Infoman's : Jurnal Ilmu-ilmu Informatika dan Manajemen Vol. 19 No.2 November (2025)
DOI: 10.33481/infomans.vxxixx

Analisis Komparatif Model Regresi Linier dan Polinomial pada *Small Dataset* untuk Prediksi Timbulan Sampah

[1]Listia Silviani, [2]Esa Firmansyah, [3]Beben Sutara
[1]Universitas Sebelas April
[2]Universitas Sebelas April
[3]Universitas Sebelas April
[1,2,3]Jl. Angkrek Situ No.19, Sumedang, Jawa Barat 45323
email : [1]220660121108@student.unsap.ac.id, [2]esa@unsap.ac.id, [3]beben@unsap.ac.id

ABSTRACT
Bandung Regency faces severe waste management challenges, generating an average daily volume of 1,300 tons without possessing an independent landfill facility. Consequently, accurate data-driven prediction is crucial for strategic infrastructure planning. However, the application of machine learning algorithms at the regional level is often significantly constrained by the scarcity of historical data (small datasets), which introduces high risks of overfitting and statistical bias. This study aims to evaluate prediction modeling strategies using a limited annual dataset (n=4) spanning from 2021 to 2024, sourced from Open Data Jabar. The methodology compares Linear Regression and Second-Degree Polynomial Regression algorithms by employing a full training approach combined with descriptive validation based on the Rate of Change (RoC) analysis. The results indicate that while Polynomial Regression achieves superior statistical performance with an R-squared of 0.9958 and an RMSE of 3,144.59 tons, trend analysis reveals clear signs of overfitting, as the model predicts an implausible deceleration in future waste growth that contradicts demographic realities. Conversely, Linear Regression (R^2 = 0.9756) provides more stable and consistent trend estimates. Therefore, this study recommends Linear Regression as a more robust model for policy planning under limited data conditions, projecting waste generation to reach 467,964.01 tons in 2025, serving as an early warning for urgent capacity expansion.

Keywords **-** Small Dataset, Waste Prediction, Linear Regression, Polynomial Regression

1. Introduction

Permasalahan pengelolaan sampah di Indonesia merupakan isu lingkungan multidimensi yang terus berkembang seiring laju urbanisasi. Berdasarkan data Sistem Informasi Pengelolaan Sampah Nasional (SIPSN) tahun 2023, total timbulan sampah nasional mencapai 56,63 juta ton/tahun, dengan tingkat pengelolaan yang belum merata di berbagai daerah [1]. Di tingkat regional, Jawa Barat menyumbang volume sampah harian yang tinggi akibat aktivitas padat di kawasan metropolitan Bandung Raya. Secara spesifik, Kabupaten Bandung menghadapi tekanan berat karena menghasilkan sekitar 1.301,5 ton sampah per hari dengan populasi lebih dari 3,7 juta jiwa [2]. Masalah menjadi semakin kritis karena wilayah ini tidak memiliki Tempat Pembuangan Akhir (TPA) mandiri dan bergantung sepenuhnya pada TPA Regional Sarimukti yang telah mengalami kelebihan kapasitas (*overload*). Ketidaktepatan dalam memprediksi volume sampah di masa depan dapat berujung pada krisis operasional dan inefisiensi anggaran daerah [2].

Penerapan *data mining* dan *machine learning* telah menjadi solusi andalan dalam manajemen lingkungan cerdas (*smart environment*). Namun, tantangan utama di pemerintahan daerah adalah ketersediaan data historis yang seringkali terbatas pada format agregat tahunan (*Small Dataset Time-Series*). Studi terbaru dari Alhathlaul et al. (2025) menyoroti bahwa keterbatasan data menghambat

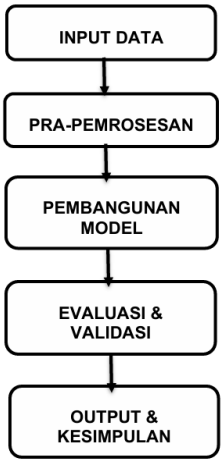
kemampuan model kompleks untuk menangkap pola tren jangka panjang secara akurat [3]. Pada kondisi sampel kecil (n < 10), algoritma kompleks cenderung mengalami *overfitting* kondisi di mana model "menghafal" data latih namun gagal memprediksi masa depan dengan tepat. Choi dan Lee (2023) juga membuktikan bahwa performa model prediksi sangat bergantung pada ukuran dataset pelatihan, di mana dataset kecil memerlukan strategi penyederhanaan model untuk menjaga stabilitas varians [4].

Penelitian terdahulu menunjukkan hasil yang beragam mengenai efektivitas model regresi. Satriyo et al. (2023) menemukan Regresi Polinomial unggul menangkap pola non-linier pada data fluktuatif seperti harga saham [5]. Namun, dalam konteks data lingkungan, Kang et al. (2025) menekankan perlunya validasi ketat saat menggunakan model polinomial untuk menghindari bias prediksi [6]. Di sisi lain, Farhan & Setiaji (2023) mencatat bahwa pada dataset kecil, Regresi Linier justru memberikan hasil yang lebih stabil dibandingkan metode kompleks seperti *Neural Network* [7].

Kesenjangan penelitian (*research gap*) terletak pada belum adanya strategi validasi yang spesifik untuk menangani prediksi sampah pada *small dataset* di level kabupaten. Penelitian ini bertujuan mengisi celah tersebut dengan mengkomparasikan performa dan stabilitas antara Regresi Linier dan Polinomial. Fokus utamanya adalah menentukan model mana yang paling *robust* (tahan uji) untuk perencanaan kebijakan, dengan mempertimbangkan validasi tren terhadap data pendukung seperti pertumbuhan penduduk dan kapasitas armada.

2. Research Method

Penelitian ini menggunakan metode kuantitatif dengan pendekatan studi dokumentasi pada data sekunder. Pendekatan ini dipilih karena efektivitasnya dalam memanfaatkan data publik yang telah terverifikasi untuk analisis prediktif [8]. Prosedur penelitian dirancang secara sistematis dalam lima fase utama yang mencakup siklus hidup data (*data lifecycle*) dari akuisisi hingga penarikan kesimpulan, sebagaimana diilustrasikan pada Gambar 1.



Gambar 1. Alur Metodologi Penelitian

2.1. Fase 1: Input Data (*Data Acquisition*)

Tahap ini merupakan proses inisialisasi data. Data utama diperoleh dari portal resmi *Open Data Jabar*, *BPS, & DLH Kabupaten Bandung*. Mengingat penelitian ini berfokus pada tantangan *small dataset*, data yang digunakan terdiri dari runtun waktu pendek periode 2021–2024. Data timbulan sampah disandingkan dengan indikator demografi dan infrastruktur untuk keperluan analisis kontekstual, sebagaimana disajikan pada Tabel 1.

Tabel 1. Dataset Timbulan Sampah dan Indikator Pendukung

DOI: 10.33481/infomans.vxxixx

Tahun	Volume Sampah (Ton)	Jumlah Penduduk (Jiwa)	Armada Angkut (Unit)
2021	288.885,04	3642196	109
2022	350.948,94	3718660	109
2023	385.212,08	3721110	119
2024	419.800,49	3839700	119

Variabel independen (X) yang digunakan dalam model adalah tahun, sedangkan variabel dependen (Y) adalah volume timbulan sampah. Data penduduk dan armada digunakan sebagai variabel kontekstual untuk validasi tren pada tahap pembahasan [8].

2.2. Fase 2: Pra-pemrosesan Data

Data mentah yang diperoleh dari sumber terbuka seringkali memiliki format angka non-standar (menggunakan pemisah ribuan). Pada fase ini dilakukan pembersihan karakter non-numerik dan konversi tipe data menjadi format *float* agar dapat diproses secara matematis. Integritas data divalidasi untuk memastikan tidak ada nilai yang hilang (*missing values*) yang dapat mendistorsi hasil prediksi [9]. Tahapan ini krusial untuk meminimalisir *noise* sebelum data masuk ke tahap pelatihan model.

2.3. Fase 3: Pembangunan Model

Mengingat jumlah observasi yang sangat terbatas (n=4), penelitian ini menerapkan strategi *Full Training* (melatih model dengan 100% data historis) guna memaksimalkan pola yang dapat dipelajari oleh algoritma. Dua model regresi dibangun menggunakan pustaka *Scikit-Learn*:

a. Regresi Linier Sederhana
   Model ini memodelkan hubungan garis lurus antara variabel tahun dan volume sampah untuk menguji asumsi tren pertumbuhan yang konstan [10]. Persamaan matematisnya adalah:

$$Y = a + bX \qquad (1)$$

Dimana:

1. Y = Variabel dependen (Volume Timbulan Sampah)
2. X = Variabel independen (Tahun)
3. a = Konstanta (*Intercept*)
4. b = Koefisien regresi (*Slope* / kemiringan garis)

b. Regresi Polinomial (Derajat 2)
   Model ini melakukan transformasi fitur (X^2) untuk memodelkan hubungan non-linier atau melengkung, guna menangkap potensi akselerasi atau perlambatan pertumbuhan sampah [6]. Persamaannya adalah:

$$Y = a + bX + cX^2 \qquad (2)$$

Dimana:

1. Y = Variabel dependen (Volume Timbulan Sampah)
2. X = Variabel independen (Tahun)
3. b_1, b_2 = Koefisien regresi untuk variabel linier dan kuadratik
4. X^2 = Fitur polinomial derajat dua

2.4. Fase 4: Evaluasi dan Validasi

Keandalan model diuji melalui tiga pendekatan validasi berlapis untuk memastikan hasil prediksi yang *robust*. Pertama, evaluasi statistik dilakukan dengan menilai akurasi model terhadap data latih menggunakan metrik *Coefficient of Determination* (R^2), *Root Mean Square Error* (RMSE), dan *Mean Absolute Error* (MAE) [6]. Kedua, dilakukan validasi manual melalui verifikasi ulang koefisien yang dihasilkan algoritma Python menggunakan perhitungan matematis manual berbasis metode matriks, sehingga validitas internal model dapat dipastikan. Ketiga, validasi tren (rate of change) diterapkan dengan mempertimbangkan prinsip peramalan pada data berskala pendek. Karena membagi dataset yang sangat kecil ($n = 4$) untuk *split test* berpotensi menghilangkan informasi penting, maka validasi digantikan dengan analisis konsistensi persentase laju pertumbuhan tahunan[11]. Pendekatan ini digunakan untuk mendeteksi potensi anomali tren maupun indikasi *overfitting* pada model[12].

2.5. Fase 5: Output dan Kesimpulan

Tahap akhir adalah sintesis hasil komparasi antara akurasi statistik dan stabilitas tren. Berdasarkan analisis tersebut, ditentukan model terbaik yang akan digunakan untuk memproyeksikan volume sampah tahun 2025–2026 sebagai dasar rekomendasi kebijakan berbasis data (*data-driven policy*) [13,14].

3. Result and Analysis

Bagian ini menyajikan temuan empiris dari penerapan model prediksi pada dataset terbatas, serta analisis kritis terhadap validitas hasil prediksi berdasarkan konteks demografi dan infrastruktur.

3.1. Hasil Fase 1 & 2: Akuisisi dan Pra-pemrosesan Data

Pada tahap akuisisi, data mentah periode 2021–2024 berhasil dikumpulkan[15,16]. Proses pra-pemrosesan data kemudian dilakukan untuk menangani format non-standar. Data mentah awalnya memiliki format teks dengan pemisah ribuan berupa titik, seperti pada nilai "288.885,04", sehingga tidak dapat langsung digunakan dalam proses analisis. Setelah dilakukan konversi, seluruh data berhasil diubah menjadi format numerik (float) yang sesuai untuk pemodelan. Proses validasi integritas data juga menunjukkan bahwa tidak terdapat missing values, sehingga dataset dinyatakan bersih dan siap digunakan pada tahap pemodelan berikutnya.

3.2. Hasil Fase 3: Pembangunan Model Regresi

Dua model dibangun menggunakan strategi *full training*, yaitu melatih algoritma dengan seluruh data historis yang tersedia. Berdasarkan hasil pelatihan, diperoleh dua parameter model utama. Pada Regresi Linier, model menghasilkan persamaan: $Y = 43.438{,}18(X) - 87.491.229{,}90$, dengan koefisien slope positif yang menunjukkan adanya tren kenaikan volume sampah secara konstan setiap tahun. Sementara itu, Regresi Polinomial menghasilkan persamaan kuadratik dengan koefisien X^2 bernilai negatif ($-3.512{,}54$). Koefisien negatif ini membentuk kurva cembung, sehingga secara matematis mengindikasikan adanya perlambatan laju pertumbuhan timbulan sampah di masa mendatang.

3.3. Hasil Fase 4: Evaluasi dan Validasi Menyeluruh

Pertama, pada evaluasi statistik dan validasi manual, perhitungan ulang secara manual menunjukkan bahwa hasil algoritma Python memiliki tingkat presisi yang konsisten. Evaluasi metrik pada data latih disajikan dalam Tabel 2.

Tabel 2. Perbandingan Metrik Evaluasi Model

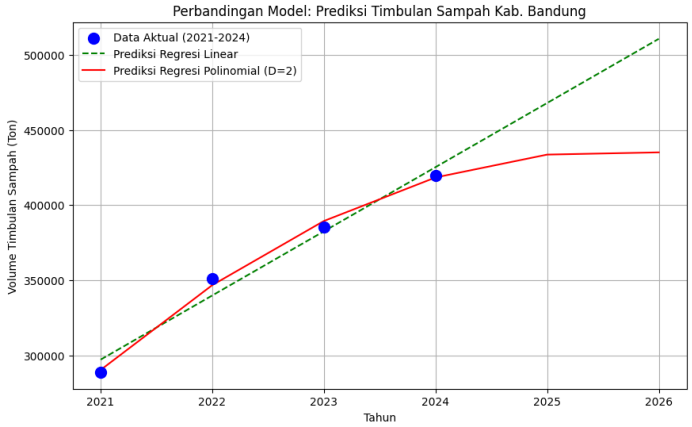
DOI: 10.33481/infomans.vxxixx

Model	R2 Score	RMSE	MAE
Regresi Linier	0.9756	7,554.4576	6,868.8725
Regresi Polinomial (D=2)	0.9958	3,144.5858	2,812.6030

Tabel 2. menyajikan perbandingan komprehensif antara kinerja model Regresi Linier dan Regresi Polinomial berdasarkan tiga metrik utama: *Coefficient of Determination* (R2), *Root Mean Square Error* (RMSE), dan *Mean Absolute Error* (MAE). Secara statistik, model Regresi Polinomial (Derajat 2) menunjukkan superioritas yang signifikan dalam kemampuan *fitting* data. Model ini menghasilkan nilai R2 yang nyaris sempurna sebesar 0,9958, yang mengindikasikan bahwa fungsi polinomial mampu menjelaskan 99,58% variasi pola data timbulan sampah historis. Tingkat kesalahan prediksi model ini juga terbukti paling rendah, dengan nilai RMSE sebesar 3.144,59 ton dan MAE sebesar 2.812,60 ton. Rendahnya nilai *error* ini menunjukkan bahwa kurva kuadratik memiliki fleksibilitas tinggi untuk meminimalisir jarak residu terhadap titik-titik data aktual.

Sebaliknya, model Regresi Linier menunjukkan performa statistik yang sedikit di bawah model polinomial, dengan nilai R2 sebesar 0,9756. Meskipun masih tergolong sangat kuat (di atas 0,90), model linier memiliki tingkat kesalahan yang lebih besar, ditandai dengan nilai RMSE mencapai 7.554,46 ton dan MAE sebesar 6.868,87 ton. Selisih *error* yang cukup lebar ini disebabkan oleh karakteristik fungsi linier yang kaku (*rigid*), sehingga cenderung mengambil garis tren rata-rata (*general trend*) dan tidak dapat mengikuti fluktuasi minor pada setiap titik data tahunan sepresisi model polinomial. Namun, perlu ditekankan bahwa dalam konteks *small dataset* (n=4), nilai akurasi statistik yang sangat tinggi pada model polinomial perlu disikapi dengan kehati-hatian karena berpotensi mengindikasikan gejala *overfitting*, di mana model terlalu sensitif terhadap data latih yang terbatas.

Kedua, pada validasi tren (*rate of change*), analisis kestabilan pola pertumbuhan memperlihatkan adanya perbedaan proyeksi kedua model. Rata-rata kenaikan aktual berada di angka 13,41%. Model Linier menghasilkan proyeksi kenaikan yang stabil sebesar 12,70%, sementara Regresi Polinomial menunjukkan kecenderungan perlambatan yang cukup signifikan pada tahun 2026. Hal ini divisualisasikan pada Gambar 2.

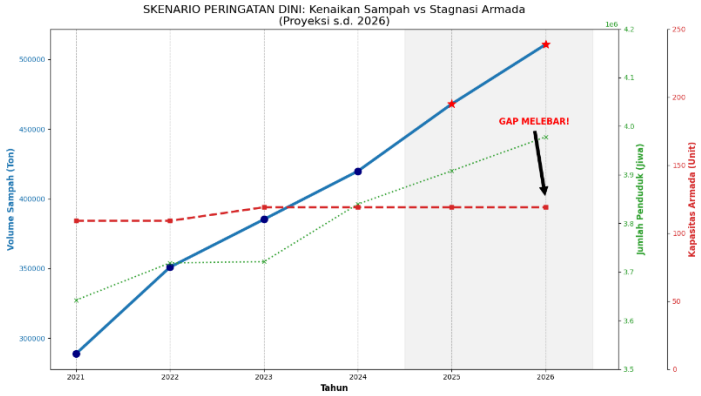


Gambar 2. Perbandingan Model Prediksi

Visualisasi pada Gambar 2. memperlihatkan divergensi tren yang krusial. Kurva Regresi Polinomial (garis merah) tampak membengkok untuk menyesuaikan diri secara agresif terhadap data historis, namun menghasilkan proyeksi perlambatan (*plateau*) yang tidak realistis di masa depan. Fenomena ini mengonfirmasi indikasi *overfitting* akibat keterbatasan data latih. Sebaliknya, Regresi Linier (garis

hijau) mempertahankan lintasan kenaikan yang positif dan stabil, yang dinilai lebih merepresentasikan kondisi nyata pertumbuhan beban sampah jangka panjang.

Ketiga, validasi kontekstual dilakukan dengan membandingkan hasil prediksi dengan data pendukung seperti dinamika populasi dan kapasitas armada pengangkut sampah. Validasi ini bertujuan memastikan bahwa tren yang dihasilkan model tetap logis dan sesuai kondisi riil di lapangan, sebagaimana terlihat pada Gambar 3**.**



Gambar 3. Disparitas Tren Timbulan Sampah, Pertumbuhan Penduduk, dan Kapasitas Armada

Visualisasi pada Gambar 3. mengonfirmasi validitas logika model: tren kenaikan timbulan sampah (batang biru) berkorelasi positif dengan laju pertumbuhan penduduk (garis hijau), yang mematahkan asumsi perlambatan pada model Polinomial. Temuan yang lebih krusial adalah teridentifikasinya kesenjangan infrastruktur yang melebar (*widening gap*), di mana lonjakan volume sampah tidak diimbangi dengan penambahan kapasitas armada (garis merah putus-putus) yang stagnan. Kondisi ini menjadi sinyal peringatan dini (*early warning*) bagi pemerintah daerah mengenai risiko krisis akumulasi sampah pada tahun 2025.

3.4. Hasil Fase 5: Sintesis Output

Berdasarkan hasil komparasi pada Fase 4, penelitian menetapkan Regresi Linier sebagai model terpilih karena memberikan stabilitas (*robustness*) yang lebih dapat diandalkan untuk perencanaan jangka panjang dibandingkan model Polinomial yang terindikasi *overfitting*. Rincian perbandingan hasil proyeksi volume timbulan sampah untuk dua tahun mendatang disajikan pada Tabel 3.

Tabel 3. Perbandingan Metrik Evaluasi Model

Tahun Prediksi	Skenario Regresi Linier (Ton)	Skenario Regresi Polinomial (Ton)
2025	467.964,01	433.619,65
2026	510.664,96	435.107,36

Berdasarkan model terpilih (Regresi Linier), diperoleh proyeksi volume timbulan sampah sebesar 467.964,01 ton pada tahun 2025 dan meningkat menjadi 510.664,96 ton pada tahun 2026. Nilai proyeksi ini menjadi dasar penting dalam penyusunan rekomendasi kebijakan, terutama karena terdapat kesenjangan signifikan antara pertumbuhan timbulan sampah dan kapasitas armada pengangkut yang masih stagnan pada angka 119 unit, sebagaimana ditunjukkan oleh garis merah putus-putus pada Gambar 2.

4. Conclusion

Penelitian ini menyimpulkan bahwa pada kondisi dataset yang sangat terbatas (n=4), penggunaan metrik statistik semata tidak cukup untuk menentukan model prediksi terbaik. Temuan utama

menunjukkan adanya dilema antara akurasi dan stabilitas. Meskipun Regresi Polinomial memiliki akurasi historis tertinggi (R^2 = 0.99), analisis tren mengungkapkan risiko instabilitas yang tinggi (*overfitting*), ditandai dengan prediksi perlambatan pertumbuhan sampah yang tidak realistis di masa depan.

Sebaliknya, Regresi Linier direkomendasikan sebagai model yang paling *robust* untuk perencanaan pengelolaan sampah daerah. Model ini menawarkan stabilitas tren yang konsisten dengan laju pertumbuhan penduduk, meskipun memiliki *error* statistik yang sedikit lebih besar pada data latih. Berdasarkan model terpilih, volume timbulan sampah diproyeksikan akan terus meningkat hingga mencapai 467.964,01 ton pada tahun 2025 dan 510.664,96 ton pada tahun 2026. Proyeksi ini menjadi sinyal peringatan dini (*early warning*) bagi pemerintah daerah untuk segera meningkatkan kapasitas armada pengangkut yang saat ini stagnan di angka 119 unit, guna mencegah krisis akumulasi sampah yang tidak terkelola.

References

[1] "SIPSN - Sistem Informasi Pengelolaan Sampah Nasional."

[2] F. M. Aprileni and A. Nurhayati, "Penanganan Sampah Rumah Tangga Di Desa X Kabupaten Bandung," *J. Ris. Kesehat. Poltekkes Depkes Bandung*, vol. 16, no. 2, pp. 702–707, 2024, doi: 10.34011/juriskesbdg.v16i2.2635.

[3] N. Alhathlaul *et al.*, "Assessing Waste Management Using Machine Learning Forecasting for Sustainable Development Goal Driven," vol. 6, pp. 1–21, 2025.

[4] L. Choi, Wonjun; Sangwon, "Performance evaluation of deep learning architectures for load and temperature forecasting under dataset size constraints and seasonality," *Energy Build.*, p. 288, 2023.

[5] S. A. L. Satriyo, Adi Rizky Pratama, and Rahmat, "Perbandingan metode linear regresi dan polynomial regresi untuk memprediksi harga saham studi kasus Bank BCA," *INFOTECH J. Inform. Teknol.*, vol. 4, no. 1, pp. 59–70, 2023, doi: 10.37373/infotech.v4i1.602.

[6] J. K. Kang, Y. J. Lee, C. Y. Son, S. J. Park, and C. G. Lee, "Alternative assessment of machine learning to polynomial regression in response surface methodology for predicting decolorization efficiency in textile wastewater treatment," *Chemosphere*, vol. 370, no. December 2024, p. 143996, 2025, doi: 10.1016/j.chemosphere.2024.143996.

[7] N. M. Farhan and B. Setiaji, "Perbandingan Algoritma Regresi Linear dengan Algoritma Backpropagation Estimasi Timbulan Sampah di Sulawesi Utara," *Indones. J. Comput. Sci.*, vol. 12, no. 2, pp. 284–301, 2023, [Online]. Available: http://ijcs.stmikindonesia.ac.id/ijcs/index.php/ijcs/article/view/3135

[8] S. S. M. A. B. Bagus Irawan1*, "Review penggunaan machine learning dalam optimalisasi pengelolaan sampah perkotaan: studi literatur terkini," *Pros. Semin. Nas. Sains dan Teknol. Seri 02*, vol. 1, no. 2, pp. 543–553, 2024.

[9] A. Widianti and I. Pratama, "Penanganan Missing Values Dan Prediksi Data Timbunan Sampah Berbasis Machine Learning," *Rabit J. Teknol. dan Sist. Inf. Univrab*, vol. 9, no. 2, pp. 242–251, 2024, doi: 10.36341/rabit.v9i2.4789.

[10] G. G. Montgomery, Douglas C., Peck, Elizabeth A., & Vining, *Introduction to Linear Regression Analysis (6th ed.)*. Wiley, 2021.

[11] N. R. S. Jayadi, Puguh; Hidayati, "Data-Driven Approach of Machine Learning Techniques for Forecasting on Small Dataset: A Case Study of City Waste," *IEEE*, 2025.

[12] U. Rohman, J. Sunupurwa Asri, H. D. Ariessanti, and S. Wahyu, "Perbandingan Algoritma Regresi Linear Sederhana dan Regresi Polinomial Dalam Prediksi Jumlah Penumpang Kereta Api Di Wilayah Jabodetabek," *ProTekInfo(Pengembangan Ris. dan Obs. Tek. Inform.*, vol. 12, no. 1, pp. 7–15, 2025, doi: 10.30656/protekinfo.v12i1.9842.

[13] R. O. Budianto and L. P. Ghanistyana, "Peran Komunikasi Politik dalam Kampanye Isu Lingkungan: Studi Kasus pada Kebijakan Pengelolaan Sampah di Indonesia," *J. Bisnis dan Komun. Digit.*, vol. 2, no. 1, p. 11, 2024, doi: 10.47134/jbkd.v2i1.3219.

[14] Rahmayani, R., Firmansyah, E., & Hikmah, H. U. (2025). Inovasi Layanan Antar Jemput Paket Surat PT Pos Indonesia Berdasarkan Penjualan dan Minat Beli. JOVISHE: Journal of Visionary Sharia Economy, 4(1), 33-47.

[15] Firmansyah, E., Rahman, A. B. A., & Subiyakto, A. A. (2023). Pengukuran Kesiapan Kota Cerdas Berdasarkan SNI ISO 37122: 2019. Infoman's: Jurnal Ilmu-ilmu Informatika dan Manajemen, 17(2).

[16] Zulfikar, W. B., Irfan, M., Ghufron, M., Jumadi, J., & Firmansyah, E. (2020). Marketplace affiliates potential analysis using cosine similarity and vision-based page segmentation. Bulletin of Electrical Engineering and Informatics, 9(6), 2492-2498.