Sistem Informasi Prediksi Penjualan E-Commerce Menggunakan Analisis Data Historis dan Algoritma MLR

I Gusti Putu Wisnu Wardhana[a1], I Wayan Santiyasa[a2]

[a]Program Studi Informatika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Udayana
Jalan Raya Kampus Udayana, Bukit Jimbaran, Kuta Selatan, Badung, Bali Indonesia
[1]igpwisnuw@gmail.com
[2]santiyasa@unud.ac.id

Abstract

Accurate sales to improve business planning and decision making. This study aims to design an information system that utilizes historical data and multiple linear regression algorithms to predict e-commerce sales. This study addresses the current challenges in forecasting uncertain sales by analyzing historical sales data and identifying relevant independent variables, such as marketing efforts, economic factors, and customer behavior. Through the implementation of a multiple linear regression algorithm, the system calculates the relationship between these variables and sales, enabling accurate predictions. The proposed information system provides valuable insights for businesses to optimize inventory management, marketing strategy and resource allocation. The experimental results show the effectiveness of the system in forecasting e-commerce sales, resulting in increased operational efficiency and revenue. This research contributes to the field of e-commerce analytics and assists businesses in making data-driven decisions for sustainable growth.

Keywords: *e-commerce, sales prediction, historical data, multiple linear regression, forecasting*

1. Pendahuluan

Perkembangan pesat industri e-commerce dalam beberapa tahun terakhir telah menciptakan tantangan baru dalam menganalisa dan memprediksi penjualan. Dalam rangka mengoptimalkan efisiensi operasional dan meningkatkan pengambilan keputusan bisnis, prediksi penjualan yang akurat menjadi krusial. Oleh karena itu, penelitian ini bertujuan untuk merancang sistem informasi yang memanfaatkan data historis dan algoritma regresi linier berganda untuk melakukan prediksi penjualan e-commerce dengan tingkat akurasi yang tinggi.

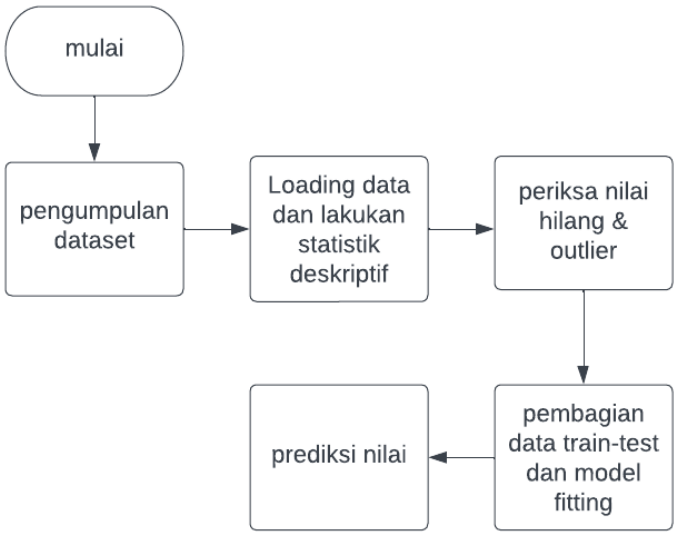
Untuk mencapai tujuan tersebut, penelitian ini menggunakan dataset "Linear Regression E-commerce" [1]. Dataset ini berisi data pelanggan yang membeli pakaian online di sebuah toko. Analisis data historis dari dataset ini bertujuan untuk mengidentifikasi pola dan tren penjualan masa lalu guna memprediksi penjualan masa depan. Dalam analisis ini, diterapkan algoritma regresi linier berganda yang mempertimbangkan faktor-faktor seperti karakteristik pelanggan, detail pembelian, dan informasi waktu. Dataset baru ini dapat memberikan wawasan relevan terkait faktor-faktor yang mempengaruhi penjualan

Beberapa penelitian sebelumnya memberikan kontribusi penting dalam pengembangan sistem informasi prediksi penjualan e-commerce. Tolstoy (2016) [2] menunjukkan bahwa algoritma regresi linier berganda dapat digunakan sebagai metode untuk memprediksi penjualan online. Penelitian lain yang sejenis dilakukan oleh Khan dan Uddin (2021) [3] yang menemukan bahwa regresi linier berganda dapat menjadi pilihan utama dalam prediksi penjualan di industri e-commerce. Sebagaimana dijelaskan oleh Basyir (2020) [4], dalam rancang bangun sistem informasi, penting untuk memperhatikan preferensi pengguna, sehingga sistem yang dibangun dapat menjadi user-friendly dan dapat diakses dengan mudah.

Ma et al. (2017) [5] mencatat bahwa data historis penjualan adalah input yang penting dalam memprediksi penjualan. Mereka menambahkan bahwa analisis pola dan tren penjualan masa lalu dapat memberikan wawasan berharga dalam memprediksi penjualan masa depan. Hasil dari analisis pola tersebut dapat mengarah pada pengembangan model prediksi yang tepat sehingga dapat membantu menjaga stok produk yang tepat dengan memaksimalkan penjualan dengan waktu yang tepat. Penelitian oleh Poon dan Zhong (2016) [6] mempelajari pentingnya mempertimbangkan interaksi antara faktor-faktor yang mempengaruhi penjualan e-commerce dalam model prediksi, karena hal ini dapat menghasilkan ramalan yang lebih akurat dan membantu perusahaan membuat keputusan yang lebih baik dalam pengelolaan persediaan dan optimasi penjualan.

Dalam proyek ini, akan dilatih model machine learning menggunakan regresi linier berganda dengan memanfaatkan dataset "Online Retail II". Selanjutnya, akan dikembangkannya sebuah antarmuka sederhana menggunakan framework Streamlit yang memungkinkan pengguna untuk memilih produk dari dataset, memilih jangka waktu tertentu, dan mendapatkan prediksi jumlah penjualan di masa mendatang. Dengan demikian, sistem informasi yang dihasilkan diharapkan dapat membantu perusahaan e-commerce dalam mengoptimalkan keputusan bisnis berdasarkan prediksi penjualan yang akurat.

2. Metode Penelitian



Gambar 1. Tahapan Penelitian

Penelitian ini dimulai dengan mengumpulkan dataset yang diperlukan untuk studi. Dataset dimuat menggunakan library pandas, dan statistik deskriptif dihitung untuk memperoleh pemahaman tentang data tersebut. Nilai yang hilang diperiksa dan ditangani dengan tepat. Outlier diidentifikasi menggunakan visualisasi dan dihapus dari dataset. Data kemudian dibagi menjadi set pelatihan dan pengujian menggunakan pembagian pelatihan-pengujian. Model regresi linear disesuaikan dengan data pelatihan untuk mempelajari pola dan hubungan yang ada. Model yang telah dilatih digunakan untuk membuat prediksi pada data pengujian. Akhirnya, nilai yang diprediksi dievaluasi dan dianalisis untuk menilai kinerja dan mengambil kesimpulan dari penelitian tersebut.

2.1 Pengumpulan Data

Data untuk penelitian ini akan dikumpulkan melalui pengunduhan dataset dari platform Kaggle. Dataset ini akan berisi informasi historis tentang penjualan e-commerce, termasuk variabel seperti jumlah transaksi, waktu transaksi, kategori produk, harga produk, dan informasi pelanggan. Dataset ini akan memberikan kerangka kerja yang diperlukan untuk melakukan analisis dan prediksi penjualan e-commerce. Penelitian Sari dan Ibrani (2017) [7] yang membahas prediksi menggunakan regresi linier memvariasikan distribusi rasio data training:data testing menjadi 70:30. 80:20. dan 90:10 untuk menentukan ukuran training data yang optimal. Maka dari itu, dilakukan perbandingan tiga variasi tersebut

2.2 Pemodelan Regresi Linier Berganda

Model regresi linear berganda berguna ketika mempelajari hubungan antara beberapa variabel independen dan satu variabel dependen. Untuk membangun model-model tersebut, para peneliti pertama-tama mengidentifikasi variabel-variabel yang kemungkinan besar mempengaruhi variabel dependen. Hal ini dapat dicapai dengan menguji variabel-variabel tersebut untuk korelasi signifikan dengan variabel dependen menggunakan tingkat signifikansi, yang sering kali ditetapkan pada $p < 0{,}20$ [8]. Variabel-variabel yang memenuhi kriteria ini kemudian dimasukkan ke dalam model, yang dibangun menggunakan algoritma seperti metode bertahap [8], kuadrat terkecil biasa (OLS), atau penurunan gradien [9].

2.3 Evaluasi dan Validasi

Setelah model regresi linier berganda dibangun, akan dilakukan evaluasi dan validasi model. Langkah-langkah yang dilakukan meliputi:

a. MSE (Mean Squared Root)

MSE merupakan rata rata kuadrat kesalahan yang dihitung dengan menjumlahkan semua kesalahan atau eror prediksi yang dihasilkan oleh suatu model kemudian dikuadratkan dan membaginya dengan jumlah periode prediksi [12]. Berikut merupakan persamaan matematis dari MSE:

$$\mathrm{MSE} = \frac{1}{\mathrm{n}} \sum_{\mathrm{i}}^{\mathrm{n}} (\mathrm{Xi} - \mathrm{Fi})^2 \quad (1)$$

Keterangan:

Xi = Data aktual pada periode ke-i
Fi = Nilai hasil prediksi atau prediksi pada period ke-i
n = Banyaknya sampel

b. MAE (Mean Absolute Error)

MAE metrik yang digunakan untuk mengukur selisih rata-rata absolut antara nilai aktual dan nilai yang diprediksi dalam suatu model regresi. MAE dihitung dengan mengambil rata-rata dari selisih absolut antara nilai yang diprediksi dan nilai sebenarnya [13]. Berikut merupakan persamaan matematis dari MAE:

$$\mathrm{MAE} = \frac{1}{\mathrm{n}} \sum_{\mathrm{i}}^{\mathrm{n}} |\mathrm{yi} - \hat{\mathrm{y}}| \quad (2)$$

Keterangan:

n = Banyaknya sampel

yi = nilai aktual pada sampel ke i
yi = nilai prediksi pada sampel ke i

c. R2 Score

R2 Score digunakan untuk mengukur sejauh mana variabilitas suatu model statistic atau machine learning dapat menjelaskan variasi data yang diamati. Range nilai dari R2 Score adalah antara 0 hingga 1. Berikut merupakan persamaan matematis dari R2 Score:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(yi - \hat{y}i)^2}{\sum_{i=1}^{n}\left(yi - \underline{yi}\right)^2} \quad (3)$$

Keterangan:

yi = Nilai aktual dari variabel dependen
ŷi = Nilai aktual dari variabel dependen

d. Integrasi Model dengan Sistem

Langkah terakhir dalam metode penelitian ini adalah integrasi model prediksi dengan sistem yang relevan, seperti sistem manajemen penjualan e-commerce. Model yang telah dibangun akan digunakan untuk memberikan prediksi penjualan e-commerce berdasarkan input yang diberikan oleh sistem. Dalam hal ini, output dari model dapat digunakan sebagai panduan dalam pengambilan keputusan bisnis dan perencanaan strategi penjualan.

3. Hasil dan Diskusi

3.1. Perancangan Model

Data yang digunakan pada penelitian ini adalah dataset customer ecommerce yang terdiri dari email, alamat, avatar, waktu sesi, waktu kunjungan, jangka langganan, serta jumlah yang dibayarkan per tahun. Adapun dataset yang digunakan dapat diakses pada link berikut: https://www.kaggle.com/datasets/iyadavvaibhav/ecommerce-customer-device-usage.

Metode yang digunakan adalah regresi linear berganda.

	Avg. Session Length	Time on App	Time on Website	Length of Membership
0	34.497268	12.655651	39.577668	4.082621
1	31.926272	11.109461	37.268959	2.664034
2	33.000915	11.330278	37.110597	4.104543
3	34.305557	13.717514	36.721283	3.120179
4	33.330673	12.795189	37.536653	4.446308
...	...	...	...	...
495	33.237660	13.566160	36.417985	3.746573
496	34.702529	11.695736	37.190268	3.576526
497	32.646777	11.499409	38.332576	4.958264
498	33.322501	12.391423	36.840086	2.336485
499	33.715981	12.418808	35.771016	2.735160

Gambar 2. Dataset yang digunakan dalam perancangan model

Data yang didapatkan dari dataset kemudian akan dibagi menjadi data latih dan data uji. Pada penelitian ini, rasio data latih:data uji didistribusikan menjadi 70:30, 80:20. dan 90:10. Pada gambar 2, rasio ini direpresentasikan dalam bentuk desimal yang menunjukkan jumlah data latih.

	Test Ratio	MSE	MAE	R2 Score
0	0.3	93.434699	7.583105	0.981536
1	0.2	89.729365	7.358401	0.979355
2	0.1	71.811240	6.612771	0.982737

Gambar 3. Performa Algoritma Regresi Linier Berganda

Berdasarkan hasil yang diperoleh dari model regresi, dapat kita lihat bahwa rasio pelatihan dan pengujian yang berbeda digunakan untuk mengevaluasi kinerja. Model ini dilatih dengan menggunakan 70% dari data pada skenario pertama, 80% pada skenario kedua, dan 90% pada skenario ketiga, sedangkan data yang tersisa digunakan untuk pengujian.

Metrik evaluasi yang digunakan untuk mengukur kinerja model adalah Mean Squared Error (MSE), Mean Absolute Error (MAE), dan R2 Score. Metrik-metrik ini memberikan wawasan tentang sejauh mana model cocok dengan data dan dapat memprediksi penjualan produk e-commerce.

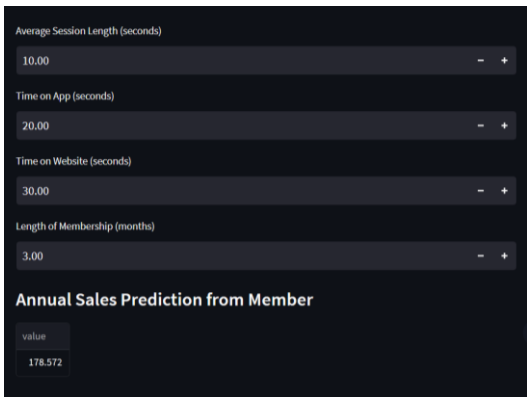
Dari hasil yang diperoleh, dapat diamati bahwa seiring dengan peningkatan ukuran set pelatihan, kinerja model juga meningkat. Hal ini terlihat dari nilai yang semakin mengecil pada MSE dan MAE, yang menunjukkan bahwa model cocok dengan data dengan baik. Dengan rasio tes 0.3, 0.2, dan 0.1, dapat dilihat bahwa Nilai MSE mengecil dari 93.4, 89.7, lalu 71.8. Untuk rasio tes yang sama, nilai MAE juga semakin berkurang, 7.59 untuk rasio 0.3, 7.36 untuk rasio 0.2, lalu 6.61 untuk rasio 0.1. Nilai MAE dan MSE yang lebih rendah menunjukkan tingkat akurasi model regresi lebih tinggi.

Selain itu, R2 Score, yang menggambarkan proporsi variansi dalam variabel target yang dijelaskan oleh model, juga meningkat seiring dengan peningkatan ukuran set pelatihan. Dengan nilai R2 sekitar 0,98, dapat disimpulkan bahwa model ini mampu menjelaskan sekitar 98% variasi dalam data penjualan. Hal ini menunjukkan bahwa model regresi linier berganda dengan menggunakan data historis dan fitur-fitur yang dipilih sangat efektif dalam memprediksi penjualan e-commerce.

Nilai MSE dan MAE rendah menunjukkan bahwa model memiliki tingkat kesalahan prediksi yang kecil. Selain itu, jika nilai R2 mendekati 1, maka model mampu menjelaskan sebagian besar variasi yang ada dalam data penjualan E-commerce. Dengan mempertimbangkan ketiga metrik ini, dapat disimpulkan bahwa model regresi linier berganda dalam proyek ini memiliki performa yang baik dalam memprediksi penjualan E-commerce berdasarkan analisis data historis.

Secara keseluruhan, hasil ini menunjukkan potensi penggunaan data historis dan algoritma regresi linier berganda dalam membangun sistem prediksi penjualan e-commerce. Analisis dan optimasi lebih lanjut dapat dilakukan untuk meningkatkan kinerja model dan mengeksplorasi fitur-fitur tambahan yang dapat berkontribusi pada prediksi penjualan yang akurat.

3.2. Integrasi Backend dengan Frontend



Gambar 4. Antarmuka aplikasi berbasis streamlit

Digunakan library streamlit untuk mengimplementasi antarmuka interaktif untuk mempermudah melakukan prediksi pendapatan e-commerce masing-masing member berdasarkan data yang dimiliki terkait member tersebut. Library streamlit dipilih karena penggunaannya yang sederhana sehingga mempercepat proses perancangan sistem.

4. Kesimpulan

Berdasarkan hasil penelitian dan analisis yang dilakukan, dapat diambil kesimpulan sebagai berikut:

a. Model regresi linier berganda yang dikembangkan dalam penelitian ini mampu memberikan prediksi penjualan E-commerce yang akurat dan mendekati nilai sebenarnya.
b. Evaluasi model menggunakan metrik MSE (Mean Squared Error), MAE (Mean Absolute Error), dan R2 (R-squared) menunjukkan hasil yang baik, dengan nilai-nilai sebagai berikut:

    1. Pada uji ratio 0.3: MSE = 93.434699, MAE = 7.583105, R2 Score = 0.981536.
    2. Pada uji ratio 0.2: MSE = 89.729365, MAE = 7.358401, R2 Score = 0.979355.
    3. Pada uji ratio 0.1: MSE = 71.811240, MAE = 6.612771, R2 Score = 0.982737.

c. Hasil pengujian dengan menggunakan tiga rasio uji yang berbeda menunjukkan konsistensi kinerja model dalam menghasilkan prediksi yang akurat.
d. MSE yang rendah pada setiap rasio uji menunjukkan tingkat kesalahan prediksi yang minim.
e. MAE yang rendah mengindikasikan bahwa model memiliki tingkat kesalahan prediksi yang kecil secara keseluruhan.
f. R2 Score yang mendekati 1 menunjukkan bahwa model dapat menjelaskan sebagian besar variasi dalam data penjualan E-commerce.
g. Hasil ini menunjukkan bahwa model regresi linier berganda yang dikembangkan memiliki kemampuan yang baik dalam memprediksi dan menjelaskan penjualan E-commerce.
h. Dengan demikian, sistem informasi prediksi penjualan yang dikembangkan dalam penelitian ini dapat digunakan sebagai alat yang efektif dalam membantu perusahaan E-commerce dalam mengoptimalkan penjualan.

Daftar Pustaka

[1] [1] S. Kolawale, "Linear Regression E-Commerce Dataset," [Online]. Available: https://www.kaggle.com/datasets/kolawale/focusing-on-mobile-app-or-website.
[2] D. Tolstoy, A. Jonsson, and D. D. Sharma, "The influence of a retail firm's geographic scope of operations on its international online sales," International Journal of Electronic

Commerce, vol. 20, no. 3, pp. 293–318, 2016. doi:10.1080/10864415.2016.1121760
[3] A. H. Khan and G. S. Uddin, "Predicting e-commerce sales using machine learning techniques," Journal of Retailing and Consumer Services, vol. 59, 2021. doi: 10.1016/j.jretconser.2020.102332.
[4] A. Basyir, M. Maskur, and I. Nuryasin, "Rancang Bangun Sistem Informasi Pertandingan Pencak Silat Berbasis Website Menggunakan Metode User Centered Design (Ucd)," JR, vol. 12, no. 2, pp. 1663–1670, 2020. doi:10.22219/repositor. v2i12.571
[5] J. Ma, Y. Liang, G. Liu, and P. Zhou, "Forecasting Fashion Retail Sales with a Hybrid Method Combining Singular Spectrum Analysis and Back-Propagation Neural Network," International Journal of Pattern Recognition and Artificial Intelligence, vol. 31, no. 8, 2017. doi:10.1142/S0218001417560101.
[6] Poon and Zhong, "A Forecasting Model of Online Retail Sales incorporating Traffic and Advertising Efforts," Journal of Business Research, vol. 69, no. 11, pp. 4958–4964, 2016. doi: 10.1016/j.jbusres.2016.05.032.
[7] Sari and Ibrani, "Prediksi Harga Saham Menggunakan Regresi Linear Berganda Berbasis Machine Learning," Jurnal Rekayasa Sistem dan Teknologi Informasi (RESTI), vol. 5, no. 2, pp. 168–174, 2021. doi:10.29207/resti. v5i2.2067
[8] D. Castilho, A. Silva, F. Gimenes, R. Nunes, A. Pires, and C. Bernardes, "Factors Related To the Patient Safety Climate In An Emergency Hospital," Rev. Latino-Am. Enfermagem, vol. 28, 2020. doi:10.1590/1518-8345.3353.3273
[9] N. Hirnschall, S. Norrby, M. Weber, S. Maedel, S. Amir-Asgari, and O. Findl, "Using Continuous Intraoperative Optical Coherence Tomography Measurements Of the Aphakic Eye For Intraocular Lens Power Calculation," Br J Ophthalmol, vol. 99, no. 1, pp. 7–10, 2014. doi:10.1136/bjophthalmol-2013-304731
[10] F. Gouzi et al., "Blunted Muscle Angiogenic Training-response In Copd Patientsversussedentary Controls," Eur Respir J, vol. 4, no. 41, pp. 806–814, 2012. doi:10.1183/09031936.00053512
[11] G. Khalaf and M. Iguernane, "Multicollinearity and A Ridge Parameter Estimation Approach," J. Mod. App. Stat. Meth., vol. 2, no. 15, pp. 400–410, 2016. doi:10.22237/jmasm/1478002980
[12] H. Setyawan, S. H. Fitriasih, and R. T. Vulandari, "Prediksi Tingkat Produksi Buah Kelapa Sawit dengan Metode Single Moving Average," J.TIKomSiN, vol. 9, no. 2, pp. 1–10, 2021. ISSN: 2338-4018 https://doi.org/10.30646/tikomsin.v9i2.53
[13] C. Willmott and K. Matsuura, "Advantages of the Mean Absolute Error (Mae) Over The Root Mean Square Error (Rmse) In Assessing Average Model Performance," Clim. Res., no. 30, pp. 79–82, 2005. doi:10.3354/cr030079

Halaman ini sengaja dibiarkan kosong