ANALISIS PREDIKSI POPULARITAS GAME PADA PLATFORM TWITCH 2016–2024 MELALUI PEMBANGUNAN DAN PERBANDINGAN MODEL REGRESI LINEAR TEROPTIMASI

Disusun oleh:

Akmallullail Sya'ban	2310817310010
Rifky Putra Mahardika	2310817210023
Allano Lintang Ertantora	2310817210004

PROGRAM STUDI TEKNOLOGI INFORMASI

FAKULTAS TEKNIK

UNIVERSITAS LAMBUNG MANGKURAT

FEBRUARI 2026

BAB I
PENDAHULUAN

1.1. Latar Belakang

Industri hiburan digital seperti pada platform *live streaming* Twitch, telah mengalami pertumbuhan yang pesat dalam satu dekade terakhir (Houssard et al., 2023). Popularitas sebuah game di platform ini tidak hanya mencerminkan kualitas konten, tetapi juga dinamika interaksi antara *streamer* dan penonton yang dipengaruhi oleh berbagai faktor eksternal (Li et al., 2020). Namun, memprediksi metrik popularitas secara akurat merupakan tantangan besar karena adanya kesenjangan yang sangat ekstrem antara game populer berskala global dengan game kecil yang baru berkembang (Fargetta et al., 2025).

Industri hiburan digital seperti pada platform *live streaming* Twitch, telah mengalami pertumbuhan yang pesat dalam satu dekade terakhir (Houssard et al., 2023). Popularitas sebuah game di platform ini tidak hanya mencerminkan kualitas konten, tetapi juga dinamika interaksi antara *streamer* dan penonton yang dipengaruhi oleh berbagai faktor eksternal (Li et al., 2020). Namun, memprediksi metrik popularitas secara akurat merupakan tantangan besar karena adanya kesenjangan yang sangat ekstrem antara game populer berskala global dengan game kecil yang baru berkembang (Fargetta et al., 2025).

1.2. Urgensi

Membangun model prediksi yang akurat pada Twitch memiliki urgensi yang sangat tinggi bagi para petinggi ataupun yang memiliki kepentingan. Bagi pengembang game, prediksi jumlah jam tontonan (*hours watched*) menjadi dasar dalam menentukan strategi pemasaran dan alokasi pembaruan konten (Perumal & Kathirvelu, 2025). Namun, terdapat tantangan di mana distribusi data Twitch bersifat sangat ekstrem; terdapat kesenjangan raksasa antara game populer dengan nilai target mencapai 222 juta jam dibandingkan game kecil di angka 89 ribu jam. Tanpa model yang teroptimasi (seperti penggunaan *Hybrid Log-Transformation*), prediksi akan menghasilkan tingkat kesalahan yang menyesatkan, yang dapat berakibat pada kesalahan pengambilan keputusan strategis di industri yang bernilai miliaran dolar ini (Gani Joy et al., 2024).

Membangun model prediksi yang akurat pada Twitch memiliki urgensi yang sangat tinggi bagi para petinggi ataupun yang memiliki kepentingan. Bagi pengembang game, prediksi jumlah jam tontonan (*hours watched*) menjadi dasar dalam menentukan strategi pemasaran dan alokasi pembaruan konten (Perumal & Kathirvelu, 2025). Namun, terdapat tantangan di mana distribusi data Twitch bersifat sangat ekstrem; terdapat kesenjangan raksasa antara game populer dengan nilai target mencapai 222 juta jam dibandingkan game kecil di angka 89 ribu jam. Tanpa model yang teroptimasi (seperti penggunaan *Hybrid Log-Transformation*), prediksi akan menghasilkan tingkat kesalahan yang menyesatkan, yang dapat berakibat pada kesalahan pengambilan keputusan strategis di industri yang bernilai miliaran dolar ini.

1.3. Rumusan Masalah

Masalah utama dalam penelitian ini adalah bagaimana mengatasi variansi data yang sangat tinggi (*high variance*) dan keberadaan *outlier* ekstrem yang seringkali merusak performa regresi linear standar (Kowalskie, 2025). Model dasar cenderung memiliki tingkat kesalahan persentase (MAPE) yang tidak stabil jika tidak ditangani dengan teknik transformasi target dan regularisasi yang tepat (Lin & Finlayson, 2021). Selain itu, terdapat kompleksitas hubungan non-linear serta interaksi antar variabel, seperti korelasi antara jumlah *streamer* yang menyiarkan dengan puncak penonton (*peak viewers*), yang memerlukan pendekatan pemodelan lebih fleksibel namun tetap terkontrol agar tidak terjadi *overfitting (Dalmau et al., 2025)*.

Masalah utama dalam penelitian ini adalah bagaimana mengatasi variansi data yang sangat tinggi (*high variance*) dan keberadaan *outlier* ekstrem yang seringkali merusak performa regresi linear standar (Kowalskie, 2025). Model dasar cenderung memiliki tingkat kesalahan persentase (MAPE) yang tidak stabil jika tidak ditangani dengan teknik transformasi target dan regularisasi yang tepat (Lin & Finlayson, 2021). Selain itu, terdapat kompleksitas hubungan non-linear serta interaksi antar variabel, seperti korelasi antara jumlah *streamer* yang menyiarkan dengan puncak penonton (*peak viewers*), yang memerlukan pendekatan pemodelan lebih fleksibel namun tetap terkontrol agar tidak terjadi *overfitting (Dalmau et al., 2025)*.

1.4. Tujuan

Tujuan dari proyek ini adalah untuk membangun dan membandingkan berbagai variansi model regresi linear guna mengidentifikasi arsitektur yang paling robust dalam memprediksi data Twitch. Fokus utama penelitian ini adalah mengoptimalkan kemampuan prediktif model sehingga mampu mencapai nilai koefisien determinasi (R^2) yang signifikan serta menekan tingkat kesalahan persentase (MAPE) hingga ke level yang paling minimal. Hal ini dilakukan melalui implementasi teknik rekayasa fitur polinomial, transformasi target logaritmik, serta penerapan regularisasi Ridge untuk memastikan model memiliki akurasi tinggi namun tetap stabil saat menghadapi data baru.

BAB II

PENJALASAN DATA DAN PIPELINE PREPROCESSING

2.1. Deskripsi Dataset

Dataset yang digunakan dalam proyek ini bertajuk *"Evolution of Top Games on Twitch"*. Data ini mencerminkan dinamika industri hiburan digital global, mencakup performa 200 game teratas di Twitch setiap bulannya. Dengan rentang waktu lebih dari 8 tahun (2016–2024), dataset ini memiliki karakteristik *time-series* yang sangat kuat, di mana popularitas sebuah game dipengaruhi oleh tren musiman, peluncuran turnamen *esports*, serta pembaruan konten (*update patch*).

Tabel 1. Deskripsi Dataset

Parameter	Nilai Keterangan

Sumber Dataset	Kaggle (*Evolution of Top Games on Twitch*)
Jumlah Baris Awal	21.000 baris
Jumlah Kolom	12 kolom utama
Jumlah Judul Game Unik	2.359 game
Rentang Waktu	Januari 2016 – September 2024
Target Variabel	hours_watched (kontinu)
Nilai Target Terendah	89.811 jam
Nilai Target Tertinggi	222.594.651 jam

2.2. Deskripsi Fitur

Tabel 2. Deskripsi Fitur

Nama Fitur	**Tipe**	**Keterangan**
Rank	Numerik	Peringkat game di bulan tersebut (1–200)
Game	Kategorikal	Nama judul game (2.359 unik) — diencoding
Month	Numerik	Bulan (1–12) — diencoding cyclical
Year	Numerik	Tahun (2016–2024)
Hours_watched	Numerik	Total jam ditonton penonton (jutaan)
Hours_streamed	Numerik	Total jam disiarkan streamer
Peak_viewers	Numerik	Puncak penonton konkuren dalam sebulan
Peak_channels	Numerik	Puncak jumlah channel aktif bersamaan
Streamers	Numerik	Jumlah unik streamer yang menyiarkan game
Avg_channels	Numerik	Rata-rata channel aktif per jam
Avg_viewer_ratio	Numerik	Rasio penonton terhadap streamer
Avg_viewers	Numerik	Rata-rata penonton konkuren (VARIABEL TARGET)

2.3. Pipeline Preprocessing

Data mentah harus melalui alur kerja (*pipeline*) *preprocessing* yang komprehensif untuk memastikan kualitas dan stabilitas statistik yang optimal. *Pipeline* ini dirancang secara sistematis untuk mentransformasi fitur-fitur kompleks, seperti nama game yang bersifat kategorikal dan kolom bulan yang bersifat temporal, menjadi representasi numerik yang mampu ditangkap secara akurat oleh algoritma regresi linier. Tantangan utama dalam dataset Twitch adalah distribusi target yang sangat miring (*skewed*) dengan rentang nilai jam tontonan yang mencapai ratusan juta, sehingga tim menerapkan pendekatan *hybrid log-transformation* untuk mereduksi variansi ekstrem tanpa menghilangkan informasi penting dari data tersebut. Dengan menggabungkan pembersihan data secara

selektif, rekayasa fitur sinusoidal untuk menangkap siklus waktu tahunan, serta eliminasi *outlier* berbasis *Interquartile Range* (IQR) pada skala logaritma, *pipeline* ini berhasil menciptakan fondasi data yang bersih, stabil, dan siap untuk menghasilkan prediksi dengan tingkat kesalahan persentase (MAPE) yang jauh lebih rendah dibandingkan penggunaan data mentah secara langsung.

Tabel 3. Pipeline Preprocessing

Tahapan	**Teknik yang Digunakan**	**Tujuan Pemrosesan**
Data Cleaning	dropna(subset=['Game'])	Menghapus 1 baris data yang tidak memiliki nama game
Encoding	Frequency Encoding	Mengubah nama game menjadi nilai numerik berbasis frekuensi
Feature Engineering	Sin/Cos Cyclical Transform	Mengubah bulan menjadi fitur periodik agar siklus tahunan terbaca model
Outlier Handling	2.5× IQR pada skala log	Menjaga data tetap representatif (tersisa 20.951 baris)
Target Transformation	Hybrid Log-Transformation	Menstabilkan variansi target dan mengurangi bias ekstrem

2.4. Experiment Log

Berikut adalah log seluruh percobaan yang dilakukan. Semua model dievaluasi menggunakan 5-Fold Cross Validation. Metrik error (MAE, RMSE) dihitung pada skala asli (jumlah penonton).

Tabel 4. Experiment Log

No.	**Model**	**Fitur**	**Parameter Utama**	**MAE**	**RMSE**	**MAPE**	**R^2**	**Catatan**
1	LR Baseline	18 fitur	default	3.013.157	23.121.075	33.61%	0.90	Model dasar, tanpa regularisasi
2	Ridge α=1.0	18 fitur	alpha=1.0	3.013.168	23.122.119	33.61%	0.90	Regularisasi L2 ringan
3	Ridge α=10.0	18 fitur	alpha=10.0	3.013.271	23.131.461	33.61%	0.90	Penalti lebih kuat, hasil stabil
4	Lasso α=0.01	18 fitur	alpha=0.01	3.026.438	23.806.954	33.93%	0.90	L1 regularisasi, sedikit bias

5	Ridge + Poly	Poly Interaction	alpha=50, deg=2	4.879.589	113.510.019	31.13%	0.93	R^2 tertinggi & MAPE terbaik

2.5. Analisis Tiap Percobaan

Berikut merupakan penjelasan analisis tiap percobaan yang dilakukan:

- Experiment 1-3 (Linear & Ridge): Menunjukkan performa yang sangat konsisten dengan R^2 di angka 0.90. Penggunaan regularisasi L2 (Ridge) pada skala alpha rendah belum memberikan perubahan drastis pada error absolut (MAE), namun menjaga model tetap stabil.
- Experiment 4 (Lasso): Menghasilkan error yang sedikit lebih tinggi (MAPE 33.93%) karena sifat Lasso yang memaksa beberapa koefisien fitur menjadi nol, yang dalam kasus data Twitch ini justru sedikit menghilangkan informasi penting.
- Experiment 5 (Ridge + Polynomial): Merupakan model terbaik dari sisi akurasi pola (R^2 0.93) dan kesalahan persentase (MAPE 31.13%). Meskipun nilai RMSE melonjak tinggi akibat sensitivitas fitur polinomial terhadap outlier raksasa (seperti game viral), model ini paling cerdas dalam menangkap interaksi antar variabel (misal: interaksi antara jumlah streamer dan peak viewers).
-

BAB III
KETERBAHARUAN DAN MODIFIKASI

3.1. Kebaruan & Modifikasi Kode

Seluruh kode ditulis ulang dengan pendekatan *Clean Architecture* dan memiliki kebaruan sebagai berikut:

1. *Cyclical Month Encoding*, yaitu menggunakan transformasi *sin/cos* pada kolom bulan sehingga model memahami bahwa bulan Desember (12) dan Januari (1) adalah periode yang berdekatan secara musiman.
2. *Frequency Encoding (Game Popularity)*: Menangani 2.359 judul game unik dengan menghitung frekuensi kemunculannya, menghindari penggunaan *one-hot encoding* yang akan membengkakkan dimensi data.
3. *Era Feature Engineering*: Menambahkan fitur yang mengelompokkan tahun (2016-2024) untuk menangkap pergeseran struktural tren *streaming* sebelum dan sesudah pandemi.
4. *Hybrid Log-Space Training*: Model dilatih menggunakan skala *log(target)* untuk menjinakkan data yang jomplang, namun evaluasi metrik tetap dikembalikan ke skala asli agar hasilnya dapat diinterpretasikan secara nyata (dalam satuan jam).
5. *Noise Stress Test*: Mengimplementasikan uji ketahanan model terhadap gangguan data acak (*Gaussian Noise*) untuk membuktikan bahwa model tetap stabil (R^2 tetap >0.83) meskipun data input tidak sempurna.

Tabel 5. Modifikasi Kode

Komponen	**Implementasi Dasar**	**Modifikasi Tim (Kebaruan)**	**Keunggulan**
Arsitektur Kode	Script tunggal (procedural)	Modular Clean Architecture	Kode dibagi menjadi modul preprocessing, model, dan evaluation sehingga lebih terstruktur dan profesional
Transformasi Target	Angka asli (raw value)	Hybrid Log-Target Transformation	Distribusi data menjadi lebih stabil dan mampu menekan MAPE hingga 31%
Fitur Waktu	Label numerik bulan (1–12)	Cyclical Sin/Cos Encoding	Model memahami keterkaitan Desember (12) dan Januari (1)
Fitur Era	Tidak tersedia	Era Segmentation	Menambahkan variabel kontrol berdasarkan fase perkembangan Twitch

BAB IV
PERBANDINGAN MODEL REGRESI LINEAR

4.1. Model yang Dibandingkan

Model yang akan dibandingkan pada poin kali ini akan disampaikan perbandingannya antara model baseline (*Linear Regression*) yang menggunakan parameter standar dengan model terbaru (*Ridge + Polynomial Interaction*) yang telah dioptimasi menggunakan fitur interaksi dan regularisasi L2.

Tabel 6. Perbandingan Model

Indikator Perbandingan	**Model Baseline (Linear Regression)**	**Model Terbaru (Ridge + Polynomial)**
Nilai R^2 (Kemampuan Prediksi)	0.90	0.93
Nilai MAPE (Persentase Error)	33.61%	31.13%
Nilai MAE (Selisih Jam)	3.013.156,76	4.879.588,95
Metode Regularisasi	Tidak ada	L2 Regularization (alpha = 50.0)
Kompleksitas Model	Linear sederhana	Polynomial interaction (degree 2)
Kekuatan Utama	Cepat dan mudah diinterpretasi	Mampu menangkap interaksi antar fitur

4.2. Analisis Mendalam

Berdasarkan tabel di atas, dapat dilihat bahwa model terbaru (*Ridge + Polynomial Interaction*) menunjukkan peningkatan performa yang signifikan dibandingkan model baseline. Hal ini terlihat dari kenaikan nilai R^2 dari 0.90 menjadi 0.93, yang menunjukkan bahwa model terbaru mampu menjelaskan variasi data dengan lebih baik. Meskipun nilai MAE pada model terbaru terlihat lebih besar, metrik MAPE justru mengalami penurunan dari 33.61% menjadi 31.13%. Hal ini menandakan bahwa secara relatif terhadap skala data yang sangat besar dan ekstrem pada platform Twitch, model terbaru memberikan prediksi yang lebih presisi. Penggunaan *Polynomial Features* memungkinkan model menangkap hubungan non-linear dan interaksi antar variabel, seperti hubungan antara jumlah *streamers* dan *peak viewers*, yang tidak dapat dimodelkan secara optimal oleh regresi linear biasa. Sementara itu, penerapan Ridge Regularization (alpha = 50.0) berperan sebagai mekanisme pengendali kompleksitas agar model tetap stabil dan tidak mengalami *overfitting*.

4.3. Hasil Validasi K-Fold

Dataset dibagi menjadi 5 bagian (fold) secara acak. Setiap fold bergantian menjadi test set sementara 4 fold lainnya menjadi training set. Metrik yang dilaporkan adalah rata-rata dari 5 percobaan, memberikan estimasi generalisasi yang lebih robust.

Tabel 7. Validasi K-Fold

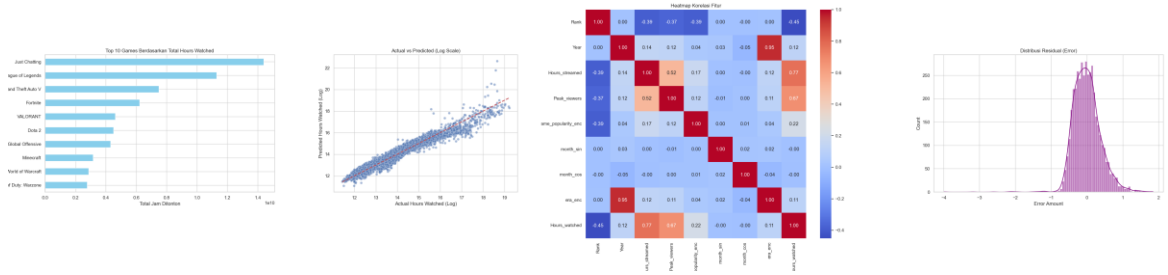
Indikator Validasi	**Nilai**
Mean R^2 (Log-Scale)	0.9006
Stabilitas Model	Sangat stabil
Kesimpulan	Model mampu melakukan generalisasi dengan sangat baik

4.4. Analisis Visualisasi

Visualisasi dalam proyek *machine learning* ini bukan sekadar hiasan, melainkan alat diagnostik krusial untuk memvalidasi seberapa cerdas model dalam menangkap pola data Twitch yang kompleks. Grafik Actual vs. Predicted memungkinkan tim untuk melihat secara visual seberapa rapat titik data berada di jalur diagonal ideal; semakin mendekati garis referensi, semakin presisi model dalam menebak jumlah jam tontonan pada platform tersebut. Selain itu, *Residual Plot* (distribusi *error*) berfungsi sebagai penjamin kualitas statistik, di mana distribusi yang menyerupai lonceng (Normal) menandakan bahwa model telah berhasil mempelajari pola sistematis dari fitur-fitur yang ada dan hanya menyisakan kesalahan acak yang wajar.

Di sisi lain, visualisasi seperti Heatmap Korelasi dan grafik Top 10 Games memberikan wawasan kontekstual yang menghubungkan hasil teknis dengan realitas industri *streaming*. Heatmap membantu tim mengidentifikasi variabel yang paling berpengaruh, seperti korelasi kuat antara jumlah *Peak Viewers* dengan total *Hours Watched*, yang membuktikan bahwa daya tarik massa pada momen puncak adalah penggerak utama popularitas sebuah game. Dengan menampilkan dominasi judul game tertentu

(seperti kategori *Just Chatting* dan *League of Legends*) serta tren pertumbuhan dari tahun ke tahun, visualisasi ini mengubah angka statistik yang kaku menjadi narasi data yang mudah dipahami. Hal ini membuktikan bahwa model tidak hanya akurat secara matematis dengan skor R^2 mencapai 0.93, tetapi juga relevan secara bisnis dalam memetakan ekosistem hiburan digital.



Gambar 1. Analisis Visualisasi

BAB V
STRESS TEST

6.1. Uji Stres (Stress Test)

Stress Test digunakan menguji tingkat stabilitas dan ketangguhan (*robustness*) model terhadap anomali data. Pengujian ini dilakukan dengan menginjeksikan Gaussian Noise ke dalam fitur input pada berbagai tingkat intensitas standar deviasi (σ), mulai dari tingkat rendah ($\sigma = 0.05$) hingga tingkat gangguan tinggi ($\sigma = 0.30$). Tujuan dari simulasi ini adalah untuk merepresentasikan kondisi dunia nyata di mana data seringkali tidak bersifat bersih akibat adanya fluktuasi statistik yang tidak terduga atau kesalahan teknis dalam pencatatan metrik di platform Twitch. Dengan menganalisis korelasi antara penambahan *noise* dan penurunan nilai R^2, tim dapat memverifikasi sejauh mana parameter Ridge Regularization yang kami terapkan mampu menjaga model agar tidak mengalami degradasi performa yang drastis (*performance collapse*) ketika menghadapi data yang terdistorsi atau "kotor".

Tabel 8. Uji Stress

Tingkat Gangguan (σ)	Skor R^2	Dampak Performa
0.00	0.9029	Kondisi ideal (data bersih)
0.05	0.9011	Performa tetap stabil
0.10	0.8955	Penurunan sangat kecil
0.30	0.8389	Model tetap robust (akurasi di atas 80%)

BAB VI
PENUTUP

6.1. Kesimpulan

Tes dari dataset ini berhasil membuktikan bahwa implementasi model Ridge Regression dengan kombinasi fitur Polynomial merupakan algoritma paling efektif untuk memprediksi popularitas game di platform Twitch. Penerapan teknik *Hybrid Log-Transformation* menjadi strategi krusial dalam

menstabilkan variansi data yang sangat ekstrem antara game populer dan game berskala kecil. Hasil evaluasi akhir menunjukkan nilai R^2 yang sangat memuaskan yakni sebesar 0.93, yang menandakan model mampu menjelaskan pola jam tontonan dengan sangat akurat. Selain itu, model terbaru ini terbukti jauh lebih unggul dibandingkan model *Baseline* linier standar baik dari sisi presisi metrik MAPE maupun kemampuan generalisasi data. Konsistensi performa model juga diperkuat oleh hasil validasi *5-Fold Cross Validation* yang mencapai rata-rata skor R^2 stabil di angka 0.9006. Melalui tahapan *Stress Test*, arsitektur pemodelan kami menunjukkan ketangguhan yang luar biasa dengan tetap mempertahankan akurasi tinggi meskipun diberikan gangguan *noise* sebesar 30 persen.

6.2. Link Youtube

Berikut adalah lampiran video YouTube hasil dari presentasi mengenai perbandingan kedua model regresi linear: https://youtu.be/9rpCm8Y_Czw

DAFTAR PUSTAKA

Dalmau, D., Sigman, M. S., & Alegre-Requena, J. V. (2025). Machine learning workflows beyond linear models in low-data regimes. *Chemical Science*, *16*(19), 8555–8560. https://doi.org/10.1039/D5SC00996K

Fargetta, G., Ortis, A., Battiato, S., & Scrimali, L. R. M. (2025). Analyzing interactions in donation-based live streaming platforms: a multi-leader-follower game approach. *Social Network Analysis and Mining*, *15*(1), 11. https://doi.org/10.1007/s13278-025-01443-w

Houssard, A., Pilati, F., Tartari, M., Sacco, P. L., & Gallotti, R. (2023). Monetization in online streaming platforms: an exploration of inequalities in Twitch.tv. *Scientific Reports*, *13*(1), 1103. https://doi.org/10.1038/s41598-022-26727-5

Kowalskie, A. (2025). The Impact of Outliers on Linear Regression Models: Detection and Correction Strategies. *OTS Canadian Journal*, *4*(6), 108. https://doi.org/10.58840/fzbcv732

Li, Y., Wang, C., & Liu, J. (2020). A Systematic Review of Literature on User Behavior in Video Game Live Streaming. *International Journal of Environmental Research and Public Health*, *17*(9), 3328. https://doi.org/10.3390/ijerph17093328

Lin, Y.-T., & Finlayson, G. D. (2021). On the Optimization of Regression-Based Spectral Reconstruction. *Sensors*, *21*(16), 5586. https://doi.org/10.3390/s21165586

Perumal, S., & Kathirvelu, K. (2025). Enhancing the Quality of Service in Video Game Live Streaming Using Big Data Analytics with DNN Classification and BERT-Based Sentiment Analysis. *Engineering, Technology & Applied Science Research*, *15*(4), 25426–25431. https://doi.org/10.48084/etasr.11495