PERBANDINGAN REGRESI POLINOMIAL DAN REGRESI LINIER UNTUK PERAMALAN JUMLAH SUMUR GAS DAN GAS KONDENSAT DI MONTANA

Muhammad Fikri Hindami[1], Diyah Rosiani[1*], Arya Dwi Candra[1*]
[1]Teknik Produksi Minyak & Gas, Politeknik Energi dan Mineral Akamigas, Jalan Gajah Mada No. 38, Cepu, Blora, Jawa Tengah, Indonesia, 58315
**E-mail*: diyahrosiani@gmail.com

ABSTRAK

Produksi gas alam di Montana, Amerika Serikat, menunjukkan tren yang tidak konsisten, sehingga diperlukan pendekatan pemodelan yang tepat untuk mendukung perencanaan energi di masa depan. Penelitian ini bertujuan memproyeksikan jumlah sumur gas dan gas kondensat di Montana hingga tahun 2030 dengan mengevaluasi kinerja tiga model regresi, yaitu Regresi Polinomial Derajat 2 (Model A), Regresi Polinomial Derajat 3 (Model B), dan Regresi Linier yang dilatih menggunakan data 10 tahun terakhir (Model C). Kinerja ketiga model dievaluasi menggunakan tiga metrik, yakni koefisien determinasi (R^2), *Root Mean Square Error* (RMSE), dan *Mean Absolute Percentage Error* (MAPE). Hasil penelitian menunjukkan bahwa Model C merupakan model yang paling akurat dan konsisten unggul, dengan R^2 sebesar 0,927, RMSE sebesar 138,7, dan MAPE sebesar 2,09%. Sebaliknya, Model B yang lebih kompleks (R^2 = 0,895) mengalami *overfitting* sehingga menghasilkan proyeksi jumlah sumur yang tidak realistis (bernilai negatif) mulai tahun 2025, sementara Model A relatif stabil namun menghasilkan rata-rata kesalahan prediksi yang cukup besar (MAPE 13,71%). Penelitian ini menyimpulkan bahwa penggunaan data yang lebih terkini lebih efektif dibandingkan model kompleks yang dilatih dengan seluruh rentang data historis, serta menegaskan bahwa nilai R^2 yang tinggi tidak dapat dijadikan satu-satunya acuan validitas ekstrapolasi jangka panjang tanpa disertai pemeriksaan kewajaran logika domain.

Kata Kunci: Gas alam, Montana, Overfitting, Peramalan jumlah sumur, Regresi polynomial, Validasi *out-of-sample*

1. PENDAHULUAN

Gas alam merupakan bahan bakar fosil berwujud gas, dengan metana sebagai komponen utamanya, yaitu sekitar 70% hingga 90% dari total volumenya. Selain metana, gas alam mentah yang dihasilkan dari batuan reservoir bawah permukaan juga mengandung hidrokarbon lain seperti etana, propana, dan butana [1]. Di samping itu, terdapat pula senyawa non-hidrokarbon atau pengotor (*impurities*), antara lain karbon dioksida, nitrogen, dan hidrogen sulfida [2], [3]. Dalam konteks transisi energi, gas alam berperan sebagai bahan bakar penghubung [1], dan pasokan gas alam Amerika Serikat sendiri terus menjadi objek kajian [4].

Tren produksi gas alam di Montana, Amerika Serikat, secara historis memperlihatkan alur perkembangan yang dinamis dan tidak sederhana. Gejolak ini dipicu oleh interaksi antara inovasi teknologi ekstraksi, dinamika ekonomi global, serta regulasi energi di tingkat nasional [5]. Pola tersebut tercermin pada jumlah sumur gas dan gas kondensat yang naik dari 2.607 unit pada 1990 menjadi puncak 7.095 unit pada 2008, lalu turun menjadi 5.043 unit pada 2020 [5].

Peramalan jumlah sumur gas dan gas kondensat perlu dilakukan karena jumlah sumur menentukan kapasitas produksi yang tersedia di suatu wilayah dan menjadi dasar berbagai keputusan perencanaan di industri migas, antara lain program pengeboran dan pemeliharaan sumur, kebutuhan fasilitas pengumpul dan pengolahan gas, serta alokasi anggaran investasi. Proyeksi yang meleset dapat berujung pada kapasitas fasilitas yang berlebih atau justru kurang.

Proyeksi jangka panjang di sektor migas juga telah disusun untuk Indonesia hingga tahun 2045 [6]. Peramalan jumlah sumur melengkapi analisis kurva penurunan produksi (*decline curve*) [7] yang bekerja pada tingkat sumur, karena peramalan jumlah sumur bekerja pada tingkat wilayah atau lapangan.

Teknologi *machine learning* memungkinkan komputer mempelajari hubungan dalam data secara otomatis tanpa perlu menuliskan aturan secara eksplisit [8]–[10]. Dalam skema *supervised learning*, model dilatih menggunakan sampel data yang diberi label untuk menghasilkan estimasi dengan akurasi yang tinggi [8], [9]. Regresi merupakan salah satu pendekatan *supervised learning* yang paling banyak dipakai untuk peramalan dan mudah diimplementasikan dengan pustaka seperti NumPy [10].

Tiga pendekatan regresi dibandingkan dalam penelitian ini. Regresi linier memodelkan hubungan sebagai garis lurus dengan dua parameter. Kelebihannya adalah sederhana dan mudah diinterpretasikan [11], [12]; dengan parameter yang sedikit, risiko *overfitting*-nya pun rendah. Pada data yang kurvanya mulus, regresi linier bahkan dapat mengungguli metode yang lebih rumit [12]. Kekurangannya, regresi linier mengasumsikan respons yang kira-kira linier pada rentang data yang dipakai, sehingga asumsi ini lebih masuk akal pada jendela waktu yang pendek dan kurang mampu menangkap hubungan nonlinier [12].

Regresi polinomial memperluas regresi linier dengan menambahkan suku berpangkat dari variabel bebas sehingga mampu memodelkan hubungan nonlinier, sementara koefisiennya tetap diestimasi secara linier [8], [13]. Kelebihannya adalah fleksibilitas dalam mengikuti pola yang melengkung, misalnya fase naik yang kemudian turun [8], [14]. Kekurangannya, semakin tinggi derajat polinomial, semakin besar risiko *overfitting* terhadap derau data dan semakin tidak stabil ekstrapolasinya di luar rentang data [15]. Studi terbaru peramalan produksi minyak mentah Amerika Serikat dengan regresi polinomial, misalnya, melaporkan R^2 data uji yang rendah (0,307) dan MAPE sekitar 20%, serta menegaskan bahwa hasilnya bersifat ekstrapolatif sehingga belum memasukkan faktor harga dan kebijakan [14].

Perbedaan sifat kedua pendekatan ini menimbulkan pertanyaan praktis, yaitu apakah model yang lebih fleksibel dengan seluruh data historis lebih baik daripada model sederhana dengan data terkini untuk meramalkan jumlah sumur. Penelitian ini bertujuan membandingkan kinerja tiga model, yaitu regresi polinomial derajat dua (Model A), regresi polinomial derajat tiga (Model B), dan regresi linier berbasis data sepuluh tahun terakhir (Model C), untuk meramalkan jumlah sumur gas dan gas kondensat di Montana hingga tahun 2030. Tren historis 1990–2020 dianalisis, kemudian kinerja model dievaluasi menggunakan R^2, RMSE, dan MAPE secara *in-sample* maupun *out-of-sample*. Data Montana digunakan sebagai studi kasus karena tersedia secara terbuka dengan rentang waktu yang panjang [5], pendekatan yang sama dapat diterapkan pada data sumur lapangan migas lain.

2. METODE

Penelitian ini bersifat kuantitatif dengan studi kasus jumlah sumur gas dan gas kondensat di Montana. Tahapan penelitian terdiri atas pengumpulan data, pembentukan tiga model regresi, evaluasi *in-sample*, proyeksi tahun 2021–2030, dan validasi *out-of-sample*.

2.1 Data Penelitian

Variabel terikat (y) adalah jumlah sumur gas dan gas kondensat Montana per tahun pada periode 1990–2020 (31 observasi) yang diperoleh dari *U.S. Energy Information Administration* (EIA) [5]. Data lengkap ditampilkan pada Tabel 1. Data disimpan dalam berkas Excel dan dibaca menggunakan pustaka *openpyxl* [16]. Variabel bebas (x) adalah tahun yang digeser menjadi x = tahun − 1990, sehingga $x = 0$ untuk 1990 dan $x = 30$ untuk 2020; pergeseran ini hanya menyederhanakan koefisien dan tidak mengubah kualitas pencocokan.

Tabel 1. Jumlah sumur gas dan gas kondensat Montana, 1990–2020 [5]

Tahun	Jumlah sumur	Tahun	Jumlah sumur	Tahun	Jumlah sumur
1990	2.607	2001	4.331	2011	6.615
1991	2.802	2002	4.544	2012	6.366
1992	2.890	2003	4.539	2013	5.870
1993	3.075	2004	4.971	2014	5.682
1994	2.940	2005	5.751	2015	5.651
1995	2.918	2006	6.578	2016	5.528
1996	2.990	2007	6.925	2017	5.440
1997	3.071	2008	7.095	2018	5.133
1998	3.423	2009	7.031	2019	4.976
1999	3.634	2010	6.059	2020	5.043
2000	3.321				

2.2 Model Regresi dan Pemilihan Derajat Polinomial

Model regresi polinomial berderajat n dirumuskan sebagai berikut [8], [13]:

$$\hat{y} = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_n x^n \tag{1}$$

dengan $\hat{y}$ = prediksi jumlah sumur, x = tahun − 1990, β_0, …, β_n = koefisien regresi, dan n = derajat polinomial. Regresi linier merupakan kasus $n = 1$. Koefisien diestimasi dengan metode kuadrat terkecil (*ordinary least squares*), yaitu meminimalkan jumlah kuadrat selisih antara data aktual dan prediksi:

$$S = \Sigma_{i=1}^{n} (y_i - \hat{y}_i)^2 \;\rightarrow\; \text{minimum} \tag{2}$$

Solusi kuadrat terkecil dinyatakan dalam bentuk matriks pada Persamaan (3), dengan X adalah matriks desain yang kolomnya berisi $1, x, x^2, \ldots, x^n$. Perhitungan dilakukan menggunakan fungsi *numpy.polyfit*(x, y, derajat) [10].

$$\hat{\beta} = (X^T X)^{-1} X^T y \tag{3}$$

Tiga model dibentuk, yaitu Model A (polinomial derajat 2) dan Model B (polinomial derajat 3) dilatih menggunakan seluruh data 1990–2020 (31 observasi), sedangkan Model C (regresi linier, derajat 1) dilatih hanya dengan data sepuluh tahun terakhir, yaitu 2011–2020 (10 observasi). Derajat 2 dipilih karena merupakan polinomial paling sederhana yang mampu menangkap satu titik balik, sesuai pola data yang naik hingga puncak pada 2008 lalu turun. Derajat 3 dipilih karena data juga memiliki fase awal yang landai (1990–1997), lonjakan tajam (2001–2008), dan penurunan sesudahnya, yaitu perubahan kelengkungan yang hanya dapat ditangkap oleh kurva berbentuk S. Derajat yang lebih tinggi tidak digunakan karena dengan 31 observasi setiap tambahan parameter meningkatkan risiko *overfitting* dan ketidakstabilan ekstrapolasi. Sebagai acuan kasar, kajian ekstrapolasi polinomial kuadrat terkecil menyarankan derajat tidak melebihi sekitar $\frac{1}{2}\sqrt{N}$ [15], atau sekitar 2,8 untuk $N = 31$, sehingga derajat 2 dan 3 dipandang sebagai batas atas yang wajar. Untuk Model C, setelah 2011 (6.615 sumur), jumlah sumur menurun bertahap hingga 2020 sehingga pola pada jendela ini mendekati linier, dan asumsi kelinieran lebih sesuai pada jendela waktu yang pendek [12]. Panjang jendela sepuluh tahun (satu dekade terakhir) dipilih secara pragmatis dan tidak dioptimalkan secara formal.

2.3 Metrik Evaluasi

Kesesuaian (*goodness of fit*) setiap model dievaluasi menggunakan koefisien determinasi:

$$R^2 = 1 - SS_{\text{res}} / SS_{\text{tot}} \tag{4}$$

$$SS_{\text{res}} = \Sigma_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{5}$$

$$SS_{\text{tot}} = \Sigma_{i=1}^{n} (y_i - \bar{y})^2 \tag{6}$$

Dengan SS_{res} = jumlah kuadrat residual dan SS_{tot} = jumlah kuadrat total terhadap rata-rata ($\bar{y}$). Pada data latih, nilai R^2 berkisar 0–1; semakin mendekati 1, semakin baik model menjelaskan data. Pada data uji, R^2 dapat bernilai negatif apabila prediksi model lebih buruk daripada sekadar menggunakan nilai rata-rata.

$$\text{RMSE} = \sqrt{[(1/n) \Sigma (y_i - \hat{y}_i)^2]} \tag{7}$$

RMSE digunakan untuk mengukur rata-rata besaran kesalahan prediksi dalam satuan yang sama dengan data asli (jumlah sumur). Dibandingkan dengan MAE, RMSE lebih peka terhadap *outlier* karena residual dikuadratkan terlebih dahulu dalam proses perhitungannya [8], [9].

$$\text{MAPE} = (100\%/n) \Sigma |(y_i - \hat{y}_i) / y_i| \tag{8}$$

MAPE menunjukkan rata-rata kesalahan prediksi dalam bentuk persentase terhadap nilai aktualnya. Dengan bentuk persentase ini, akurasi model lebih mudah diinterpretasikan secara relatif dan dapat diperbandingkan meskipun skala datanya berbeda [8], [17].

$$\text{MAE} = (1/n) \Sigma |y_i - \hat{y}_i| \tag{9}$$

MAE (*Mean Absolute Error*) mengukur rata-rata besaran mutlak kesalahan prediksi dalam satuan yang sama dengan data asli. Berbeda dengan RMSE, MAE tidak mengkuadratkan residual sehingga kurang sensitif terhadap *outlier*; MAE digunakan sebagai metrik tambahan pada validasi *out-of-sample* untuk melengkapi interpretasi RMSE.

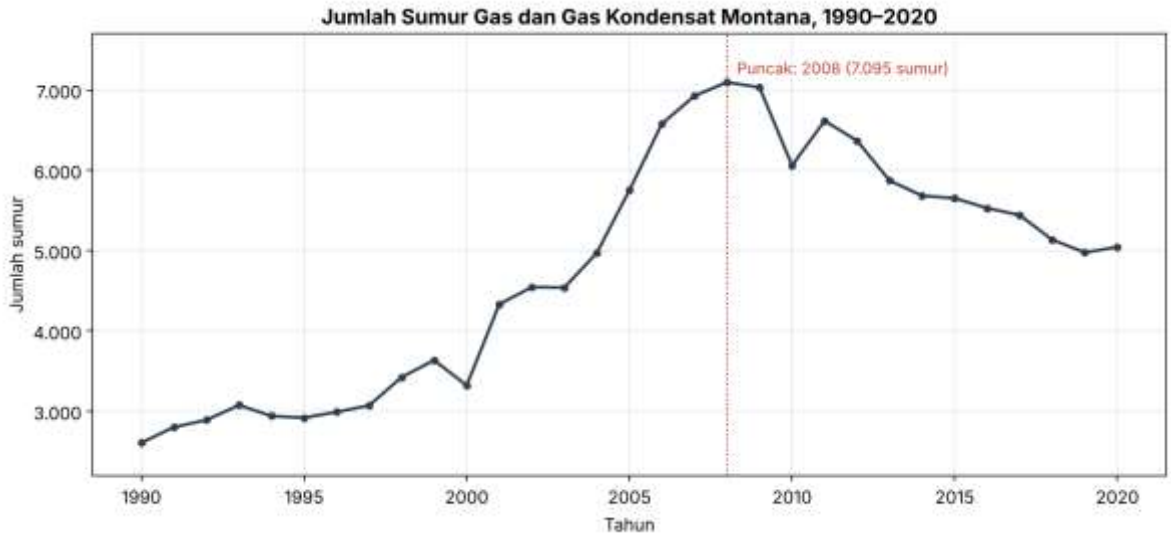
2.4 Skema Validasi dan Proyeksi

Evaluasi dilakukan dalam dua skema. Pertama, evaluasi *in-sample* menghitung metrik pada data yang sama dengan data latih. Kedua, validasi *out-of-sample* melatih ulang ketiga model dengan data 1990–2015 (Model C memakai 2006–2015) dan mengujinya pada data 2016–2020. Proyeksi tahun 2021–2030 dihasilkan oleh model yang dilatih ulang dengan seluruh data yang tersedia (Model C: 2011–2020). Seluruh perhitungan dijalankan dengan bahasa pemrograman Python.

3. HASIL DAN PEMBAHASAN

3.1 Pola fluktuasi dan perkembangan sumur dalam rentang 1990 hingga 2020

Gambar 1 menampilkan deret waktu tahunan jumlah sumur gas dan gas kondensat di Montana. Jumlah sumur berkembang dari 2.607 unit pada 1990 menjadi puncak 7.095 unit pada 2008, diikuti oleh tren penurunan secara keseluruhan hingga 4.976 unit pada 2019, sebelum naik kecil kembali menjadi 5.043 unit pada 2020 [5]. Dari puncak 2008 hingga 2020, jumlah sumur berkurang sekitar 29%.



Gambar 1. Jumlah sumur gas dan gas kondensat Montana, 1990–2020 [5]

3.2 Hasil Pembentukan Model Regresi

Pembentukan model dilakukan melalui lima langkah: (1) variabel bebas dibentuk dengan x = tahun − 1990; (2) matriks desain X disusun sesuai derajat polinomial setiap model; (3) koefisien β diestimasi dengan metode kuadrat terkecil menurut Persamaan (2) dan (3); (4) nilai pencocokan (*fitted*) $\hat{y}$ dihitung dari persamaan yang diperoleh; dan (5) residual $e = y - \hat{y}$ dihitung, kemudian R^2, RMSE, dan MAPE ditentukan menurut Persamaan (4)–(8). Persamaan regresi hasil estimasi disajikan pada Tabel 2.

Tabel 2. Persamaan regresi hasil estimasi (x = tahun − 1990)

Model	Data latih	Persamaan regresi
A (derajat 2)	1990–2020 (n = 31)	$\hat{y} = -8{,}9296x^2 + 390{,}45x + 1.634{,}56$
B (derajat 3)	1990–2020 (n = 31)	$\hat{y} = -0{,}8958x^3 + 31{,}3807x^2 - 85{,}3955x + 2.725{,}62$
C (linier)	2011–2020 (n = 10)	$\hat{y} = -172{,}19x + 10.021{,}35$

Pola ketiga model terbentuk dari tanda dan besarnya koefisien. Pada Model A, koefisien x^2 yang negatif (−8,9296) menghasilkan parabola terbuka ke bawah dengan puncak pada x = 390,4454 / (2 × 8,9296) ≈ 21,9, yaitu sekitar tahun 2012, sedangkan puncak aktual terjadi pada 2008; selisih ini menjelaskan residual positif yang besar pada 2006–2009 (Gambar 6). Pada Model B, suku kubik yang negatif ($-0{,}8958x^3$) membentuk kurva berbentuk S yang mengikuti kenaikan 2001–2008 dengan lebih baik, tetapi dominasi suku kubik pada ujung periode membuat kurva menurun tajam menjelang 2020 (nilai pencocokan 4.220 sumur terhadap 5.043 sumur aktual). Pada Model C, kemiringan −172,19 menunjukkan bahwa jumlah sumur berkurang rata-rata sekitar 172 unit per tahun selama 2011–2020.

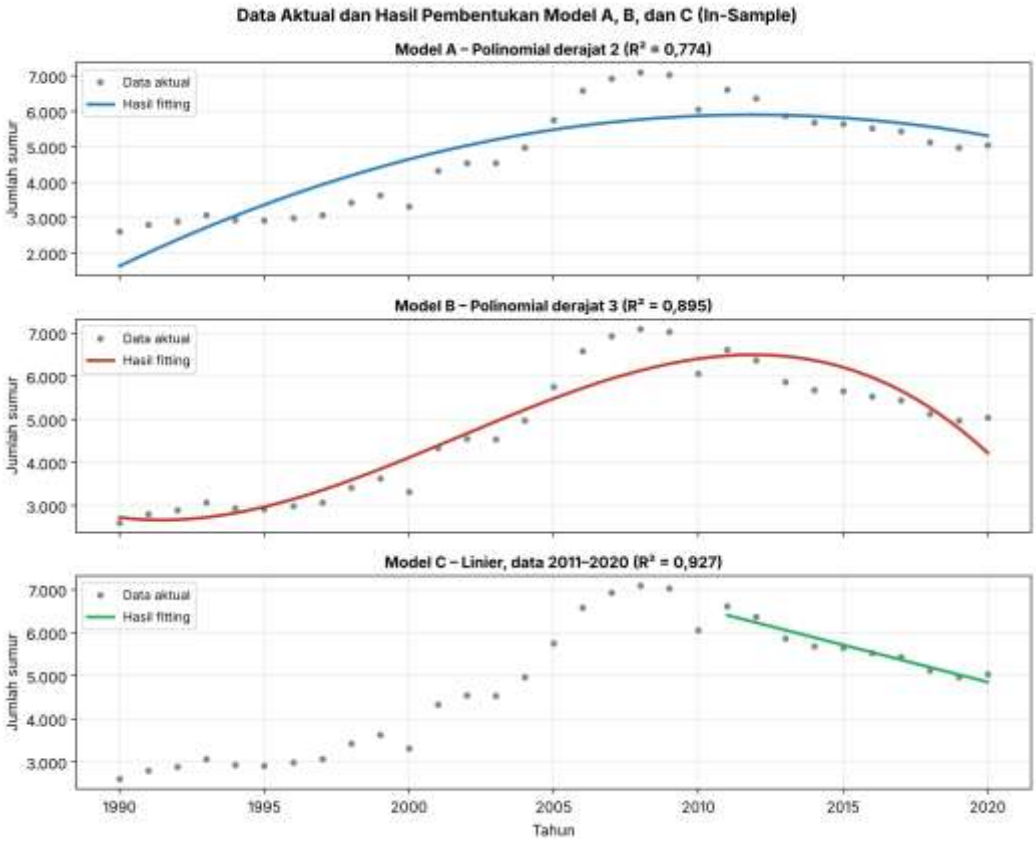
Cara perhitungan metrik diilustrasikan pada Tabel 3 untuk Model C. Untuk setiap tahun, residual $e = y - \hat{y}$ dihitung, lalu dikuadratkan dan dinyatakan sebagai persentase terhadap nilai aktual. Dengan n = 10 diperoleh $SS_{res} = \Sigma e^2$ = 192.275 dan $SS_{tot} = \Sigma(y - \bar{y})^2$ = 2.638.462 dengan $\bar{y}$ = 5.630,4. Selanjutnya, R^2 = 1 − 192.275 / 2.638.462 = 0,9271; RMSE = √(192.275 / 10) = 138,7 sumur; MAPE = (100% / 10) × $\Sigma|e|/y$ = 2,09%; dan MAE = 119,6 sumur. Untuk Model A dan Model B, perhitungan yang sama dilakukan terhadap 31 observasi dengan SS_{tot} = 64.430.586, menghasilkan SS_{res} masing-masing sebesar 14.563.844 dan 6.793.661, sehingga R^2 = 0,7740 (Model A) dan 0,8946 (Model B).

Grafik hasil pembentukan ketiga model ditampilkan pada Gambar 2. Model A tampak terlalu landai sehingga menyimpang dari data pada kedua ujung periode; Model B mengikuti

pola dengan lebih baik pada 1990–2005 tetapi menyimpang pada 2013–2020; sedangkan Model C mengikuti data 2011–2020 dengan penyimpangan yang kecil.

Tabel 3. Rincian perhitungan R^2, RMSE, dan MAPE untuk Model C (data 2011–2020)

Tahun	x	y (aktual)	$\hat{y}$ (fitted)	$e = y - \hat{y}$	e^2	$\lvert e \rvert / y$ (%)
2011	21	6.615	6.405,3	209,7	43.986	3,17
2012	22	6.366	6.233,1	132,9	17.668	2,09
2013	23	5.870	6.060,9	−190,9	36.437	3,25
2014	24	5.682	5.888,7	−206,7	42.721	3,64
2015	25	5.651	5.716,5	−65,5	4.290	1,16
2016	26	5.528	5.544,3	−16,3	266	0,29
2017	27	5.440	5.372,1	67,9	4.609	1,25
2018	28	5.133	5.199,9	−66,9	4.478	1,30
2019	29	4.976	5.027,7	−51,7	2.675	1,04
2020	30	5.043	4.855,5	187,5	35.146	3,72
Jumlah / rata-rata				**0,0**	**192.275**	**2,09**



Gambar 2. Data aktual dan hasil pembentukan Model A, Model B, dan Model C (*in-sample*)

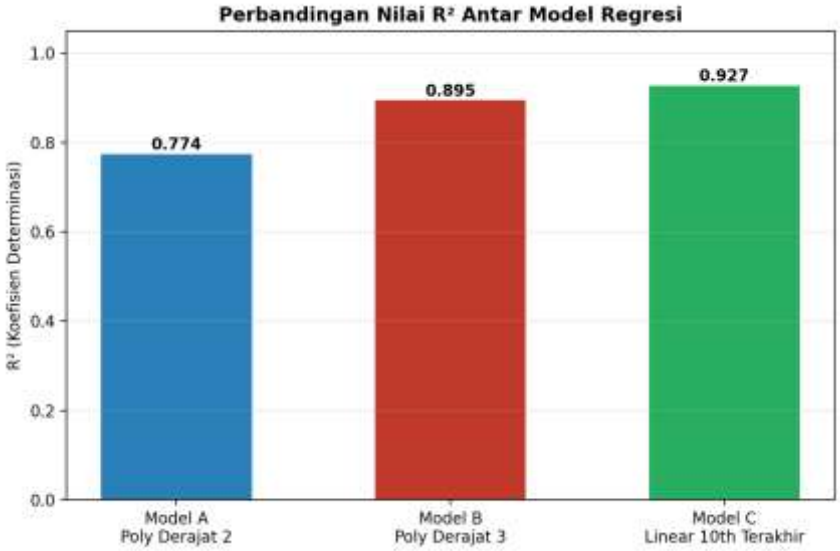
3.3 Evaluasi dan Perbandingan Kinerja Model: R^2, RMSE, dan MAPE

Berdasarkan perhitungan pada subbab sebelumnya, nilai R^2 yang diperoleh adalah 0,7740 (Model A), 0,8946 (Model B), dan 0,9271 (Model C). Hasil evaluasi lengkap in-sample ketiga model disajikan pada Tabel 4.

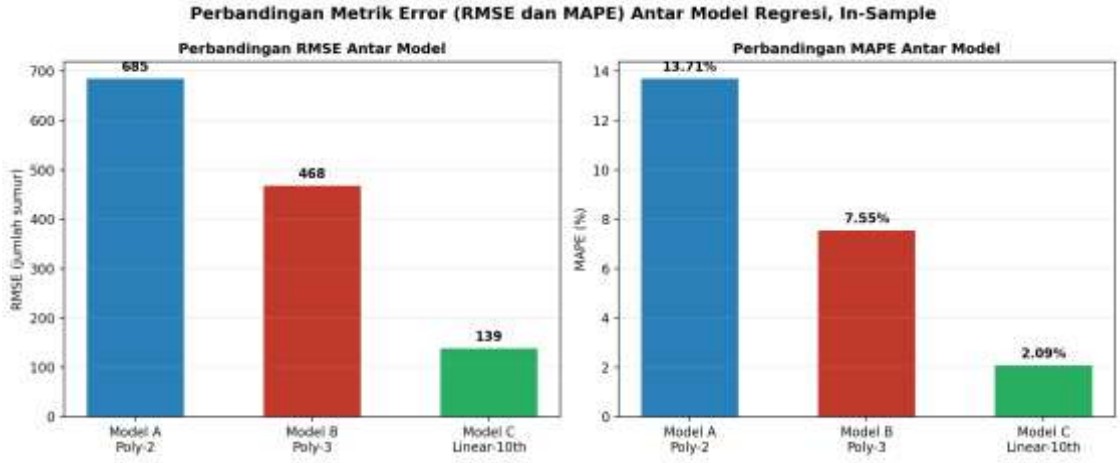
Tabel 4. Hasil evaluasi in-sample model: R², RMSE, dan MAPE

Model	R²	RMSE (Sumur)	MAPE (%)
A (Poly-2)	0,774	685,4	13,71
B (Poly-3)	0,895	468,1	7,55
C (Linear)	0,927	138,7	2,09

Konsistensi arah kesimpulan tampak dari ketiga metrik yang digunakan: Model C secara sistematis mengungguli kedua model lainnya di seluruh metrik (R² paling tinggi, RMSE dan MAPE paling rendah), sementara Model A justru menunjukkan performa paling lemah. Hal yang patut dicermati, MAPE Model A mencapai 13,71%, yang berarti rata-rata prediksinya meleset lebih dari 13% dari nilai aktual, tingkat kesalahan yang cukup besar bila dipakai untuk kebutuhan perencanaan operasional. Bandingkan dengan Model C yang MAPE-nya hanya 2,09%, mencerminkan presisi yang sangat tinggi, padahal model ini dilatih dengan jumlah data yang jauh lebih sedikit (10 observasi, dibandingkan dengan 31 observasi pada Model A). Perbandingan metrik tersebut divisualisasikan pada Gambar 3 dan Gambar 4.

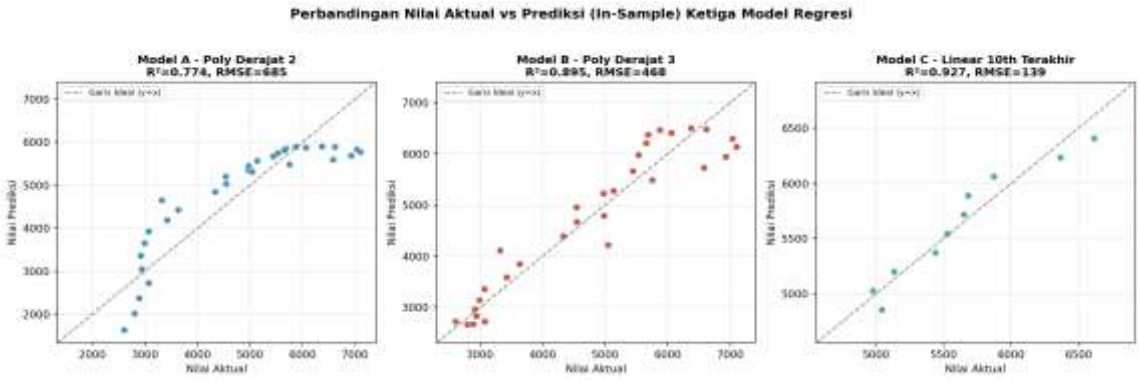


Gambar 3. Perbandingan nilai *R²* antar model regresi



Gambar 4. Perbandingan RMSE dan MAPE antar model regresi

Visualisasi *goodness-of-fit* melalui *scatter plot* antara nilai aktual dan prediksi dari ketiga model ditampilkan pada Gambar 5. Tingkat akurasi tinggi dari Model C yang sebelumnya terbukti secara numerik kembali diperkuat oleh persebaran titiknya yang paling presisi di sepanjang garis $y = x$. Di sisi lain, Model A memperlihatkan deviasi yang cukup signifikan dari garis ideal tersebut, terutama pada rentang 3.000 hingga 5.000 sumur yang bertepatan dengan fase peralihan tren historis dari pertumbuhan ke penurunan.



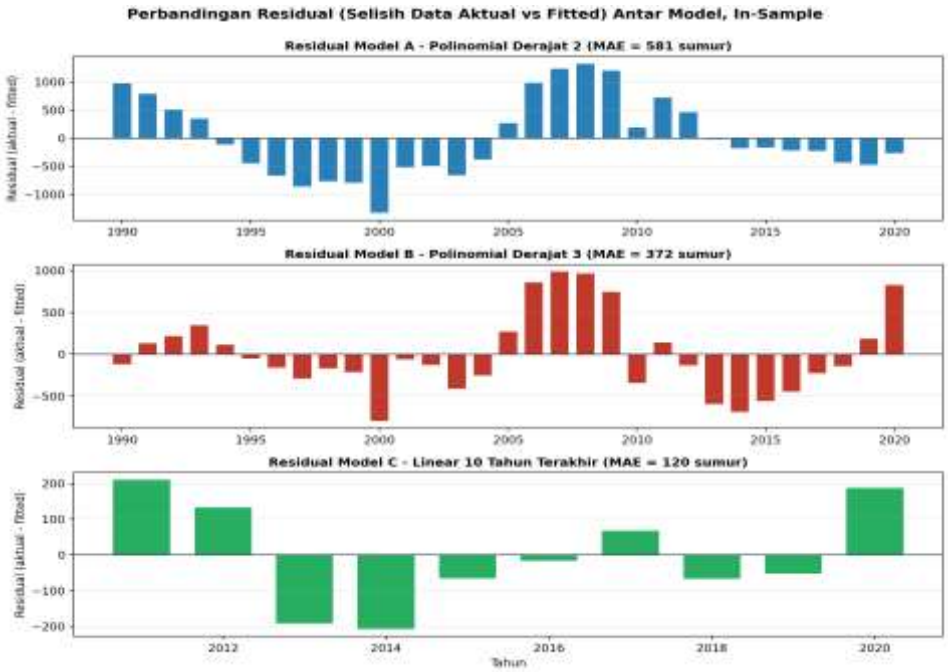
Gambar 5. *Scatter plot* nilai aktual vs prediksi (*in-sample*) ketiga model regresi terhadap garis ideal $y = x$

3.4 Analisis Residual Model

Meskipun ketiga model dikategorikan unbiased karena rata-rata residualnya mendekati nol, variabilitas error yang dihasilkan menunjukkan perbedaan yang tajam (Tabel 5 dan Gambar 6). Model A mencatat tingkat fluktuasi tertinggi dengan standar deviasi residual mencapai 685,4 sumur (hampir lima kali lipat dibandingkan Model C sebesar 138,7 sumur) serta rentang penyimpangan yang terlebar. Adanya pola non-acak yang sistematis pada residual Model A dan B mengindikasikan keterbatasan keduanya dalam memodelkan dinamika nonlinier [18]. Sebaliknya, residual Model C tersebar di sekitar nilai nol tanpa pola yang jelas, yang menunjukkan bahwa garis linier sudah cukup merepresentasikan data 2011–2020.

Tabel 5. Hasil analisis residual model

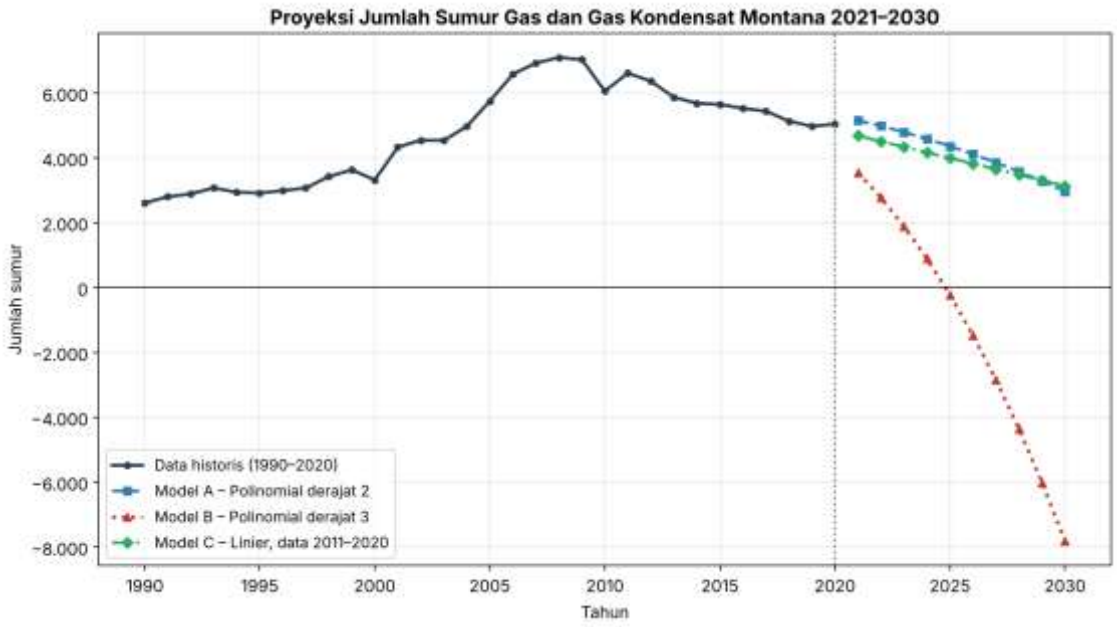
Model	Mean Residual	Std. Dev.	Residual Min	Residual Max
Model A (Poly-2)	0,0	685,4	−1.325,1	1.325,6
Model B (Poly-3)	0,0	468,1	−793,0	983,1
Model C (Linear-10th)	0,0	138,7	−206,7	209,7



Gambar 6. Perbandingan residual (selisih data aktual dan *fitted*) antar model, evaluasi *in-sample*

3.5 Perbandingan Performa Model Regresi dan Kecenderungan *Overfitting*

Proyeksi tahun 2021–2030 pada Gambar 7 menunjukkan bahwa Model C, regresi linier periode 2011 hingga 2020, unggul dengan nilai R^2 mencapai 0,927. Nilai ini lebih tinggi dibandingkan Model B, polinomial derajat tiga, yang mendapatkan 0,895, dan Model A, polinomial derajat dua, yang hanya 0,774. Meskipun Model B awalnya terlihat menjanjikan dengan R^2 yang tinggi, proyeksinya hingga 2030 menunjukkan gejala *overfitting* yang serius. Hasil prediksinya menjadi negatif sejak 2025, dan kondisi ini tidak masuk akal dalam konteks jumlah sumur. Di sisi lain, Model A dan Model C memberikan proyeksi yang jauh lebih stabil dan sesuai dengan logika domain. Keunggulan Model C terlihat dari capaian R^2 tertinggi serta tren ekstrapolasinya yang landai dan sejalan dengan pola penurunan bertahap selama sepuluh tahun terakhir [6], [13].



Gambar 7. Perbandingan hasil proyeksi tiga model regresi terhadap data historis, 1990–2030

3.6 Validasi Out-of-Sample (Data Uji 2016–2020)

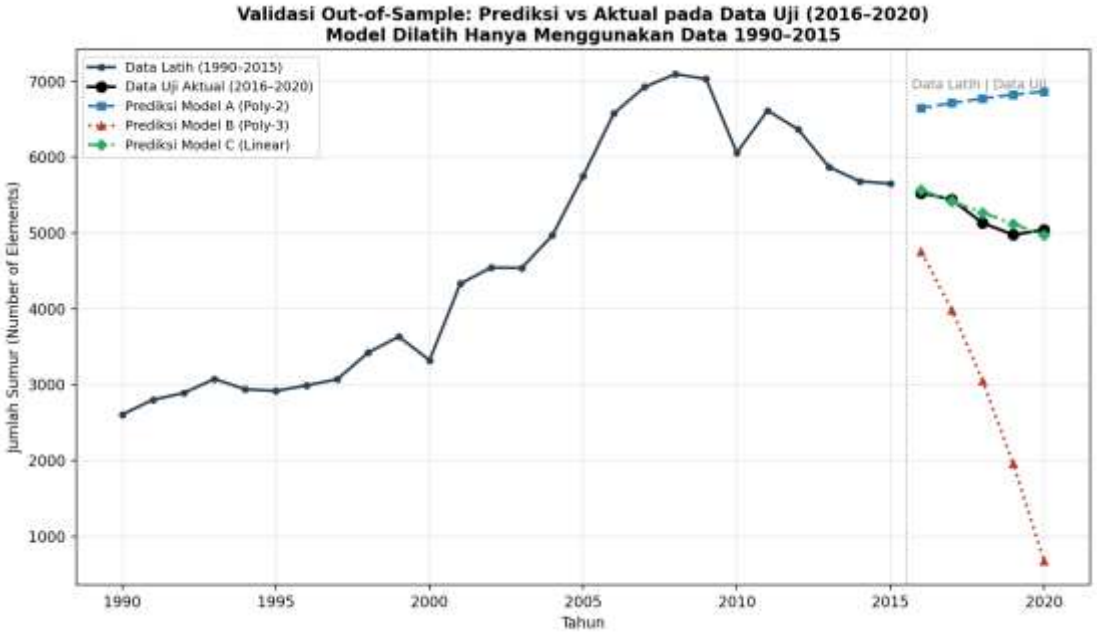
Evaluasi pada bagian sebelumnya bersifat *in-sample*, yaitu model diuji pada data yang sama dengan data yang digunakan untuk melatihnya, sehingga akurasi tinggi yang ditunjukkan Model C berpotensi hanya mencerminkan kesesuaian model terhadap data yang telah dilihatnya, bukan kemampuan generalisasi yang sesungguhnya. Untuk menguji hal ini secara lebih ketat, dilakukan validasi *out-of-sample* dengan menyisihkan data lima tahun terakhir (2016–2020) sebagai data uji yang sama sekali tidak dilibatkan dalam proses pelatihan model. Ketiga model dilatih ulang hanya menggunakan data periode 1990–2015 (26 observasi), dengan Model C menggunakan jendela sepuluh tahun terakhir dari data latih tersebut (2006–2015).

Tabel 6. Hasil evaluasi *out-of-sample* model terhadap data uji 2016–2020 (model dilatih hanya dengan data 1990–2015)

Model	R^2	RMSE (Sumur)	MAPE (%)	MAE (Sumur)
A (Poly-2)	−49,95	1569,1	29,79	1541,2
B (Poly-3)	−144,63	2652,8	45,67	2337,2
C (Linear)	0,808	96,2	1,62	82,6

Tabel 6 menyajikan hasil evaluasi ketiga model terhadap data uji 2016–2020 yang belum pernah dilihat model selama pelatihan. Hasil ini berbeda drastis dari evaluasi in-sample: Model A dan Model B justru mencatat nilai R^2 negatif (masing-masing −49,95 dan −144,63), yang menunjukkan bahwa performa prediksi kedua model tersebut jauh lebih buruk dibandingkan sekadar menggunakan nilai rata-rata data uji. Secara visual (Gambar 8), Model A terus memprediksi kenaikan jumlah sumur padahal data aktual justru menurun, sedangkan Model B mengalami keruntuhan prediksi yang jauh lebih ekstrem hingga mendekati nol pada tahun 2020. Sebaliknya, Model C tetap mempertahankan performa yang baik pada data uji (R^2 = 0,808, RMSE = 96,2, MAPE = 1,62%, MAE = 82,6). RMSE dan MAPE-nya bahkan lebih rendah daripada hasil in-sample, sedangkan R^2 lebih rendah (0,808 dibanding 0,927) karena data uji hanya terdiri atas lima titik dengan sebaran yang sempit sehingga R^2 sangat sensitif. Temuan ini mengindikasikan bahwa keunggulan Model C bukan sekadar artefak *overfitting* terhadap

data pelatihannya sendiri, melainkan menunjukkan kemampuan generalisasi pada data yang belum pernah dilihat.



Gambar 8. Validasi *out-of-sample*: prediksi vs aktual pada data uji 2016–2020, model dilatih hanya menggunakan data 1990–2015

Hasil validasi *out-of-sample* ini memperkuat kesimpulan bahwa Model C merupakan pendekatan yang paling andal untuk digunakan. Model final untuk proyeksi 2021–2030 (Gambar 7) selanjutnya dilatih ulang menggunakan keseluruhan data historis, mengikuti praktik umum dalam pemodelan prediktif: validasi kemampuan generalisasi dilakukan lebih dahulu melalui skema data latih-uji, kemudian model terpilih dilatih ulang dengan seluruh data yang tersedia untuk menghasilkan proyeksi akhir.

3.7 Implikasi bagi Perencanaan Industri Migas

Hasil proyeksi memiliki implikasi langsung bagi perencanaan di industri migas. Model C memproyeksikan jumlah sumur gas dan gas kondensat Montana turun dari 5.043 unit pada 2020 menjadi sekitar 3.134 unit pada 2030, atau berkurang sekitar 38% dengan laju rata-rata sekitar 172 sumur per tahun. Apabila penurunan ini tidak diimbangi dengan pengeboran sumur baru atau reaktivasi sumur, kapasitas produksi wilayah akan menyusut sehingga kebutuhan fasilitas pengumpul dan pengolahan gas perlu ditinjau ulang. Sebaliknya, proyeksi Model B yang bernilai negatif sejak 2025 tidak dapat dipakai untuk perencanaan karena melanggar batas fisik jumlah sumur. Perlu dicatat bahwa jumlah sumur belum mencerminkan produktivitas setiap sumur; karena itu peramalan ini sebaiknya dikombinasikan dengan analisis kurva penurunan produksi per sumur [7] serta faktor eksternal seperti harga dan kebijakan [14]. Metode yang sama dapat diterapkan pada data sumur lapangan migas di Indonesia, misalnya untuk mendukung proyeksi produksi migas nasional [6].

4. SIMPULAN

Dari tiga model yang dibandingkan, regresi linier yang dilatih dengan data sepuluh tahun terakhir (Model C) paling andal dalam meramalkan jumlah sumur gas dan gas kondensat Montana, dengan $R^2 = 0{,}927$, RMSE = 138,7 sumur, dan MAPE = 2,09% pada data latih serta $R^2 = 0{,}808$ dan MAPE = 1,62% pada data uji 2016–2020. Sebaliknya, polinomial derajat dua dan

tiga yang dilatih dengan seluruh data historis menghasilkan R^2 negatif pada data uji, dan polinomial derajat tiga memproyeksikan jumlah sumur bernilai negatif sejak 2025. Model sederhana berbasis data terkini dengan demikian lebih layak untuk ekstrapolasi jangka menengah, dan nilai R^2 yang tinggi pada data latih harus selalu disertai validasi *out-of-sample* serta pemeriksaan kewajaran fisik. Model C memproyeksikan sekitar 3.134 sumur pada 2030; penelitian lanjutan disarankan untuk memasukkan faktor harga dan kebijakan serta menggunakan data sumur lapangan migas Indonesia.

5. DAFTAR PUSTAKA

[1] D. T. Da S. Barreto and P. Khinanta, "Peran Gas Alam Sebagai Bahan Bakar Penghubung dalam Transisi Energi," *CITIZEN: Jurnal Ilmiah Multidisiplin Indonesia*, vol. 5, no. 4, pp. 959–964, 2025.

[2] M. Ismail, "Proses Pemurnian Gas Bumi Sebagai Bahan Baku Kilang LNG Mini (Studi kasus Ladang Marginal Cikarang-Area Operasi Barat EP Pertamina)," *M.P.I.*, vol. 8, no. 1, pp. 25–38, Apr. 2014.

[3] A. Nurjani and Yanto, "Pemodelan dan Simulasi Proses Flash Separation Condensate untuk Optimasi RVP Produk CPF Lapangan-X," *Jurnal Praktik Keinsinyuran*, vol. 2, no. 4, pp. 461–466, Sep. 2025.

[4] Z. Zhang dkk., "Carbon Intensity of United States Natural Gas Supply," *Environmental Science & Technology*, vol. 60, no. 7, pp. 5416–5429, Feb. 2026, https://pubs.acs.org/esthag/article-abstract/60/7/5416/5082577/Carbon-Intensity-of-United-States-Natural-Gas (accessed Sep. 9, 2026).

[5] U.S. Energy Information Administration (EIA), "Montana Natural Gas Number of Gas and Gas Condensate Wells," *eia.gov*. https://www.eia.gov/dnav/ng/hist/na1170_smt_8a.htm (accessed Aug. 9, 2026).

[6] M. E. Dewanto, "Proyeksi Produksi Migas Indonesia Sampai Dengan Tahun 2045," *Jurnal Ilmu Ekonomi (JIE)*, vol. 7, no. 02, pp. 195–210, Mei 2023.

[7] J. J. Arps, "Analysis of Decline Curves," *Transactions of the AIME*, vol. 160, no. 1, pp. 228–247, Dec. 1945, https://onepetro.org/TRANS/article/160/01/228/161823 (accessed Sep. 9, 2026).

[8] U. Rohman, J. S. Asri, H. D. Ariessanti, dan S. Wahyu, "Perbandingan algoritma regresi linear sederhana dan regresi polinomial dalam prediksi jumlah penumpang kereta api di wilayah Jabodetabek," *Jurnal ProTekInfo*, vol. 12, no. 1, hlm. 7–15, Feb. 2025.

[9] A. Wijoyo, A. Y. Saputra, S. Ristanti, S. R. Sya'Ban, M. Amalia, dan R. Febriansyah, "Pembelajaran Machine Learning," *OKTAL: Jurnal Ilmu Komputer dan Science*, vol. 3, no. 2, hlm. 375–380, Feb. 2024.

[10] C. R. Harris *et al.*, "Array programming with NumPy," *Nature*, vol. 585, no. 7825, pp. 357–362, Sep. 2020, https://www.nature.com/articles/s41586-020-2649-2 (accessed Sep. 9, 2026).

[11] S. Lestari, "Analisis Algoritma Regresi Linear Sederhana dalam Memprediksi Tingkat Penjualan Album KPOP," *INSOLOGI: Jurnal Sains dan Teknologi*, vol. 2, no. 1, pp. 199–209, Feb. 2023.

[12] M. A. Fernandes, M. M. Furlanetti, E. Gildin, and M. A. Sampaio, "Data-driven models for production forecasting and decision supporting in petroleum reservoirs," *arXiv preprint* arXiv:2508.18289, 2025. [Online]. Available: https://arxiv.org/abs/2508.18289 (accessed Oct. 25, 2026).

[13] J. S. Malensang, H. Komalig, dan D. Hatidja, "Pengembangan Model Regresi Polinomial Berganda Pada Kasus Data Pemasaran," *Jurnal Ilmiah Sains*, vol. 12, no. 2, hlm. 149–152, Okt. 2012.

[14] M. I. Ilham, D. Rosiani, and P. Rahalintar, "Prediksi Produksi Minyak Mentah di Amerika Serikat Menggunakan Regresi Polinomial," *Prosiding Seminar Nasional Teknologi Energi dan Mineral (SNTEM)*, vol. 5, no. 1, pp. 1135–1141, 2025, doi: 10.53026/prosidingsntem.v5i1.701.

[15] L. Demanet and A. Townsend, "Stable extrapolation of analytic functions," *arXiv preprint* arXiv:1605.09601, 2016. [Online]. Available: https://arxiv.org/abs/1605.09601 (accessed Oct. 25, 2026).

[16] *openpyxl - A Python library to read/write Excel 2010 xlsx/xlsm files*, openpyxl Documentation. https://openpyxl.readthedocs.io (accessed Sep. 9, 2026).

[17] A. Chugh, "MAE, MSE, RMSE, Coefficient of Determination, Adjusted R Squared — Which Metric is Better?," *Analytics Vidhya*, Dec. 7, 2020. https://medium.com/analytics-vidhya/mae-mse-rmse-coefficient-of-determination-adjusted-r-squared-which-metric-is-better-cd0326a5697e (accessed Sep. 9, 2026).

[18] Y. Diao, L. Yan, and K. Gao, "Improvement of the machine learning-based corrosion rate prediction model through the optimization of input features," *Materials & Design*, vol. 198, p. 109326, 2021. https://www.sciencedirect.com/science/article/pii/S0264127520308625 (accessed Aug. 10, 2026).