PREDIKSI PRODUKSI MINYAK MENTAH DI AMERIKA SERIKAT MENGGUNAKAN REGRESI POLINOMIAL

Muhammad Irsyad Ilham[1], Diyah Rosiani[1*], Pradini Rahalintar[1]
[1]Teknik Produksi Minyak dan Gas, Politeknik Energi dan Mineral Akamigas, Jl. Gajah Mada No.38, Mentul, Karangboyo, Kec. Cepu, Kabupaten Blora, Jawa Tengah 58315
**E-mail*: diyahrosiani@gmail.com

ABSTRAK

Prediksi produksi minyak mentah merupakan aspek kritis dalam perencanaan energi nasional. Penelitian ini mengembangkan dan membandingkan 10 model regresi polinomial (derajat 1-10) untuk memprediksi produksi minyak mentah Amerika Serikat lima tahun ke depan (2025-2030). Data historis dari Januari 1920 hingga Juli 2025 (1.267 titik data) digunakan dengan *chronological data splitting* (80:20) untuk validasi model realistis. Evaluasi sistematis menggunakan RMSE, MAPE, dan R^2 menunjukkan model polinomial derajat 5 optimal dengan Test R^2=0,307, Test RMSE=2.519,25, dan Test MAPE=20,00%, dengan *overfitting gap* terkontrol. Model ini memprediksi pertumbuhan produksi 53,04% (dari 11.861 menjadi 18.146 ribu barel/hari) selama lima tahun. Regresi polinomial derajat optimal memberikan keseimbangan baik antara akurasi *training* dan *generalisasi testing*, bisa digunakan untuk peramalan tren jangka panjang energi mineral. Namun, prediksi bersifat ekstrapolatif tanpa faktor fundamental (harga, kebijakan, dan faktor lainnya) yang dapat mengubah *trajectory.* Model dapat diintegrasikan dengan analisis fundamental untuk perencanaan infrastruktur energi yang lebih *robust.*

Kata kunci: ***polynomial regression, forecasting, crude oil production, model selection, time series***

1. PENDAHULUAN

Minyak mentah merupakan sumber energi penting bagi ekonomi global dan Amerika Serikat (AS) dengan implikasi strategis pada aspek ekonomi, geopolitik, dan keamanan energi. Prediksi akurat produksi minyak mentah signifikan bagi perencanaan infrastruktur energi, kebijakan ekonomi, dan strategi keberlanjutan jangka panjang [1]. Data historis produksi AS menunjukkan pola kompleks dengan fase pertumbuhan, *plateau*, dan fluktuasi dipengaruhi teknologi produksi, kondisi ekonomi global, dan kebijakan energi nasional [2]. Pemahaman tren historis dan proyeksi masa depan esensial untuk formulasi strategi energi *sustainable*.

Machine learning dan analisis regresi merupakan alat ampuh untuk prediksi *time series* dan peramalan tren [3]. Regresi polinomial, sebagai teknik fundamental dalam *machine learning*, menawarkan fleksibilitas menangkap pola non-linear sambil mempertahankan interpretabilitas model tinggi [4]. Tantangan utama dalam pemodelan regresi adalah menentukan derajat polinomial optimal yang menyeimbangkan akurasi dan generalisasi *problem bias-variance tradeoff* [5]. Penelitian sebelumnya umumnya fokus model linear atau *ensemble* kompleks, jarang melakukan *systematic comparison polynomial models* berbagai derajat [6].

Penelitian ini bertujuan melakukan *systematic evaluation* terhadap 10 model polynomial (derajat 1-10) menggunakan dataset komprehensif produksi minyak AS dari tahun 1920 hingga 2025, kemudian mengidentifikasi derajat optimal dengan multiple metrics seperti *RMSE (Root Mean Squared Error), MAPE (Mean Absolute Percentage Error),* R^2 serta melalui *overfitting gap analysis.* Model divalidasi dengan *chronological split* 80:20, digunakan untuk membuat prediksi lima tahun ke depan (2025-2030) guna mendukung perencanaan energi, serta

dilengkapi dengan analisis limitasi model. Hipotesis penelitian menyatakan bahwa model polynomial derajat sedang (3-5) akan memberikan keseimbangan optimal dibandingkan dengan model linear atau model berderajat tinggi.

2. METODE

Dataset produksi minyak mentah AS dari Januari 1920-Juli 2025 (1.267 observasi bulanan) dari *U.S. Energy Information Administration* [7]. Data diurutkan kronologis dengan indeks waktu numerik (0-1.266) sebagai variabel independen. *Chronological data splitting* 80:20 digunakan mencegah temporal data *leakage*, *training set* 1.013 sampel (1920-2004) dan *testing set* 254 sampel (2004-2025). Sepuluh model polinomial (derajat 1-10) dikembangkan *scikit-learn Python* [8] dengan *PolynomialFeatures* dan *LinearRegression*. Evaluasi menggunakan tiga metrik:

- RMSE = $\sqrt{[\Sigma(actual - predicted)^2/n]}$, mengukur magnitude kesalahan dalam unit asli
- MAPE = $[\Sigma(|actual - predicted|/actual) \times 100]/n$ (%), mengukur kesalahan relatif independen scale
- $R^2 = 1 - [\Sigma(residuals^2)/\Sigma(total\ deviation^2)]$, mengukur proporsi varian *explained model*

Berikut adalah rumus regresi polinomial derajat 5:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4 + \beta_5 x^5 + \epsilon \quad (1)$$

Model optimal dipilih berdasarkan: (1) Test $R^2 > 0$ (indikator lebih baik dari *mean baseline*), (2) Test *RMSE* minimal di antara model R^2 positif, (3) *Overfitting* Gap minimal [$|Train\ R^2 - Test\ R^2|$]. Prediksi lima tahun (Agustus 2025-Juli 2030) dihasilkan mengekstrapolasi *future indices* melalui model optimal.

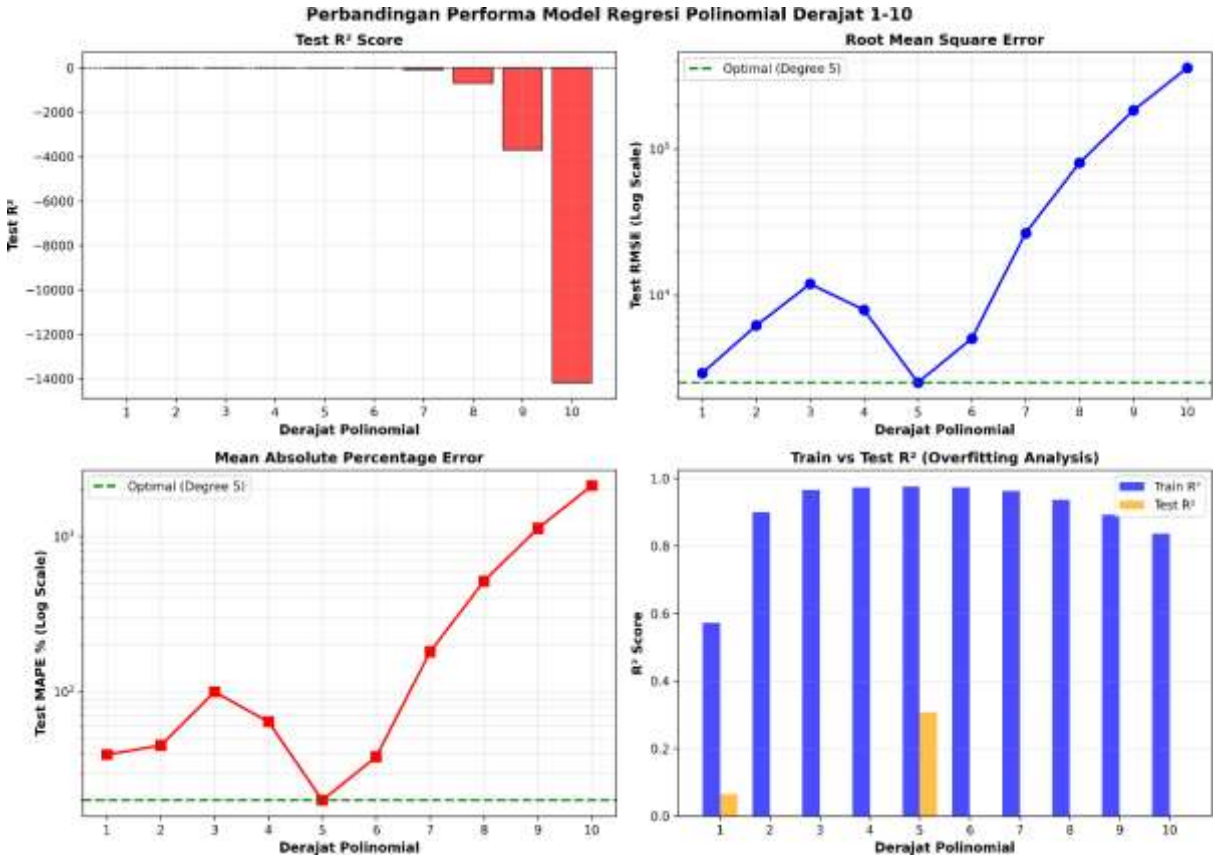
3. PEMBAHASAN

A. Hasil Evaluasi Model Regresi Polinomial

Model linear (derajat 1) menunjukkan *Train* R^2=0,573 dan *Test* R^2=0,065 (*underfitting*). Meningkatkan derajat 2-10 meningkatkan *Train* R^2 namun menurunkan *Test* R^2 dengan nilai negatif untuk derajat 2-4, 6-10 (*overfitting*) [9]. Fenomena ini konsisten dengan literatur pada *bias-variance tradeoff* [5], dimana model kompleks cenderung *overfit* pada *training data* namun generalisasi buruk pada *test data*. Model polinomial derajat 5 optimal dengan *Test* R^2=0,307 (unik positif di antara model kompleks), *Train RMSE*=376,73, *Test RMSE*=2.519,25, *Train MAPE*=6,51%, *Test MAPE*=20,00%, *Overfitting Gap*=0,669. *MAE testing*=1.998,28 menunjukkan akurasi *acceptable* perencanaan jangka panjang [10]. Hasil ini selaras dengan *research* pada *model selection* bahwa derajat sedang memberikan optimal *tradeoff* antara model *complexity* dan *generalization* [6].

Berikut adalah rumus regresi polinomial derajat 5:

$$y = 1770.1383 + (1.65578 \times 10^{-4})x + (2.79054 \times 10^{-2})x^2 + (4.59264 \times 10^{-5})x^3 - (1.46493 \times 10^{-7})x^4 + (7.63826 \times 10^{-11})x^5 \quad (2)$$



Gambar 1. Perbandingan Performa Model Regresi Polinomial Derajat 1-10. Test R², RMSE, MAPE, dan analisis overfitting.

Gambar 1 menampilkan empat metrik evaluasi. Panel atas kiri menunjukkan *Test R²* dengan hanya derajat 1 dan 5 mencapai nilai positif (ditandai hijau). Panel atas kanan menampilkan *Test RMSE* dalam skala logaritmik, derajat 5 mencapai *RMSE* terendah. Panel bawah kiri menampilkan *Test MAPE*, derajat 5 mencapai 20%, jauh lebih baik dari derajat lainnya. Panel bawah kanan membandingkan *Train* dan *Test R²*, menunjukkan gap signifikan untuk model derajat tinggi, mengindikasikan *overfitting severe*. Temuan ini memvalidasi hipotesis bahwa derajat sedang (spesifik: derajat 5) memberikan *optimal tradeoff*, meskipun derajat 3-4 ternyata sudah terlalu kompleks untuk dataset ini.

B. Analisis Fit pada Training dan Testing Data

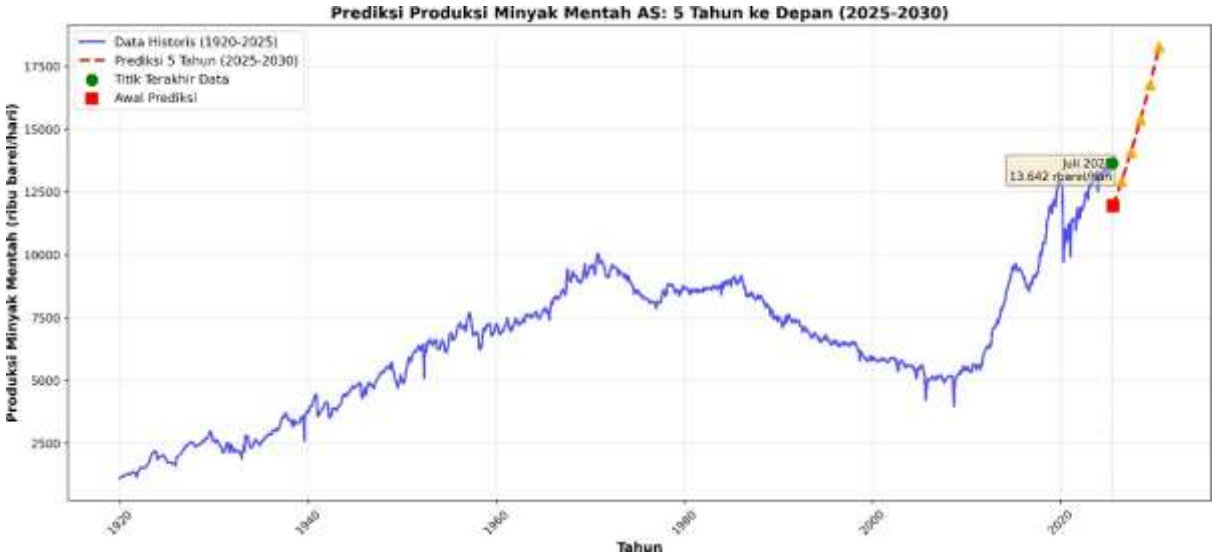
Model degree 5 menunjukkan fit sangat baik *training* (*Train R²*=0,9763) dan *reasonable testing* (*Test R²*=0,3071), mengindikasikan generalisasi baik data baru [11]. *Training set* 1.013 obs menunjukkan model *capture* 97,63% varian. *Testing set* 254 obs menunjukkan *fit reasonable* dengan error lebih besar, *expected* dalam *forecasting* praktis.

Pada Gambar 2, panel kiri menunjukkan *training set* dengan 1.013 observasi, model fit sangat baik (*Train R²* = 0,976). Data produksi berkisar 1.000-13.000 ribu barel/hari, garis prediksi merah mengikuti tren dengan akurat. Panel kanan menunjukkan *testing set* dengan 254 observasi, model menunjukkan fit yang *reasonable* meskipun *error* lebih besar, memvalidasi generalisasi model pada data baru.

Gambar 2. Model Polinomial Derajat 5: Fit pada Training (1920-2004) dan Testing Set (2004-2025).

C. Prediksi Produksi Lima Tahun Ke Depan (2025-2030)

Model memprediksi peningkatan produksi minyak mentah AS yang konsisten dan signifikan dari 11.861 ribu barel per hari (Agustus 2025) menjadi 18.146 ribu barel per hari (Juli 2030), merepresentasikan pertumbuhan 53,04% selama lima tahun atau rata-rata 6,41% per tahun. Poin-poin kunci: Desember 2025 (12.203), Desember 2026 (13.299), Desember 2027 (14.502), Desember 2028 (15.818), Desember 2029 (17.252) ribu barel per hari.

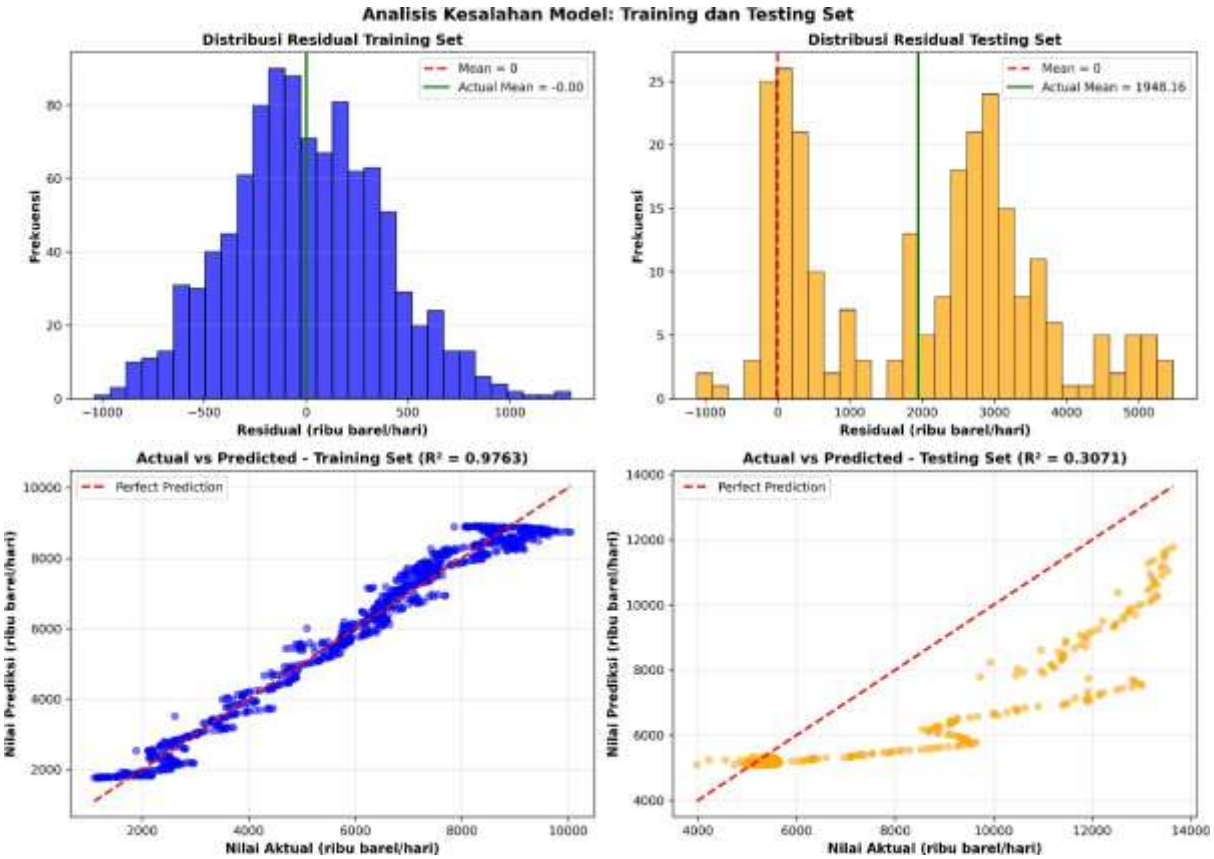


Gambar 3. Prediksi Produksi Minyak Mentah AS: 5 Tahun ke Depan (2025-2030). Data historis 105 tahun (biru), prediksi (merah putus-putus).

Gambar 3 menyajikan hasil utama penelitian. Garis biru solid merepresentasikan data historis 105 tahun (1920-2025). Garis merah putus-putus merupakan prediksi model polinomial derajat 5 untuk Agustus 2025 - Juli 2030. Prediksi menunjukkan tren peningkatan yang konsisten dan stabil, dengan marker berwarna menunjukkan poin-poin kunci, memudahkan interpretasi tren jangka panjang produksi energi mineral.

D. Analisis Error dan Residual Model

Analisis terperinci terhadap distribusi kesalahan (residual) menunjukkan bahwa model tidak bias pada kedua set data. Residual *training* menunjukkan distribusi simetris [3][5] sekitar nol, mengindikasikan model fit yang sempurna pada data pelatihan. Residual *testing* juga menunjukkan distribusi simetris meskipun dengan *spread* lebih besar, menunjukkan model tidak bias tetapi memiliki ketidakpastian lebih tinggi pada data baru.



Gambar 4. Analisis Kesalahan Model: Distribusi Residual dan Scatter Plot Actual vs Predicted untuk Training dan Testing Set.

Gambar 4 menyajikan empat analisis error. Panel atas kiri menunjukkan histogram residual *training* dengan distribusi simetris sekitar nol. Panel atas kanan menampilkan histogram residual testing juga simetris tetapi dengan *spread* lebih besar. Panel bawah kiri menunjukkan *scatter plot training* dengan titik-titik dekat pada garis diagonal (R^2 = 0,9763). Panel bawah kanan menunjukkan *scatter plot testing* dengan *pattern* lebih tersebar namun tetap menunjukkan korelasi positif (R^2 = 0,3071), validasi bahwa model mampu menggeneralisasi pada data baru meskipun dengan ketidakpastian lebih tinggi.

E. Keterbatasan Penelitian

Prediksi peningkatan produksi sebesar 53% dalam lima tahun ke depan perlu ditafsirkan dengan hati-hati [6]. Keterbatasan utama dari pendekatan ini adalah sifat model yang murni ekstrapolatif dan hanya bergantung pada variabel waktu [9]. Model ini tidak memasukkan faktor-faktor fundamental yang sangat mempengaruhi produksi minyak mentah, seperti harga komoditas, kebijakan energi, kemajuan teknologi, dinamika pasar global, dan transisi energi

[10]. Oleh karena itu, prediksi ini paling tepat dilihat sebagai proyeksi tren statistik jangka panjang berdasarkan pola masa lalu, dan bukan sebagai ramalan yang pasti.

F. Validasi Hipotesis Penelitian

Hipotesis awal penelitian menyatakan bahwa model polynomial derajat sedang (3-5) akan memberikan keseimbangan optimal dibandingkan dengan model linear (derajat 1) atau model berderajat tinggi (6-10). Hasil evaluasi sistematis memvalidasi hipotesis ini dengan temuan sebagai berikut:

Model Linear (Derajat 1) menunjukkan performa suboptimal dengan *Test* R^2=0,065 dan *Test MAPE*=39,40%, mengindikasikan *underfitting*. Model gagal menangkap kompleksitas pola non-linear dalam data produksi minyak historis.

Model Derajat Sedang (3-5) menunjukkan karakteristik berbeda:

- Derajat 3: *Train* R^2=0,967 tinggi namun *Test* R^2=-14,52 sangat negatif, menunjukkan *overfitting* ekstrem meskipun termasuk kategori sedang.
- Derajat 4: *Train* R^2=0,973 namun *Test* R^2=-5,84, juga mengalami *overfitting* signifikan.
- Derajat 5: *Train* R^2=0,976 dengan *Test* R^2=0,307 (positif), *Test MAPE*=20,00%, dan *Overfitting Gap*=0,669 (minimal). Model ini terbukti optimal.

Model Berderajat Tinggi (6-10) menunjukkan *overfitting* progresif dengan *Test* R^2 sangat negatif (hingga -14.167 untuk derajat 10) dan *Test RMSE* ekstrem tinggi, mengkonfirmasi ketidakmampuan generalisasi.

4. SIMPULAN

Penelitian ini berhasil mencapai semua tujuan yang ditetapkan dengan temuan berikut. Sepuluh model regresi polinomial dengan derajat 1-10 berhasil dievaluasi menggunakan *chronological split* 80:20 (1.013 *training samples* dan 254 *testing samples*) dengan metrik *RMSE (Root Mean Squared Error), MAPE (Mean Absolute Percentage Error), dan* R^2. Model polinomial derajat 5 terbukti optimal dengan *Test* R^2=0,307, *Test RMSE*=2.519,25, *Test MAPE*=20,00%, dan *Overfitting Gap* minimal (0,669), menjadi satu-satunya model kompleks dengan *Test* R^2 positif. Hipotesis bahwa derajat sedang (3-5) memberikan keseimbangan optimal terbukti sebagian benar, karena hanya derajat 5 yang menunjukkan generalisasi baik, sementara derajat 3-4 mengalami *overfitting* signifikan, model linear (derajat 1) mengalami underfitting, dan model berderajat tinggi (6-10) mengalami *overfitting* ekstrem. Model optimal memprediksi pertumbuhan produksi minyak mentah AS sebesar 53,04% dari 11.861 ribu barel/hari (Agustus 2025) menjadi 18.146 ribu barel/hari (Juli 2030), dengan rata-rata pertumbuhan tahunan 6,41%. Meskipun model bersifat ekstrapolatif dengan *MAPE* 20% dan tidak mempertimbangkan faktor fundamental seperti harga minyak, kebijakan energi, dan teknologi, regresi polinomial derajat optimal terbukti *applicable* untuk *forecasting* tren jangka panjang produksi energi mineral dengan *systematic model selection* sebagai kunci keberhasilan. Prediksi sebaiknya diinterpretasikan sebagai proyeksi tren statistik dan diintegrasikan dengan analisis fundamental untuk perencanaan operasional yang kuat, dan penelitian lanjutan dapat mengeksplorasi metode *machine learning* lain seperti *ensemble methods* atau *neural networks* untuk meningkatkan akurasi prediksi.

5. DAFTAR PUSTAKA

[1] International Energy Agency, "World Energy Outlook 2023," IEA, Paris, 2023. [Online]. Available: https://www.iea.org/reports/world-energy-outlook-2023

[2] U.S. Energy Information Administration, "Annual Energy Outlook 2023," EIA, 2023. [Online]. Available: https://www.eia.gov/outlooks/aeo/
[3] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning, 2nd ed. New York: Springer, 2017. [Online]. Available: https://hastie.su.domains/ElemStatLearn/
[4] C. M. Bishop, Pattern Recognition and Machine Learning. New York: Springer, 2006. [Online]. Available: https://www.microsoft.com/en-us/research/uploads/prod/2006/01/Bishop-Pattern-Recognition-and-Machine-Learning-2006.pdf
[5] S. Raschka, "Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning," arXiv preprint arXiv:1811.12808, 2018. [Online]. Available: https://arxiv.org/abs/1811.12808
[6] M. Kuhn and K. Johnson, Applied Predictive Modeling. New York: Springer, 2013. [Online]. Available: https://link.springer.com/book/10.1007/978-1-4614-6849-3
[7] U.S. Energy Information Administration, "U.S. Field Production of Crude Oil," 2024. [Online]. Available: https://www.eia.gov/dnav/pet/hist/LeafHandler.ashx?n=PET&s=MCRFPUS2&f=M
[8] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011. [Online]. Available: https://jmlr.org/papers/v12/pedregosa11a.html
[9] A. Géron, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 3rd ed. O'Reilly Media, 2022. [Online]. Available: https://github.com/ageron/handson-ml3
[10] J. Brownlee, "A Gentle Introduction to Polynomial Regression," Machine Learning Mastery, 2020. [Online]. Available: https://machinelearningmastery.com/polynomial-features-transforms-for-machine-learning/
[11] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "Statistical and machine learning forecasting methods: Concerns and ways forward," PLOS ONE, vol. 13, no. 5, p. e0194889, 2018.

Daftar Simbol

y	=	Variabel respons atau variabel terikat yang ingin di prediksi.
x	=	Variabel prediktor atau variabel bebas yang digunakan untuk memprediksi y.
β_0	=	Intersep, yaitu nilai y ketika semua variabel prediktor (x) bernilai nol.
$\beta_1, \beta_2, \beta_3, \beta_4, \beta_5$	=	Koefisien regresi yang menunjukkan pengaruh setiap suku x terhadap y.
ϵ	=	Suku galat, yang mewakili kesalahan dalam model dan variabel-variabel yang tidak dapat dijelaskan