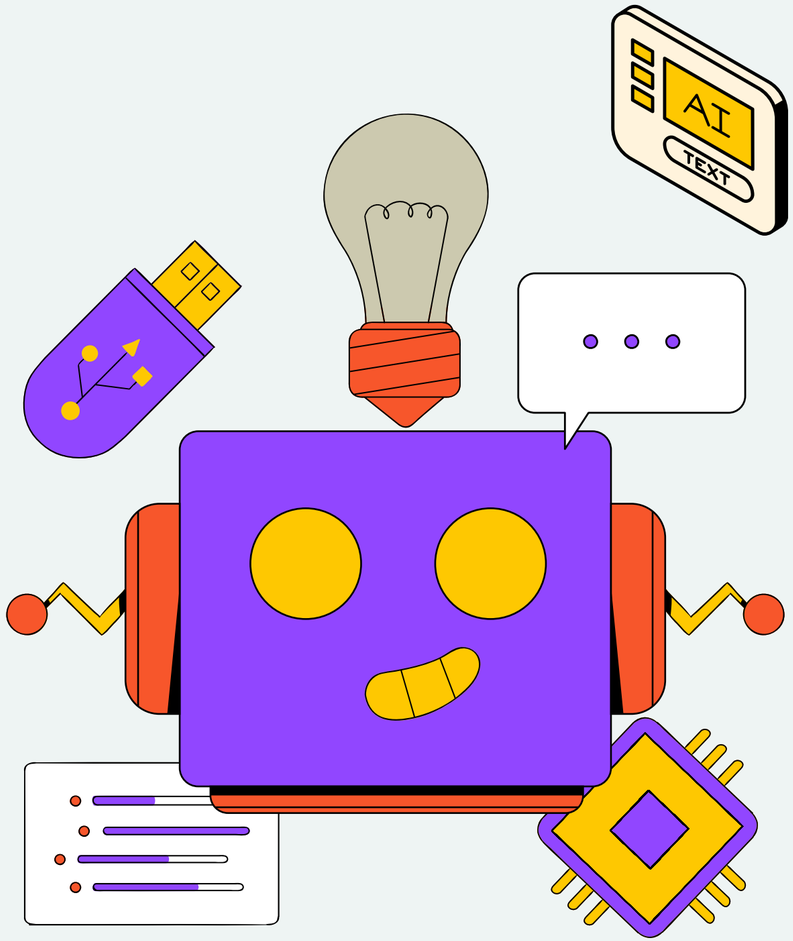


MEMBANGUN DAN MEMBANDINGKAN MODEL REGRESI LINIER: PREDIKSI POPULARITAS GAME DI TWITCH.

PRESENTED BY:

KELOMPOK 4 (BALANGAN)

ANGGOTA KELOMPOK

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

LATAR BELAKANG

Industri hiburan digital seperti pada platform live streaming Twitch, telah mengalami pertumbuhan yang pesat dalam satu dekade terakhir. 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. 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.

URGENSI

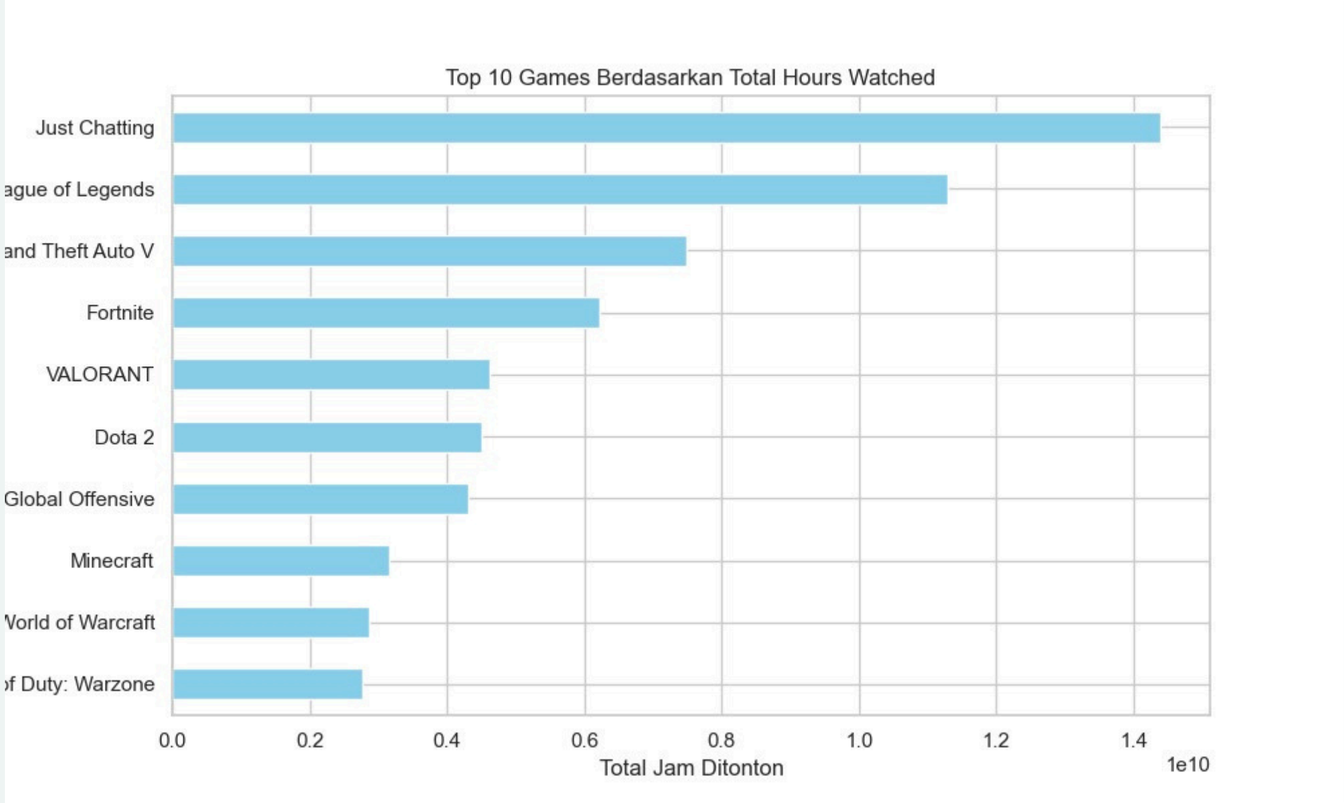
1. Kegagalan memprediksi jam tontonan secara presisi berisiko merusak efektivitas strategi pemasaran serta menyebabkan kesalahan alokasi konten bagi pengembang game
2. Kesenjangan distribusi data yang ekstrem menuntut penggunaan teknik Hybrid Log-Transformation untuk memitigasi risiko kesalahan prediksi yang menyesatkan
3. Optimalisasi model regresi sangat krusial dalam menjamin presisi analisis guna menghindari kerugian finansial pada industri hiburan digital global

DESKRIPSI DATASET

Dataset : Evolution of Top Games on Twitch" dari Kaggle
Karakteristik : Data Time-Series yang mencerminkan dinamika hiburan digital global
Statistik Dataset :

1. **Jumlah Baris: 21.000**
2. **Jumlah Fitur: 12 kolom**
3. **Jumlah Game: 2.359 judul game**
4. **Target Variable: hours_watched**
5. **Rentang Waktu: Januari 2016 – September 2024**

EXPLORATORY DATA ANALYSIS (EDA) - TOP GAMES



Tahapan Exploratory Data Analysis (EDA) melalui visualisasi Top Games menunjukkan dominasi mutlak kategori "Just Chatting" sebagai pemimpin pasar, yang diikuti oleh judul-judul besar seperti "League of Legends" dan "Grand Theft Auto V". Proses ini berhasil mengubah angka statistik yang kaku menjadi narasi data yang mudah dipahami, sehingga menghubungkan hasil teknis secara langsung dengan realitas industri streaming. Melalui pendekatan ini, tim membuktikan bahwa model yang dibangun tidak hanya akurat secara matematis, tetapi juga memiliki relevansi bisnis yang tinggi dalam memetakan dinamika ekosistem hiburan digital.

PIPELINE PROCESSING

1. Pembersihan: menghapus data tanpa nama
2. Outlier : Eliminasi berbasis 2.5x IQR pada skala logaritma untuk menjaga representasi data
3. Target: menerapkan hybrid log-transformation untuk variansi dan mengurangi bias ekstrim

KEBARUAN DAN MODIFIKASI KODE

1. Arsitektur menggunakan Modular Clean Architecture yang terbagi modul preprocessing, model, dan evaluasi
2. encoding: Sin/Cos Cyclical Transform untuk fitur bulan (memahami siklus Desember–Januari)
3. Fitur Era: klasifikasi tahun (2016 – 2024) untuk menangkap perubahan struktural tren streaming

EXPERIMENT LOG

Evaluasi dilakukan menggunakan 5-Fold Cross Validation dengan Metrik pada skala asli

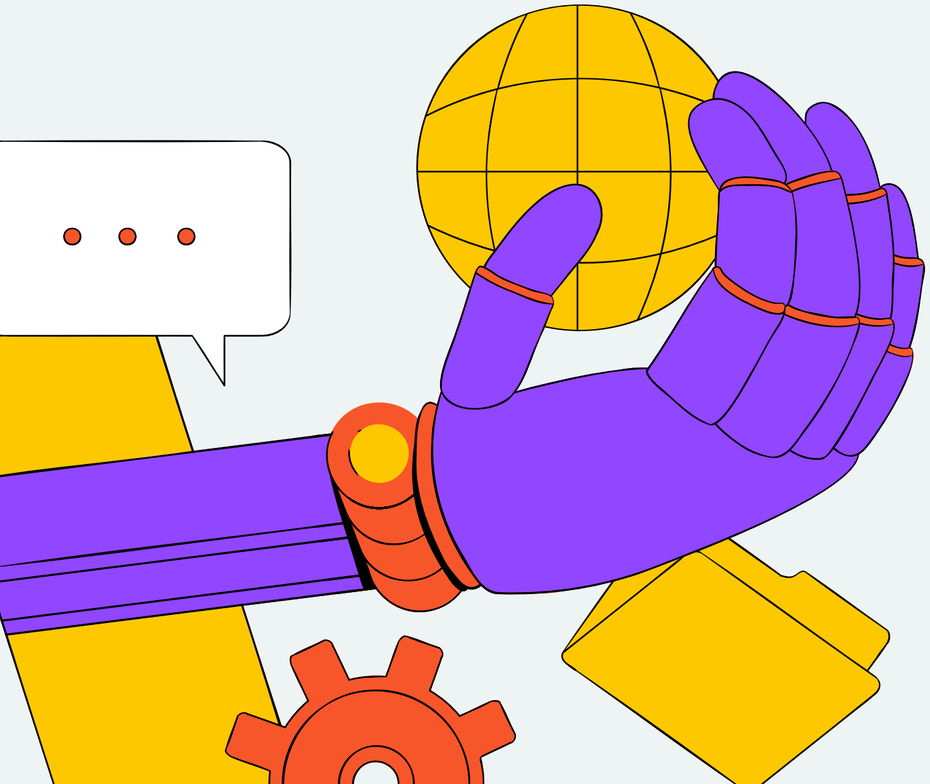
No.	Model	Fitur	Parameter Utama	MAE	RMSE	MAPE	R^2	Catatan
1	LR Baseline	18 fitur	default	3.013.157	23.121.075	33.61%	0.9	Model dasar, tanpa regularisasi
2	Ridge α=1.0	18 fitur	alpha=1.0	3.013.168	23.122.119	33.61%	0.9	Regularisasi L2 ringan
3	Ridge α=10.0	18 fitur	alpha=10.0	3.013.271	23.131.461	33.61%	0.9	Penalti lebih kuat, hasil stabil
4	Lasso α=0.01	18 fitur	alpha=0.01	3.026.438	23.806.954	33.93%	0.9	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

PERBANDINGAN BASELINE VS MODEL TERBARU

Metode : Membandingkan Linear Regression standar dengan Ridge + Polynomial Interaction

Regularisasi : Penggunaan L2 Regularization (α=50.0) pada model terbaru untuk mencegah overfitting

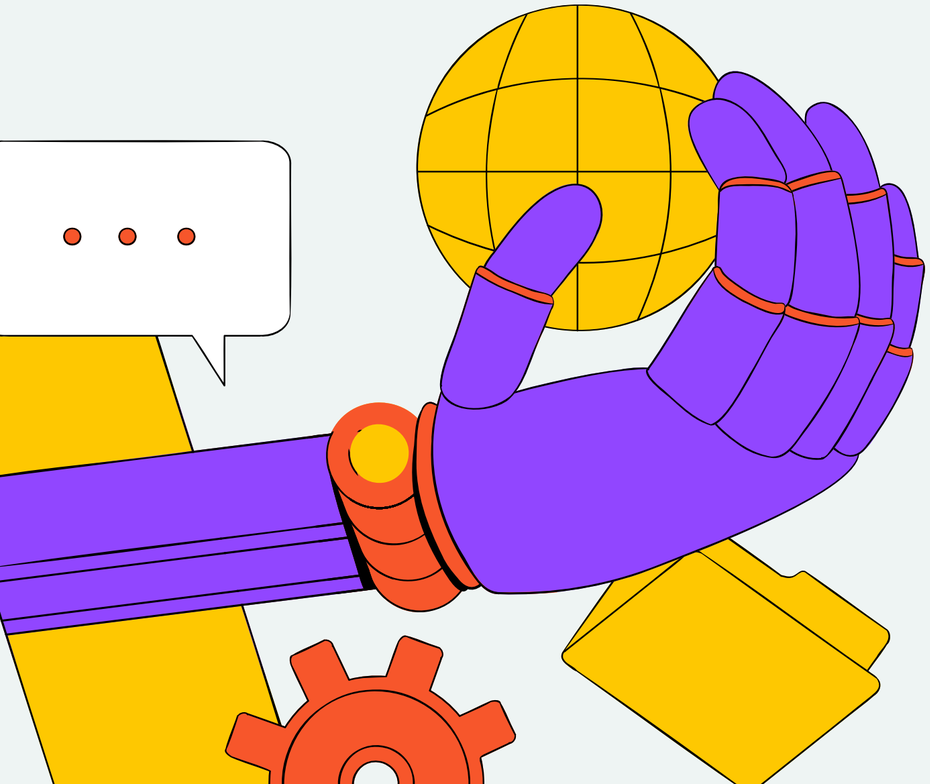
ANALISIS MENDALAM PERFORMA



Kenaikan R^2 (0.90 ke 0.93) membuktikan model terbaru lebih baik dalam menjelaskan variasi data

Penurunan MAPE (33,61% ke 31.13%) menunjukkan prediksi yang lebih presisi pada skala ekstrem di Twitch

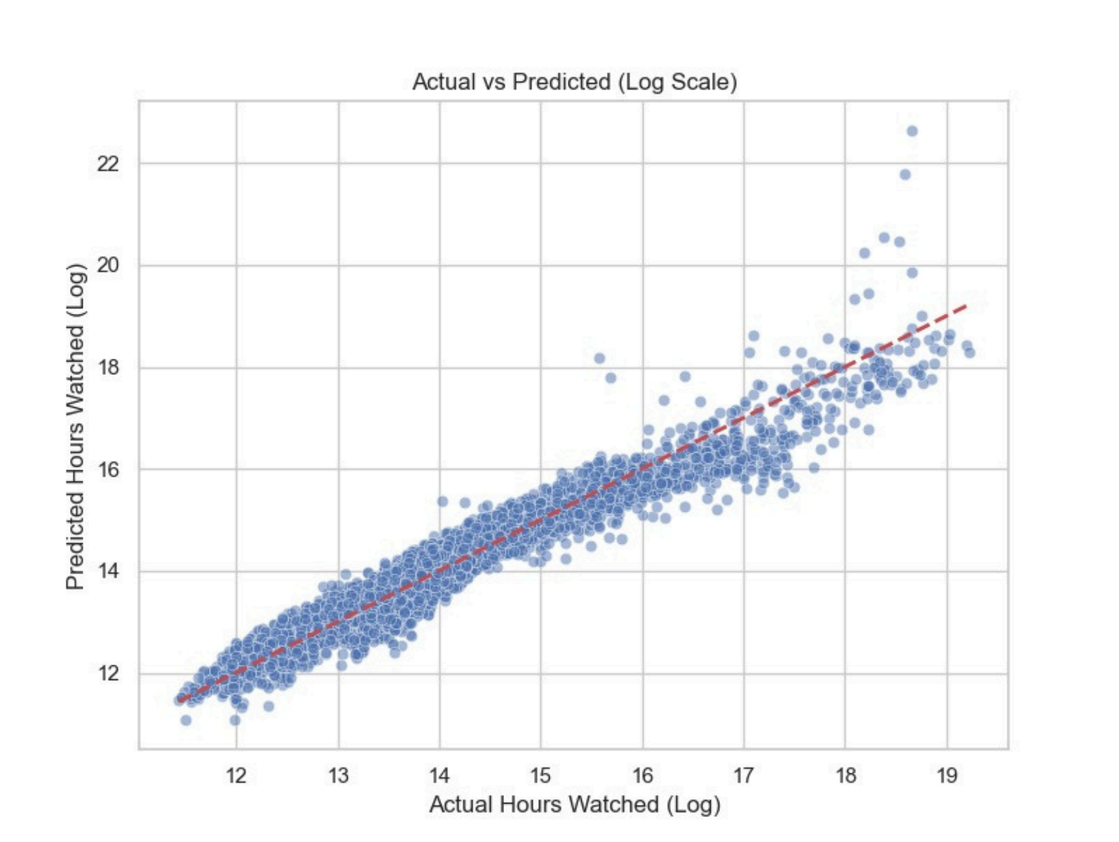
HASIL VALIDASI K-FOLD



Metodelogi 5-Fold Cross Validation untuk Estimasi generalisasi yang robust

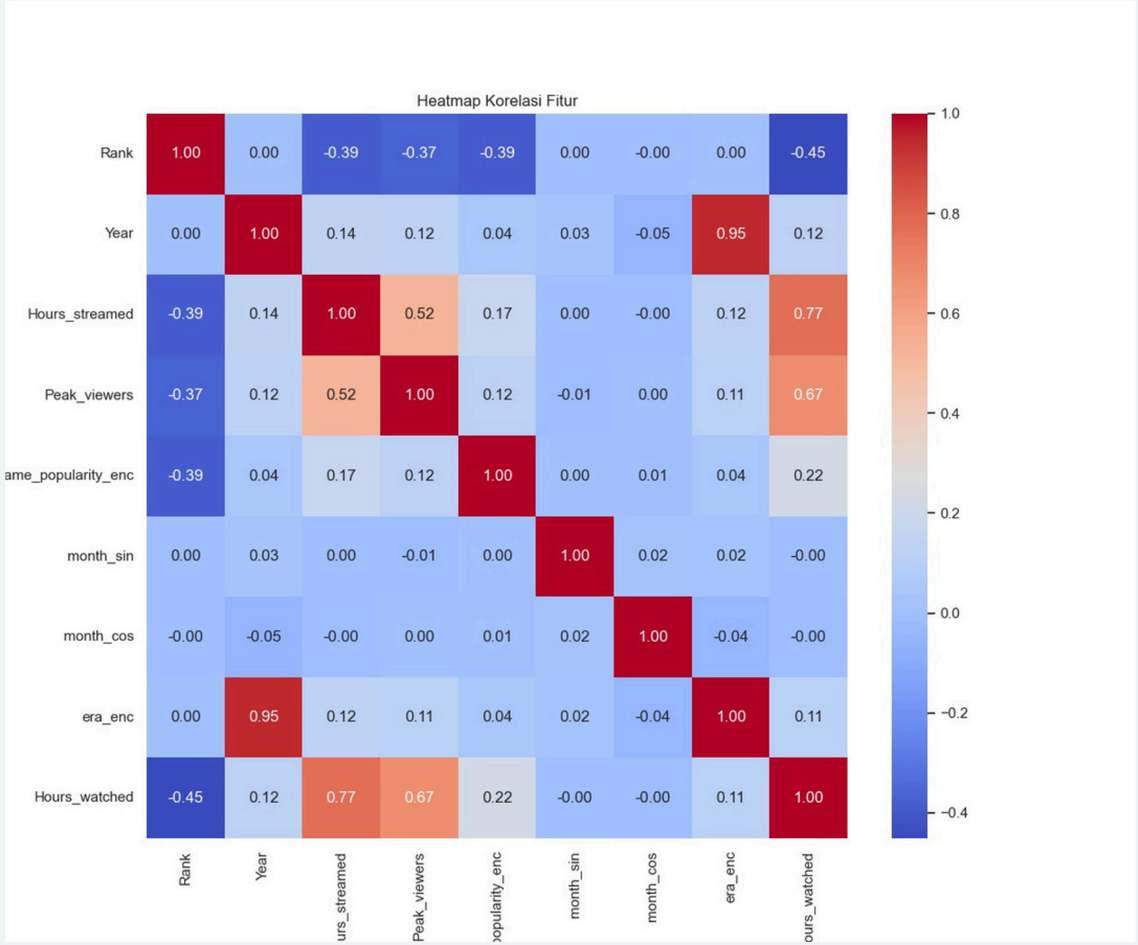
Skor: Rata-rata Mean R^2 (log-scale) mencapai 0.9006, menunjukkan stabilitas model yang tinggi

EVALUASI PREDIKSI (SCATTER PLOT)



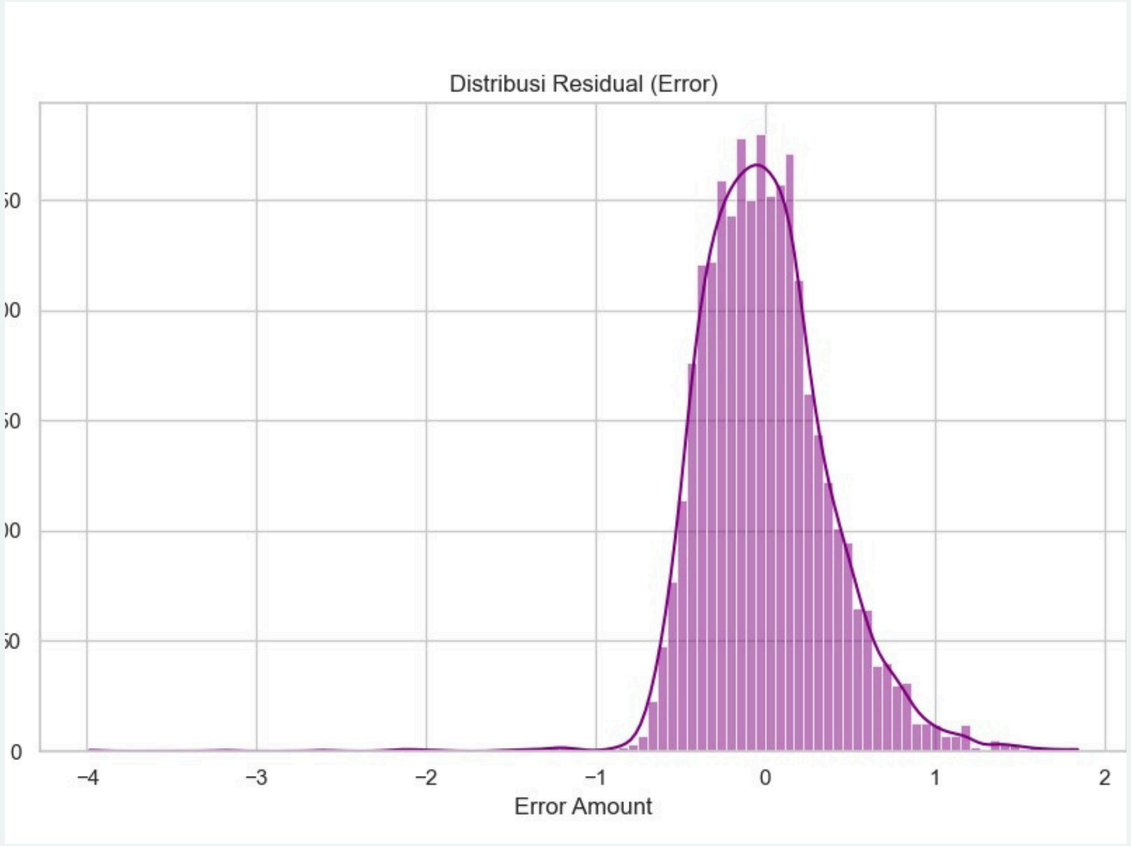
Titik data mengikuti garis diagonal merah dengan konsisten, membuktikan akurasi prediksi model pada berbagai skala jumlah tontonan.

HEATMAP KORELASI FITUR



Implementasi Heatmap Korelasi Fitur dalam proyek ini berfungsi sebagai alat diagnostik krusial yang membantu tim untuk mengidentifikasi variabel-variabel dengan pengaruh paling signifikan terhadap variabel target Hours Watched. Melalui visualisasi data tersebut, ditemukan indikator terkuat yaitu fitur Hours_streamed dengan nilai korelasi sebesar 0.77 dan fitur Peak_viewers sebesar 0.67 terhadap total jam ditonton. Fenomena ini membuktikan secara ilmiah bahwa daya tarik massa pada momen puncak merupakan penggerak utama dari popularitas sebuah judul game di platform Twitch. Adanya korelasi yang sangat kuat antar fitur inilah yang kemudian menjadi fondasi utama bagi model untuk mencapai akurasi prediksi dengan skor R^2 yang sangat memuaskan, yakni mencapai 0.93.

ANALISIS RESIDUAL (ERROR)



Error terdistribusi secara normal (lonceng) di sekitar angka 0, membuktikan model sehat secara statistik dan tidak memiliki bias sistematis.

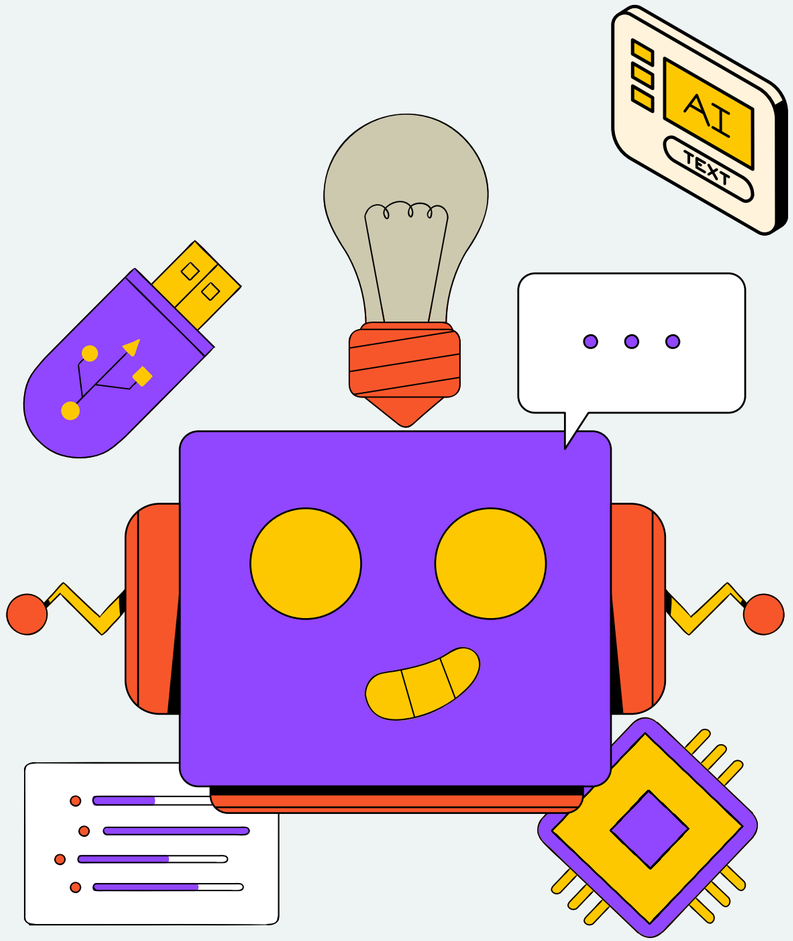
STRESS TEST

Metode injeksi Gaussian noise (σ = 0.05 hingga 0.30) pada fitur input
Hasilnya metode tetap robust dengan R^2 0.8389 bahkan saat data diberikan gangguan sebesar 30%

Tingkat Gangguan (σ)	Skor R^2	Dampak Performa
0	0.9029	Kondisi ideal (data bersih)
0.05	0.9011	Performa tetap stabil
0.1	0.8955	Penurunan sangat kecil
0.3	0.8389	Model tetap robust (akurasi di atas 80%)

KESIMPULAN

- Model Ridge + Polynomial adalah algoritma paling efektif untuk dataset Twitch.
- Teknik Hybrid Log-Transformation krusial dalam menstabilkan variansi ekstrem.
- Model mencapai akurasi R^2 0.93 dan terbukti tangguh terhadap gangguan data melalui Stress Test.



TERIMA KASIH