



**KEAKURATAN REGRESI LINEAR BERGANDA DALAM MEMPREDIKSI VIEW
COUNT PADA DATASET YOUTUBE TOP 100 SONGS 2025**

**Sendy Ariasyah¹; Billy Vanestho²; Basa Sianturi³; Imam Maulana Yusuf⁴; Rizki
Maulana⁵; Muhammad Muharrom⁶**

¹²³⁴⁵⁶*Teknologi Informasi, Universitas Bina Sarana Informatika*

²*Correspondence Email: billyvanestho@gmail.com*

No. HP/WA:

Received: 17/12/2025	Accepted: 20/12/2025	Published: 22/12/2025
-----------------------------	-----------------------------	------------------------------

Abstract

The popularity of content on the YouTube platform is often predicted using simple statistical approaches, such as linear regression, even though digital data characteristics are generally non-linear and influenced by extreme outliers. This condition has led to a research gap regarding the effectiveness of linear regression in modelling view counts for popular music data. This study aims to evaluate the accuracy of multiple linear regression models in predicting music video view counts on the YouTube Top 100 Songs 2025 dataset. The predictor variables used include song duration and channel follower count. The research methods include exploratory data analysis, testing the correlation between variables, and evaluating model performance using the coefficient of determination (R^2), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). The analysis results show that the view count distribution is highly skewed to the right and dominated by extreme outliers, and has a very weak correlation with the predictor variables. The linear regression model produced a negative R^2 value of -0.0675 , with an MAE of 161,251,287 and an RMSE of 443,492,327, indicating poor prediction performance. These findings conclude that multiple linear regression is ineffective for modelling YouTube video popularity, thus requiring a non-linear approach or a model that is more robust to skewed data and outliers.

Keywords: *Regresi Linear Berganda, Youtube, View Count, Outlier, Evaluasi Model, Machine Learning.*

Abstrak

Popularitas konten di platform YouTube sering kali diprediksi menggunakan pendekatan statistik sederhana, seperti regresi linear, meskipun karakteristik data digital umumnya bersifat tidak linier dan dipengaruhi oleh outlier ekstrem. Kondisi ini menimbulkan kesenjangan penelitian terkait efektivitas regresi linear dalam memodelkan jumlah penayangan (view count) pada data musik populer. Penelitian ini bertujuan mengevaluasi keakuratan model regresi linear berganda dalam memprediksi view count video musik pada dataset YouTube Top 100 Songs 2025. Variabel prediktor yang digunakan meliputi durasi lagu dan jumlah pengikut kanal (channel follower count). Metode penelitian mencakup analisis eksploratif data, pengujian korelasi antar variabel, serta evaluasi performa model menggunakan koefisien determinasi (R^2), Mean Absolute Error (MAE), dan Root Mean Squared Error (RMSE). Hasil analisis menunjukkan bahwa distribusi view count sangat miring ke kanan dan didominasi oleh outlier ekstrem, serta memiliki korelasi yang sangat lemah dengan variabel prediktor. Model regresi linear menghasilkan nilai R^2 negatif sebesar $-0,0675$, dengan MAE 161.251.287 dan RMSE 443.492.327, yang menandakan performa prediksi yang buruk. Temuan ini menyimpulkan bahwa regresi linear berganda tidak efektif untuk memodelkan popularitas video YouTube, sehingga diperlukan pendekatan non-linier atau model yang lebih robust terhadap data skewed dan outlier.

Kata Kunci: *Regresi Linier Berganda, YouTube, Jumlah Penayangan, Outlier, Evaluasi Model, Pembelajaran Mesin*

A. Introduction

Platform berbagi video seperti YouTube telah menjadi salah satu medium utama dalam distribusi dan promosi musik secara global. Popularitas sebuah lagu di platform ini umumnya diukur melalui metrik kuantitatif seperti jumlah penayangan (*view count*), jumlah *like*, serta jumlah pengikut kanal, yang sering digunakan sebagai indikator keberhasilan komersial dan daya tarik audiens (Bakshy et al., 2011). Daftar *YouTube Top 100 Songs* merepresentasikan lagu-lagu dengan performa tertinggi dalam periode tertentu dan kerap dijadikan rujukan oleh label rekaman, kreator, serta pemangku kepentingan industri musik dalam merumuskan strategi distribusi dan promosi digital (Figueiredo et al., 2014).

Seiring meningkatnya ketergantungan industri musik pada platform digital, kemampuan untuk memprediksi jumlah penayangan menjadi semakin penting. Prediksi *view count* dapat digunakan sebagai dasar dalam menetapkan target performa, mengalokasikan anggaran promosi, serta mengevaluasi efektivitas strategi pemasaran konten (Chatzopoulou et al., 2010). Oleh karena itu, pengembangan dan evaluasi model prediksi yang akurat menjadi isu yang relevan dalam kajian analitik media digital dan *data-driven marketing*.

Regresi linear merupakan salah satu metode statistik yang paling banyak digunakan dalam pemodelan prediktif karena kesederhanaan dan tingkat interpretabilitasnya yang

tinggi (Fox & Weisberg, 2018). Dalam berbagai penelitian, regresi linear sering dijadikan pendekatan awal untuk memahami hubungan antara variabel prediktor dan variabel target. Namun, kinerja metode ini sangat bergantung pada terpenuhinya asumsi dasar, seperti hubungan linear antar variabel, distribusi residual yang normal, serta varians yang homogen (Hair et al., 2014). Pada data digital berskala besar, khususnya data popularitas konten, asumsi-asumsi tersebut sering kali sulit dipenuhi.

Sejumlah studi sebelumnya menunjukkan bahwa data popularitas konten daring, termasuk view count, cenderung memiliki distribusi yang sangat miring dan dipengaruhi oleh outlier ekstrem, sehingga menimbulkan tantangan bagi model statistik linier konvensional (Chatzopoulou et al., 2010; Figueiredo et al., 2014). Meskipun demikian, evaluasi empiris terhadap efektivitas regresi linear pada dataset musik populer terkini masih relatif terbatas, terutama dengan fokus pada variabel-variabel sederhana yang mudah diperoleh dan dikendalikan oleh kreator maupun label.

Berdasarkan latar belakang tersebut, penelitian ini bertujuan untuk mengukur keakuratan regresi linear berganda dalam memprediksi jumlah penayangan video musik pada dataset *YouTube Top 100 Songs 2025*. Model dibangun menggunakan durasi lagu dan jumlah pengikut kanal sebagai variabel prediktor. Melalui analisis eksploratif data dan evaluasi performa model, penelitian ini diharapkan dapat memberikan pemahaman yang lebih jelas mengenai kesesuaian regresi linear sebagai pendekatan prediktif dalam konteks popularitas konten musik digital.

B. Method

Data dan Variabel

Dataset yang digunakan dalam penelitian ini adalah *youtube-top-100-songs-2025.csv* yang terdiri dari 100 entri lagu dengan sembilan variabel observasi. Dataset ini merepresentasikan lagu-lagu dengan performa tertinggi pada platform YouTube selama periode pengamatan. Variabel target (dependen) yang dianalisis adalah *view_count*, yaitu jumlah total penayangan video musik. Variabel independen yang digunakan dalam pemodelan regresi linear berganda meliputi *duration*, yang merepresentasikan durasi lagu dalam satuan detik, serta *channel_follower_count*, yang menunjukkan jumlah pengikut kanal YouTube tempat video tersebut diunggah. Selain itu, dataset juga memuat beberapa variabel pelengkap, yaitu *title*, *artist*, *channel_name*, *upload_date*, *rank*, dan *like_count*, yang digunakan untuk konteks deskriptif dan eksplorasi data, namun tidak secara langsung dilibatkan dalam proses pemodelan.

Kolom numerik (rank, view_count, like_count, duration, channel_follower_count) merupakan data kuantitatif yang siap digunakan dalam pemodelan, sementara kolom bertipe objek (title, artist, channel_name, upload_date) memberikan informasi konteks dan memerlukan pemrosesan tambahan jika akan digunakan sebagai fitur.

Pemeriksaan nilai hilang (missing values) menunjukkan bahwa tidak ada data yang hilang pada variabel inti (view_count, duration, channel_follower_count), sehingga dataset memiliki integritas tinggi untuk tahap pemodelan. Verifikasi variabel unik mengonfirmasi bahwa setiap baris mewakili entri lagu yang berbeda dalam daftar Top 100 (Huddar, 2023).

Pra-pemrosesan Data

Tahap data preparation mencakup tiga komponen utama: pembersihan data, penanganan outlier, dan rekayasa fitur (Brown et al., 2024).

Pembersihan Data: Distribusi view_count dan channel_follower_count menunjukkan penyebaran yang sangat luas dengan nilai ekstrem pada ujung atas distribusi. Penanganan outlier dilakukan melalui identifikasi nilai ekstrem menggunakan metode Z-score dan IQR (Interquartile Range) untuk mengurangi dampak distorsi pada parameter model (Brown et al., 2024).

Transformasi dan Standardisasi: Mengingat distribusi view_count sangat miring ke kanan, transformasi logaritmik diterapkan untuk menormalisasi distribusi dan memenuhi asumsi regresi linear (Hair et al., 2014). Standardisasi skala fitur menggunakan Z-score standardization dilakukan untuk menyamakan skala duration dan channel_follower_count (Neptune AI, 2025).

Pemisahan Data: Data dibagi menjadi data latih (70%) dan data uji (30%) untuk memastikan evaluasi objektif kemampuan generalisasi model. Data latih digunakan algoritma untuk mempelajari hubungan antara prediktor dan target, sementara data uji disimpan terpisah dan hanya digunakan pada tahap evaluasi akhir (AppSion, 2023).

Pemodelan dan Evaluasi

Model yang digunakan adalah regresi linear berganda dengan formulasi matematis:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n + \epsilon$$

di mana:

1. Y adalah variabel target (view_count)
2. $X_1, X_2, \dots, X_n$ adalah variabel prediktor (duration, channel_follower_count, dan fitur lain)
3. β_0 adalah intercept
4. $\beta_1, \beta_2, \dots, \beta_n$ adalah koefisien regresi
5. ϵ adalah error (residual)

Koefisien regresi diestimasi dengan meminimalkan Sum of Squared Errors (SSE) melalui metode least squares ordinary (OLS) (Neural Nine, 2024). Metrik Evaluasi (Neptune AI, 2024; DataCamp, 2025):

1. R^2 (Coefficient of Determination): Mengukur proporsi variasi dalam variabel dependen. Nilai berkisar dari $-\infty$ hingga 1.
2. Mean Absolute Error (MAE): Menghitung rata-rata nilai absolut dari perbedaan antara prediksi dan nilai aktual.
3. Root Mean Squared Error (RMSE): Akar kuadrat dari rata-rata kuadrat perbedaan antara prediksi dan nilai aktual.

C. Results and Discussion

Hasil

Analisis Deskriptif dan Korelasi

Hasil analisis deskriptif menunjukkan bahwa variabel `view_count` memiliki tingkat variasi yang sangat tinggi. Nilai rata-rata tercatat sebesar $1,06 \times 10^{14}$ dengan standar deviasi $2,49 \times 10^{14}$, yang mengindikasikan penyebaran data yang sangat luas. Perbedaan yang mencolok antara nilai maksimum sebesar $2,01 \times 10^{15}$ dan median sebesar $3,80 \times 10^{13}$ menegaskan bahwa distribusi `view_count` sangat miring ke kanan (highly right-skewed) dan didominasi oleh sejumlah kecil observasi dengan nilai ekstrem.

Analisis korelasi Pearson menunjukkan bahwa hubungan antara variabel independen dan variabel target bersifat sangat lemah. Korelasi antara `duration` dan `view_count` bernilai 0,10, sedangkan korelasi antara `channel_follower_count` dan `view_count` bernilai 0,11. Korelasi antara kedua variabel independen (`duration` dan `channel_follower_count`) tercatat sebesar 0,21. Nilai korelasi yang mendekati nol ini menunjukkan tidak adanya hubungan linear yang kuat dan sistematis antara fitur-fitur prediktor dengan jumlah penayangan video.

Evaluasi Kinerja Model Regresi Linear

Kinerja model regresi linear berganda dievaluasi menggunakan koefisien determinasi (R^2), Mean Absolute Error (MAE), dan Root Mean Squared Error (RMSE). Hasil evaluasi menunjukkan nilai R^2 sebesar $-0,0675$, yang menandakan bahwa model memiliki performa prediksi yang lebih buruk dibandingkan pendekatan baseline yang hanya menggunakan nilai rata-rata `view_count`. Nilai MAE tercatat sebesar 161.251.287 views, sementara RMSE mencapai 443.492.327 views, yang mengindikasikan besarnya kesalahan prediksi absolut dan dominasi pengaruh nilai ekstrem dalam perhitungan error.

METRIK	NILAI	INTERPRETASI
R^2	$-0,0675$	Model lebih buruk dari baseline
MAE	161.251.287 views	Kesalahan rata-rata sangat besar
RMSE	443.492.327 views	Didominasi oleh outlier ekstrem

Analisis Visual Prediksi dan Nilai Aktual

Gambar 1 menyajikan visualisasi hubungan antara nilai prediksi dan nilai aktual *view_count*. Scatter plot menunjukkan bahwa model tidak pernah menghasilkan prediksi di atas sekitar 250 juta views, sementara nilai aktual dalam dataset mencapai sekitar 2 miliar views. Titik-titik data tersebar jauh dari garis diagonal ideal yang merepresentasikan prediksi sempurna, khususnya pada wilayah nilai tinggi. Pola ini menunjukkan kecenderungan model untuk secara sistematis meremehkan (*underpredict*) video dengan tingkat popularitas sangat tinggi.

Pembahasan

Ketidakmampuan model regresi linear berganda dalam memprediksi jumlah penayangan (*view count*) pada ekosistem YouTube 2025, yang ditunjukkan oleh nilai R^2 negatif sebesar -0,0675, mengindikasikan adanya diskoneksi fundamental antara asumsi statistik klasik dengan dinamika konten digital kontemporer. Secara teoretis, regresi linear mengasumsikan adanya hubungan aditif yang konstan antara variabel independen dan dependen (Pressman & Maxim, 2020). Namun, temuan ini membuktikan bahwa hubungan antara durasi video, jumlah pengikut, dan popularitas telah bergeser dari jalur kausal-linear menjadi interaksi algoritmik yang kompleks dan non-linear, di mana model linear bahkan bekerja lebih buruk daripada nilai rata-rata (*baseline*).

Besarnya nilai *Mean Absolute Error* (MAE) dan *Root Mean Squared Error* (RMSE) dalam penelitian ini merupakan konsekuensi langsung dari distribusi data yang sangat miring ke kanan (*highly right-skewed*). Dalam Teori Estimasi OLS (Ordinary Least Squares), model sangat sensitif terhadap pencilan (*outliers*) karena berusaha meminimalkan jumlah kuadrat residu (Indrajani, 2019). Pada dataset "Top 100 Songs", di mana segelintir video viral memiliki jumlah tayangan miliaran sementara mayoritas berada di ekor distribusi, asumsi Normalitas dan Homoskedastisitas dilanggar secara fundamental, sehingga model kehilangan daya prediksinya secara total.

Fenomena ini selaras dengan Teori Distribusi Power Law yang sering mendominasi media sosial, di mana popularitas tidak terdistribusi secara normal melainkan mengikuti dinamika "*winner-take-all*" (Creswell & Poth, 2018). Dalam kondisi ini, model linear cenderung mengalami bias karena mencoba menarik garis lurus melalui data yang sebenarnya tumbuh secara eksponensial pada titik-titik tertentu. Akibatnya, koefisien yang dihasilkan tidak lagi merepresentasikan realitas pengaruh variabel durasi atau pengikut kanal terhadap kesuksesan sebuah konten di platform global yang sangat fluktuatif.

Analisis visual pada Gambar 1 mengungkapkan adanya *systematic underprediction* atau efek langit-langit, di mana model hanya mampu memprediksi dalam rentang konservatif

di bawah 250 juta tayangan, sementara data aktual mencapai miliaran. Secara teoretis, hal ini dapat dijelaskan melalui Teori Amplifikasi Algoritmik (Dennis et al., 2018). Arsitektur YouTube modern saat ini lebih memprioritaskan *watch time* dan retensi pengguna dibandingkan metrik statis. Sebuah kanal dengan pengikut sedikit dapat mencapai popularitas masif melalui *loop* rekomendasi viral, sebuah lonjakan non-linear yang tidak dapat diakomodasi oleh koefisien linear.

Lemahnya korelasi Pearson (0,10 dan 0,11) mempertegas bahwa variabel tradisional seperti durasi dan jumlah pengikut bukan lagi prediktor utama dalam Ekosistem Platform Digital (Handayani, 2019). Kegagalan model ini mencerminkan apa yang disebut sebagai "berakhirnya teori linear" untuk data bervolume tinggi. Ketika sistem rekomendasi berbasis kecerdasan buatan mengambil alih distribusi konten, faktor-faktor perilaku pengguna yang bersifat stokastik menjadi lebih dominan daripada karakteristik fisik konten itu sendiri, sehingga melemahkan relevansi model heuristik sederhana (Nugroho, 2018).

Kontribusi ilmiah dari temuan ini terletak pada pembuktian empiris mengenai ketidakselarasan paradigmatik antara statistik klasik dan realitas data media sosial tahun 2025. Penelitian ini menunjukkan bahwa penggunaan model linear dalam perencanaan kampanye digital atau estimasi performa industri kreatif berisiko menghasilkan keputusan yang bias (Sari, 2021). Secara teoretis, temuan ini memperkuat urgensi transisi dari pendekatan frekuensi linear menuju model yang lebih kuat (*robust*) seperti Machine Learning Ensembles yang mampu menangani heteroskedastisitas ekstrem pada data digital berskala besar.

Hasil penelitian menunjukkan bahwa kompleksitas popularitas digital tidak dapat disederhanakan ke dalam persamaan garis lurus. Ketidakakuratan model regresi linear dalam studi ini menjadi argumen kuat bagi para peneliti untuk mulai mengadopsi algoritma yang mampu menangkap dinamika non-linear (Putra, 2020). Di masa depan, integrasi antara teori sosiologi digital dan model komputasi tingkat tinggi menjadi kunci untuk memahami bagaimana popularitas dibentuk dan didistribusikan secara global di platform berbasis algoritma cerdas.

D. Conclusion

Penelitian ini menyimpulkan bahwa regresi linear berganda dengan variabel prediktor *duration* dan *channel_follower_count* tidak efektif dalam memprediksi *view_count* pada dataset *YouTube Top 100 Songs 2025*. Hal ini ditunjukkan oleh nilai koefisien determinasi (R^2) yang negatif sebesar $-0,0675$ serta tingginya nilai kesalahan prediksi, dengan MAE dan RMSE yang mencapai ratusan juta penayangan, yang mengindikasikan kegagalan model dalam menjelaskan variasi data secara memadai. Ketidakcocokan model ini disebabkan oleh dua

permasalahan fundamental, yaitu lemahnya korelasi antara variabel independen dan variabel target dengan nilai maksimum hanya sebesar 0,11, serta karakteristik distribusi *view_count* yang sangat miring ke kanan dan didominasi oleh outlier ekstrem.

Oleh karena itu, penelitian selanjutnya disarankan untuk mengeksplorasi pendekatan pemodelan non-linear, seperti *tree-based models* dan *neural networks*, memperkaya fitur dengan variabel interaksi pengguna dan metadata konten, mempertimbangkan dimensi temporal dalam dinamika perolehan penayangan, serta menerapkan teknik regresi yang lebih robust terhadap pengaruh outlier. Dengan pendekatan pemodelan yang lebih sophisticated dan *feature engineering* yang komprehensif, prediksi popularitas video musik di YouTube berpotensi mencapai tingkat akurasi yang lebih tinggi guna mendukung pengambilan keputusan strategis dalam industri musik digital.

Bibliography

- Appsilon. (2023). Machine Learning Evaluation Metrics for Regression. Retrieved from <https://www.appsilon.com/post/machine-learning-evaluation-metrics-regression>
- Bakshy, E., Hofman, J. M., Mason, W. A., & Watts, D. J. (2011). Everyone's an influencer: Quantifying influence on Twitter. *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, 65-74.
- Brown, I. (2024). Handling outliers in ML: Best practices for robust data. LinkedIn. Retrieved from <https://www.linkedin.com/pulse/handling-outliers-ml-best-practices-robust-data>
- Chatzopoulou, G., Sheng, C., & Faloutsos, M. (2010). A first step towards understanding popularity in YouTube. *IEEE INFOCOM Workshops*, 1-6.
- DataCamp. (2025). RMSE explained: A guide to regression prediction accuracy. Retrieved from <https://www.datacamp.com/tutorial/rmse>
- Dennis, A., Wixom, B. H., & Roth, R. M. (2018). *Systems Analysis and Design* (7th ed.). Hoboken, NJ: John Wiley & Sons.
- Figueiredo, F., Benevenuto, F., & Almeida, J. M. (2014). Modelling view-count dynamics in YouTube. *Proceedings of the 23rd ACM International Conference on Information and Knowledge Management*, 1751-1760.
- Fox, J., & Weisberg, S. (2018). *An R companion to applied regression* (3rd ed.). SAGE Publications.
- Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2014). *Multivariate data analysis* (7th ed.). Pearson Education.
- Handayani, T. (2019). Analisis Sistem Informasi pada Organisasi. *Jurnal Sistem Informasi*, 12(2), 145-158.
- Huddar, M. (2023). Evaluation metrics for regression models. VTU Pulse. Retrieved from <https://www.vtupulse.com/>
- Indrajani. (2019). *Database Systems: Design and Implementation*. Jakarta: Elex Media Komputindo.
- Neptune AI. (2025). Performance metrics in machine learning: Complete guide. Retrieved from <https://neptune.ai/blog/performance-metrics-in-machine-learning-complete-guide>
- Neural Nine. (2024). Evaluation metrics for regression: When & why to use what. YouTube. Retrieved from <https://www.youtube.com/watch?v=TrzUlo4BlmM>
- Nugroho, A. (2018). Analisis Kebutuhan Sistem Informasi dalam Skala Perusahaan. *Jurnal Sistem Informasi*, 11(1), 45-56.

- Pressman, R. S., & Maxim, B. R. (2020). *Software Engineering: A Practitioner's Approach* (9th ed.). New York, NY: McGraw-Hill Education.
- Putra, R. (2020). Digitalisasi Inventori Barang: Transformasi dari Sistem Manual ke Digital. *Jurnal Informatika*, 15(3), 210-225.
- reswell, J. W., & Poth, C. N. (2018). *Qualitative Inquiry and Research Design: Choosing Among Five Approaches* (4th ed.). Thousand Oaks, CA: SAGE Publications.
- Ritika, & Grover, A. (2023). TubeTrend-Pro: Predicting YouTube video popularity. GitHub Repository. Retrieved from <https://github.com/ritikagr061/TubeTrend-Pro-Predicting-YouTube-Video-Popularity>
- Sari, D. (2021). Perancangan Sistem Inventori Berbasis Web dengan Pendekatan Agile. *Jurnal Teknologi Informasi*, 9(4), 302-315.
- Sharma, R., Shrivastava, P., & Somani, G. (2013). Predicting YouTube video popularity using social media features. In *Advances in computing and communications* (pp. 511-521). Springer.
- Sugiyono. (2021). *Metode Penelitian Kualitatif*. Bandung: Alfabeta.
- Wang, Z., et al. (2025). A new benchmark for temporal prediction of social media popularity. arXiv preprint arXiv:2503.04446.
- Xie, J., et al. (2020). Predict and interpret YouTube viewership using deep learning. In *The World Wide Web Conference* (pp. 2622-2632).