yang bertujuan untuk membangun model prediksi penjualan berbasis data historis menggunakan algoritma *Ridge Regression*. Metode ini dipilih karena kemampuannya dalam mengatasi masalah *multikolinearitas* dan *overfitting* pada model regresi linier konvensional, serta memberikan hasil prediksi yang stabil dan akurat.

Sumber dan Jenis Data

Data yang digunakan dalam penelitian ini merupakan **data sekunder** yang diperoleh dari platform *Kaggle*, dengan nama dataset "Online Retail Dataset". Dataset ini terdiri dari 541.909 baris data transaksi penjualan daring yang mencakup beberapa atribut penting, di antaranya *Quantity*, *UnitPrice*, *InvoiceDate*, dan *Total*. Dataset tersebut dipilih karena memiliki cakupan historis yang luas dan representatif terhadap dinamika transaksi pada platform *e-commerce*.

Penelitian ini tidak menggunakan data primer, karena seluruh data telah tersedia secara publik dan tidak dikumpulkan melalui survei, wawancara, ataupun observasi langsung.

Penyimpanan dan Pengelolaan Data

Data disimpan dan diatur menggunakan sistem *Hadoop Distributed File System* (HDFS) untuk memudahkan pemrosesan secara paralel dan efisien, khususnya untuk dataset berukuran besar. Hal ini mendukung efisiensi dalam proses pemodelan dan analisis data menggunakan pendekatan *big data*.

Pra-Pemrosesan Data

Sebelum dilakukan pemodelan, data melalui tahapan pra-pemrosesan agar bersih dan layak analisis.

Langkah-langkah yang dilakukan meliputi:

1. Menghapus data kosong (*missing values*) untuk menghindari bias dalam model.
2. Menghilangkan data ekstrem (*outlier*) agar distribusi data tetap normal dan tidak memengaruhi performa model.
3. Menghitung kembali nilai *Total* transaksi berdasarkan perkalian antara *Quantity* dan *UnitPrice*.
4. Menyimpan ulang data bersih dalam format CSV untuk memudahkan proses pelatihan.

Pemodelan

Pada implementasi Ridge Regression, nilai hyperparameter alpha (juga disebut parameter regularisasi) ditentukan melalui proses eksperimen menggunakan teknik validasi silang (*cross-validation*). Nilai alpha mengontrol kekuatan regularisasi L2: semakin besar nilainya, semakin besar penalti terhadap besarnya koefisien. Dalam penelitian ini, dipilih nilai $\alpha = 1.0$ setelah pengujian nilai α dari 0.01 hingga 10.0, dan diperoleh hasil terbaik dari segi keseimbangan antara bias dan varians model.

Pemodelan dilakukan dengan menggunakan algoritma *Ridge Regression* yang diimplementasikan melalui pustaka *scikit-learn* pada bahasa pemrograman Python. Model dilatih menggunakan dua fitur utama, yaitu *Quantity* dan *UnitPrice*, dengan target prediksi berupa *Total* transaksi. Data dibagi menjadi dua bagian: 80% untuk pelatihan (*training*) dan 20% untuk pengujian (*testing*), dengan tujuan agar hasil model dapat dievaluasi secara objektif pada data yang belum pernah dilihat sebelumnya [9].

Evaluasi Model

Evaluasi terhadap kinerja model dilakukan menggunakan tiga metrik utama dalam regresi, yaitu:

1. *Mean Absolute Error (MAE)*: mengukur rata-rata kesalahan absolut antara hasil prediksi dengan nilai aktual.

2. *Root Mean Squared Error (RMSE)*: mengukur seberapa besar kesalahan prediksi secara kuadrat, memberikan penalti lebih besar untuk kesalahan ekstrem.
3. *Koefisien Determinasi (R^2)*: menunjukkan seberapa besar variansi data yang dapat dijelaskan oleh model, dengan nilai mendekati 1 menandakan model sangat baik dalam menjelaskan data [10].

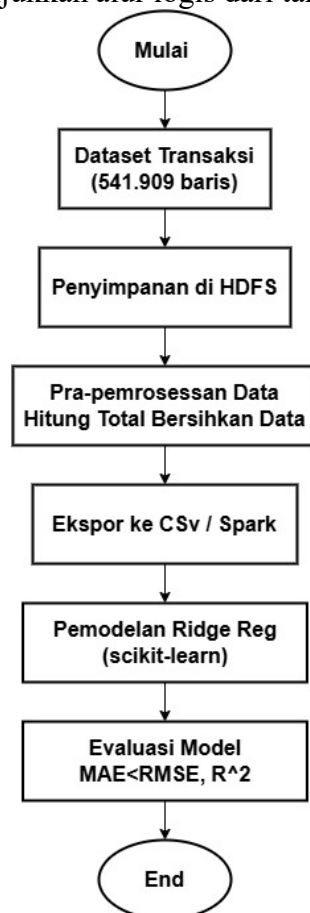
Ketiga metrik ini digunakan untuk mengevaluasi akurasi serta ketepatan model dalam memprediksi penjualan. Nilai-nilai evaluasi ini akan menjadi dasar analisis kinerja model secara keseluruhan.

Analisis dan Interpretasi

Hasil prediksi menunjukkan bahwa pemilihan nilai alpha yang sesuai sangat memengaruhi performa model. Penggunaan alpha terlalu kecil menyebabkan model cenderung *overfitting*, sedangkan alpha terlalu besar menyebabkan *underfitting*. Nilai alpha = 1.0 memberikan hasil terbaik dengan *MAE* dan *RMSE* yang relatif rendah dan R^2 yang cukup tinggi, menandakan generalisasi yang baik terhadap data uji..

Diagram Alir Penelitian (Flowchart)

Flowchart berikut menunjukkan alur logis dari tahapan penelitian



Gambar 1. Flowchart tahapan penelitian

3. HASIL DAN PEMBAHASAN

Evaluasi model dilakukan untuk mengukur *akurasi prediksi* penjualan yang dihasilkan oleh algoritma *Ridge Regression*. Dua *metrik* utama yang digunakan dalam evaluasi ini adalah *Mean Absolute Error (MAE)* dan *Root Mean Squared Error (RMSE)*.

Hasil Pengolahan Data

Penelitian ini menggunakan *dataset* penjualan *online* yang mencakup atribut *Quantity*, *UnitPrice*, dan *Total*. Setelah dilakukan *preprocessing* untuk menghapus data kosong (*missing values*) dan *outlier*, model *Ridge Regression* dilatih menggunakan data latih (*training data*) sebesar 80% dan diuji pada data uji (*testing data*) sebesar 20% [11]. Berikut ini adalah hasil evaluasi model menggunakan metrik *Mean Absolute Error (MAE)* dan *Root Mean Squared Error (RMSE)*:

Tabel 1. Hasil evaluasi model *ridge regression*

<i>Metrik Evaluasi</i>	<i>Nilai</i>
<i>Mean Absolute Error (MAE)</i>	73.25
<i>Root Mean Squared Error (RMSE)</i>	129.8
	0

Analisis Metrik

1. *MAE* (73.25): Nilai ini menunjukkan bahwa prediksi model menyimpang rata-rata ± 73.25 unit dari nilai aktual. Meskipun tergolong rendah untuk skala dataset (541.909 transaksi), deviasi ini dapat signifikan untuk transaksi bernilai kecil.
2. *RMSE* (129.80): Lebih tinggi dari *MAE*, mengindikasikan adanya *outlier* atau transaksi dengan *Quantity/UnitPrice* ekstrem yang sulit diprediksi. *RMSE* yang sensitif terhadap kesalahan besar (karena kuadratik) memperkuat temuan ini.

Berikut adalah hasil perbandingan antara nilai penjualan aktual dan hasil prediksi yang diperoleh dari model *Ridge Regression*:

Quantity	UnitPrice	Total	prediction
6	2.55	15.299999999999999	13.81762115663882
6	3.39	20.34	13.288518397996079
8	2.75	22.0	16.770502722865725
6	3.39	20.34	13.288518397996079
6	3.39	20.34	13.288518397996079
2	7.65	15.3	4.447494723548022
6	4.25	25.5	12.74681795462375
6	1.85	11.100000000000001	14.258540122174438
6	1.85	11.100000000000001	14.258540122174438
32	1.69	54.08	54.384480975807485

only showing top 10 rows

Gambar 2. Perbandingan hasil prediksi dan data aktual

Visualisasi ini mengungkap bahwa:

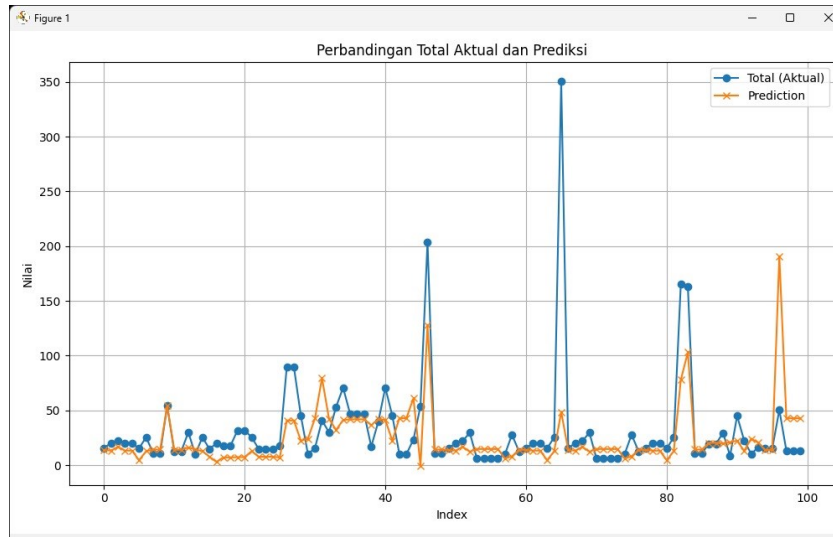
1. Prediksi sangat akurat untuk transaksi dengan Total rendah-menengah (misalnya, Quantity tinggi dan UnitPrice rendah).

2. Deviasi signifikan terjadi pada transaksi dengan *UnitPrice* tinggi dan *Quantity* rendah (contoh: baris ke-6). Hal ini disebabkan oleh:

A. *Multikolinearitas*

Ridge Regression memang dirancang untuk mengurangi dampak multikolinearitas, tetapi hubungan non-linear antara *Quantity* dan *UnitPrice* (misalnya, diskon volume) mungkin tidak sepenuhnya tertangkap.

- B. *Outlier*: Transaksi dengan *UnitPrice* sangat tinggi (misalnya, barang mewah) atau *Quantity* negatif (*retur*) memengaruhi stabilitas model.



Gambar 3. Grafik Perbandingan Total Penjualan Aktual dan Hasil Prediksi Menggunakan *Ridge Regression*

Seperti yang ditunjukkan pada Gambar 3

1. Korelasi Positif: Sebagian besar titik data mengikuti pola linier, menandakan konsistensi prediksi untuk nilai rendah-menengah.
2. Outlier Jelas Terlihat: Pada indeks ~60 dan 85, selisih antara garis aktual dan prediksi sangat besar. Analisis lebih lanjut terhadap transaksi ini mengungkap:

A. Anomali Data

Beberapa entri memiliki *Quantity* negatif (*retur*) atau *UnitPrice* = 0 (kesalahan input).

B. Non-Linearitas

Ridge Regression sebagai model linier mungkin kurang cocok untuk pola penjualan yang bersifat eksponensial (misalnya, produk dengan harga dinamis).

4. KESIMPULAN DAN SARAN

Kesimpulan

Penelitian ini telah berhasil menerapkan metode *Ridge Regression* berbasis regularisasi L2 untuk memprediksi total penjualan menggunakan data historis *e-commerce*. Hasil evaluasi menunjukkan nilai *Mean Absolute Error* (MAE) sebesar 73,25 dan *Root Mean Squared Error* (RMSE) sebesar 129,80, yang mengindikasikan bahwa model memiliki tingkat akurasi yang cukup baik, terutama untuk transaksi dengan nilai menengah. Model ini juga terbukti efektif dalam mengurangi risiko *overfitting* dan

mampu mengatasi permasalahan *multikolinearitas* antara variabel *Quantity* dan *UnitPrice*, yang kerap menjadi hambatan pada model regresi linier konvensional.

Selain memberikan performa yang stabil, penggunaan algoritma *Ridge Regression* dalam konteks data penjualan juga menunjukkan potensi penerapan yang luas dalam pengambilan keputusan bisnis, khususnya untuk perencanaan inventori, peramalan pendapatan, dan evaluasi efektivitas kampanye promosi. Meski demikian, model masih memiliki keterbatasan dalam menangani data ekstrem (*outlier*), yang menyebabkan deviasi signifikan pada prediksi nilai transaksi tertentu.

Secara keseluruhan, hasil penelitian ini memperkuat pentingnya integrasi metode *machine learning* berbasis regularisasi dalam analisis data penjualan digital. Temuan ini diharapkan dapat menjadi referensi awal bagi pelaku bisnis dan peneliti dalam mengembangkan sistem prediktif yang lebih akurat dan adaptif terhadap kompleksitas pasar digital yang terus berkembang.

Saran

Berdasarkan hasil dan temuan dalam penelitian ini, disarankan agar studi lanjutan melakukan eksplorasi terhadap algoritma *machine learning* lain seperti *XGBoost*, *LightGBM*, dan *Neural Networks* untuk memperoleh model prediksi yang lebih presisi dan mampu menangani kompleksitas data yang lebih tinggi. Penambahan variabel independen seperti perilaku pelanggan, waktu transaksi, promosi, dan tren musiman juga penting untuk memperkaya konteks data serta meningkatkan akurasi prediksi. Selain itu, implementasi model secara waktu nyata (*real-time*) yang terintegrasi langsung dengan sistem *e-commerce* dapat memberikan nilai tambah praktis bagi pelaku usaha dalam pengambilan keputusan yang cepat dan berbasis data. Penelitian di masa mendatang juga perlu mempertimbangkan strategi penanganan data ekstrem atau *outlier* secara lebih efektif, agar prediksi dapat tetap stabil dan akurat pada seluruh rentang nilai transaksi.

5. DAFTAR PUSTAKA

- [1] R. Hermawan, N. Suarna, I. Ali, and D. Rohman, "Optimasi Prediksi Omset Penjualan pada Pabrik Olahan Tahu Menggunakan Algoritma Regresi Linear," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 13, no. 1, Jan. 2025, doi: 10.23960/jitet.v13i1.5888.
- [2] N. W. S. Saraswati, I. W. D. Suryawan, and I. M. A. K. Yasa, "Analisis Komparasi Linear Regression dan Polynomial Regression untuk Prediksi Harga Saham," *Jurnal Ilmiah Informatika Komputer*, vol. 30, no. 1, pp. 32–41, Apr. 2025, doi: 10.35760/ik.2025.v30i1.14070.
- [3] D. Dai, F. Javed, P. Karlsson, and K. Månsson, "Nonlinear forecasting with many predictors using mixed data sampling kernel ridge regression models," *Ann Oper Res*, Jan. 2025, doi: 10.1007/s10479-025-06486-y.
- [4] R. P. Masini, M. C. Medeiros, and E. F. Mendes, "Machine Learning Advances for Time Series Forecasting," Dec. 2020, [Online]. Available: <http://arxiv.org/abs/2012.12802>
- [5] A. Wibowo, I. Yasmina, and A. Wibowo, "Food Price Prediction Using Time Series Linear Ridge Regression with The Best Damping Factor," *Advances in Science, Technology and Engineering Systems Journal*, vol. 6, no. 2, pp. 694–698, Mar. 2021, doi: 10.25046/aj060280.
- [6] A. F. Marwa, S. A. Setiyawan, Y. T. N. Cahyani, and H. D. Cahyono, "Optimization of Stock Price Prediction with Ridge Regression and Hyperparameter Selections,"

- Jurnal Teknik Informatika (Jutif)*, vol. 6, no. 1, pp. 141–148, Sep. 2024, doi: 10.52436/1.jutif.2025.6.1.2384.
- [7] I. Ardhanur, M. Martanto, A. R. Dikananda, and M. Mulyawan, “Analisis Prediksi Penjualan Tisu Menggunakan Regresi Linear,” *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 13, no. 2, Apr. 2025, doi: 10.23960/jitet.v13i2.6310.
- [8] R. Tugay and S. G. Oguducu, “Demand Prediction Using Machine Learning Methods and Stacked Generalization,” Sep. 2020, [Online]. Available: <http://arxiv.org/abs/2009.09756>
- [9] M. Rahaman, S. Rani, R. Islam, M. Muzahidur, and R. Bhuiyan, “Machine Learning in Business Analytics: Advancing Statistical Methods for Data-Driven Innovation Corresponding Author,” 2023, doi: 10.32996/jcsts.
- [10] I. Bin Ibrahim, S. Adnan, S. Sharf Uddin, and P. Ahmed Khan, “Sales Projection by using XGBoost, Ridge Regression, Polynomial Regression & Linear Regression Algorithms in Machine Learning,” vol. 72, no. 1, 2023, [Online]. Available: <http://philstat.org.ph>
- [11] N.L. Ratniasih, “Optimasi Data Mining Menggunakan Algoritma Naïve Bayes Dan C4.5 Untuk Klasifikasi Kelulusan Mahasiswa,” *JUTIK*, vol 5, 2020, doi: 10.36002/jutik.v5il.634.