

LINEAR DAN POLYNOMIAL REGRESSION

Oleh: Kelompok 3



NAMA ANGGOTA



01

Meutia Aini
(2208107010005)

02

Akhsania Maisa Rahmah
(2208107010017)

03

Fadli Ahmad Yazid
(2208107010032)

04

Muhammad Mahathir
(2208107010056)

05

Muhammad Aufa Zaikra
(2208107010070)

DATASET

Dataset “Salary Data” berisi informasi gaji berdasarkan beberapa fitur seperti umur, jenis kelamin, tingkat pendidikan, jabatan pekerjaan, tahun pengalaman kerja, dan gaji tahunan dalam dolar. Tujuan analisis adalah membangun model prediksi gaji menggunakan regresi linear dan polinomial. Evaluasi model dilakukan dengan metrik MAE, MSE, dan R² score.

METODOLOGI

1. Pemahaman Dataset
2. Eksplorasi dan Pra-pemrosesan Data
3. Implementasi Model
4. Evaluasi Model
5. Analisis Model

PEMAHAMAN DATASET

Sumber Data:

- Dataset: "Salary Data"
- Jumlah entri: 6.704

Variabel dalam Dataset:

- Age: Usia karyawan (numerik, dalam tahun).
- Gender: Jenis kelamin (kategorikal: Male, Female, Other).
- Education Level: Tingkat pendidikan (kategorikal: High School, Bachelor's, Master's, PhD).
- Job Title: Jabatan pekerjaan (kategorikal, misalnya: Software Engineer, Data Analyst).

- Years of Experience: Tahun pengalaman kerja (numerik).
- Salary: Gaji tahunan dalam dolar (numerik, variabel target).

Informasi Dataset:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6704 entries, 0 to 6703
Data columns (total 6 columns):
 #   Column            Non-Null Count  Dtype  
--- 
 0   Age               6702 non-null    float64
 1   Gender            6702 non-null    object  
 2   Education Level   6701 non-null    object  
 3   Job Title         6702 non-null    object  
 4   Years of Experience 6701 non-null    float64
 5   Salary            6699 non-null    float64
dtypes: float64(3), object(3)
memory usage: 314.4+ KB
```

Kolom	Jumlah Non-Null	Tipe Data
Age	6702	float64
Gender	6702	object
Education Level	6701	object
Job Title	6702	object
Years of Experience	6701	float64
Salary	6699	float64

Table 1: Informasi Dataset "Salary Data"

PEMAHAMAN DATASET

Statistik deskriptif untuk fitur numerik

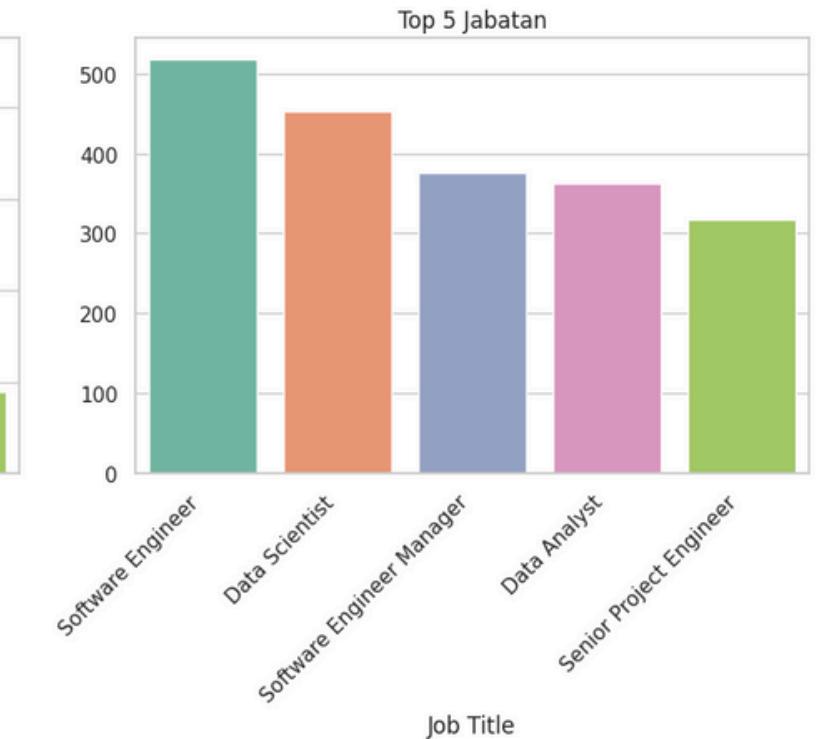
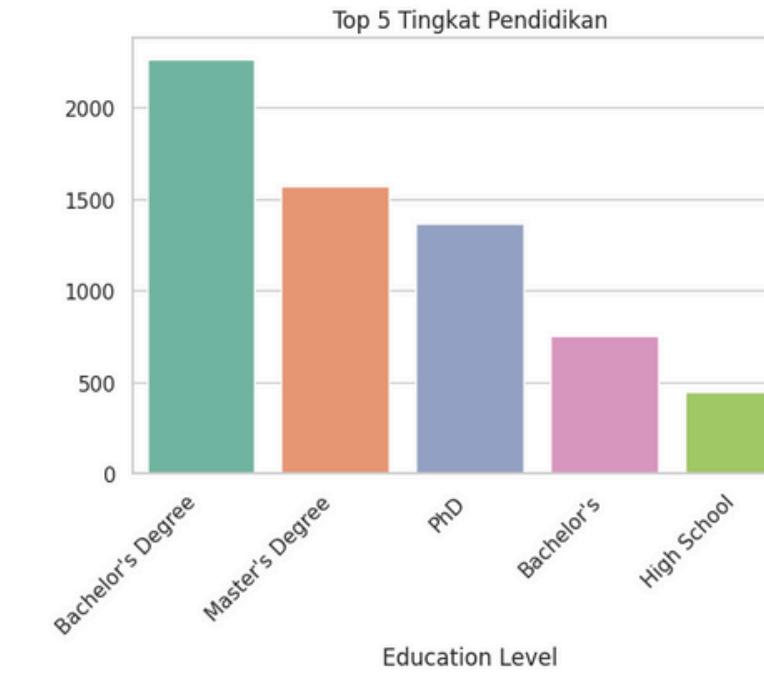
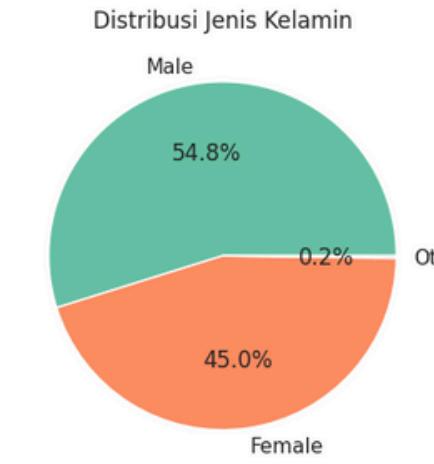
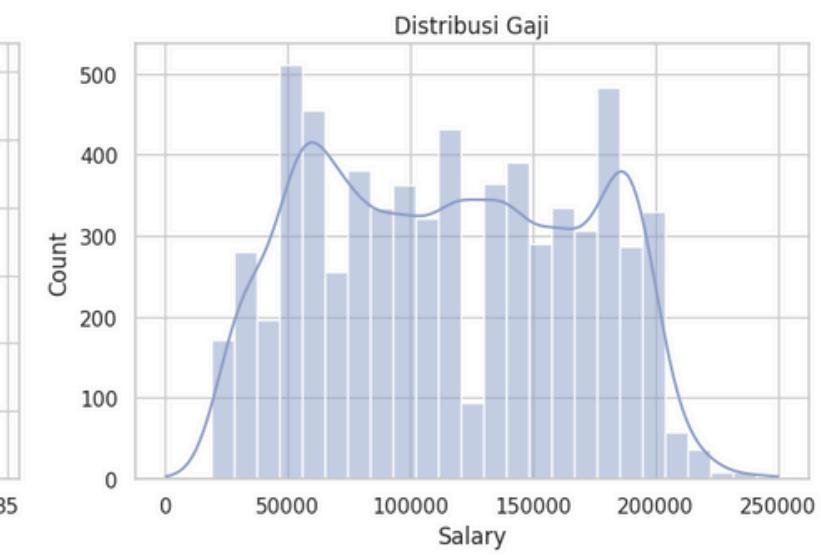
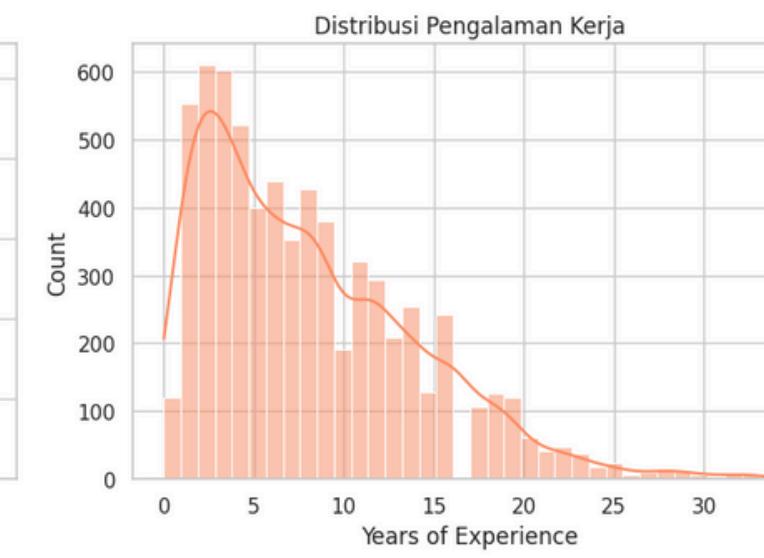
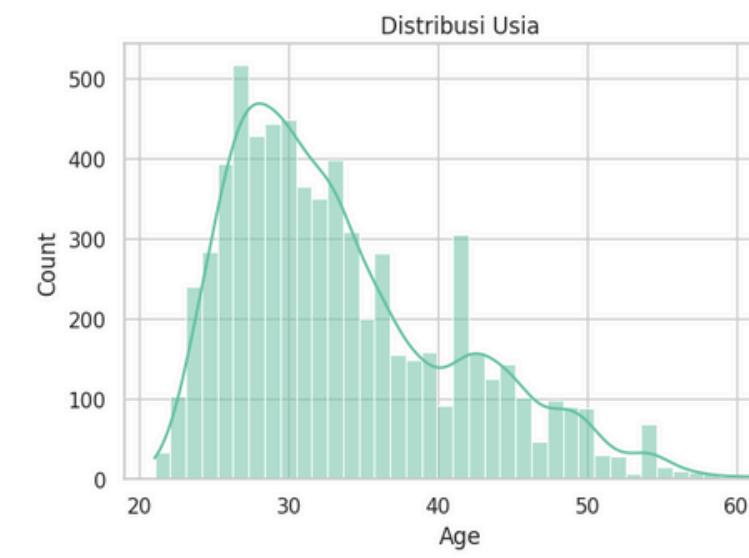
```
          Age  Years of Experience      Salary
count    6702.000000    6701.000000  6699.000000
mean     33.620859     8.094687  115326.964771
std      7.614633     6.059003  52786.183911
min     21.000000     0.000000   350.000000
25%    28.000000     3.000000  70000.000000
50%    32.000000     7.000000 115000.000000
75%    38.000000    12.000000 160000.000000
max     62.000000    34.000000 250000.000000
```

Contoh Data Awal:

Index	Age	Gender	Education Level	Job Title	Years of Experience	Salary
0	32.0	Male	Bachelor's	Software Engineer	5.0	90000.0
1	28.0	Female	Master's	Data Analyst	3.0	65000.0
2	45.0	Male	PhD	Senior Manager	15.0	150000.0
3	36.0	Female	Bachelor's	Sales Associate	7.0	60000.0
4	52.0	Male	Master's	Director	20.0	200000.0

EKSPLORASI DATA

Distribusi Fitur Numerik dan Kategorikal



Fitur Numerik

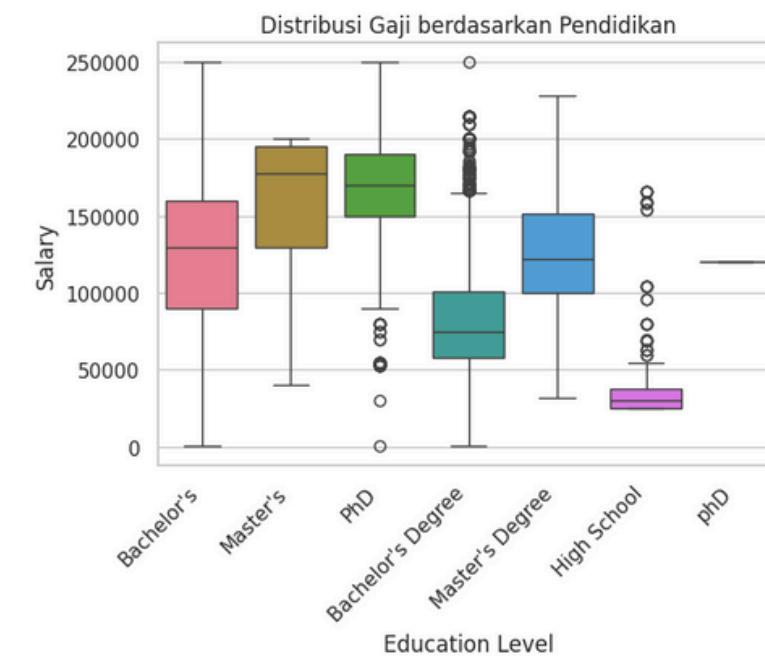
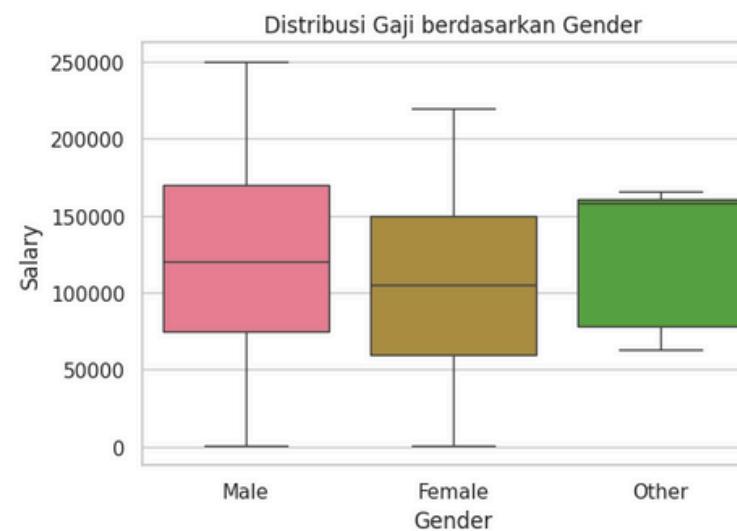
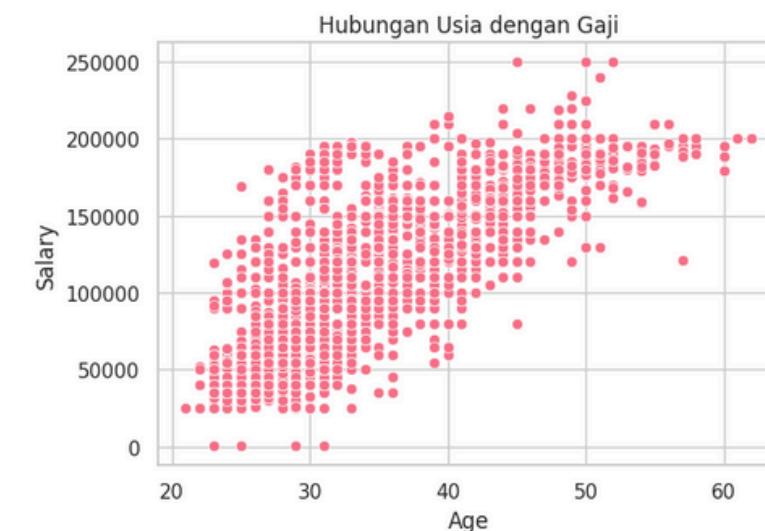
- Distribusi Usia: Usia karyawan berkisar antara 21 hingga 62 tahun, dengan mayoritas berada di rentang 28–38 tahun.
- Distribusi Pengalaman Kerja: Pengalaman kerja berkisar dari 0 hingga 34 tahun, dengan mayoritas karyawan memiliki pengalaman 3–12 tahun.
- Distribusi Gaji: Gaji berkisar dari \$350 hingga \$250.000, dengan distribusi multimodal yang menunjukkan puncak di sekitar \$50.000, \$100.000, dan \$150.000– \$200.000.

Fitur Kategorikal

- Distribusi Jenis Kelamin: Distribusi hampir seimbang dengan 54.8% Male, 45.0% Female, dan 0.2% Other.
- Tingkat Pendidikan: Mayoritas karyawan memiliki gelar Bachelor's, diikuti oleh Master's, PhD, dan High School. •
- Jabatan: Jabatan teratas meliputi Software Engineer, Data Scientist, Senior Manager, Data Analyst, dan Project Engineer

EKSPLORASI DATA

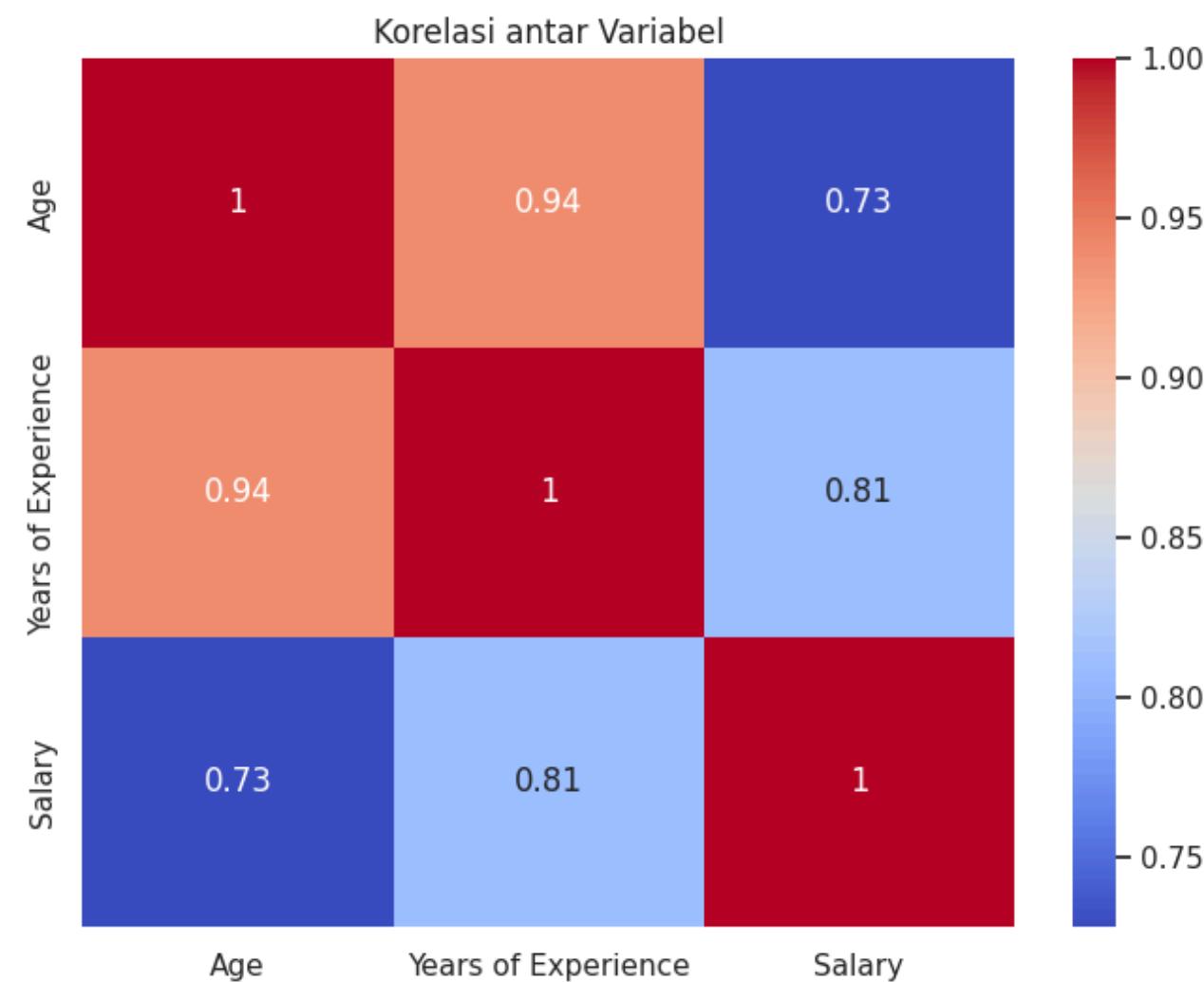
Hubungan antar Variabel



- Usia vs Gaji: Terdapat hubungan positif antara usia dan gaji, meskipun variasinya cukup besar.
- Pengalaman Kerja vs Gaji: Hubungan non-linier yang jelas terlihat, dengan peningkatan gaji yang lebih cepat pada tahun-tahun awal pengalaman.
- Gender vs Gaji: Distribusi gaji antara Male dan Female relatif seimbang, dengan sedikit variasi.
- Pendidikan vs Gaji: Gaji meningkat seiring dengan tingkat pendidikan, dengan PhD memiliki median gaji tertinggi.
- Jabatan vs Gaji: Jabatan seperti Director dan Senior Manager memiliki rata-rata gaji tertinggi.

PRA-PEMROSESAN DATA

Analisis Korelasi antar Variabel



Korelasi antara Years of Experience dan Salary adalah 0.81, menunjukkan hubungan positif yang kuat. Age juga memiliki korelasi positif dengan Salary (0.73), tetapi lebih lemah dibandingkan pengalaman kerja

Pemeriksaan Missing Value

Output awal:

```
Age           2
Gender        2
Education Level 3
Job Title     2
Years of Experience 3
Salary         5
dtype: int64
```

Output setelah penghapusan:

```
Age           0
Gender        0
Education Level 0
Job Title     0
Years of Experience 0
Salary         0
dtype: int64
```

Alasan Penghapusan:

Jumlah missing values sangat kecil (<1% dari total 6.704 entri), sehingga penghapusan tidak signifikan memengaruhi representativitas dataset.

PRA-PEMROSESAN DATA

Standardisasi Nilai Kategorikal

Nilai unik pada kolom Education Level:

```
["Bachelor's" "Master's" 'PhD' "Bachelor's Degree" "Master's Degree"  
 'High School' 'phD']
```

Setelah standardisasi Education Level:

```
["Bachelor's Degree" "Master's Degree" 'PhD' 'High School']
```

Interpretasi:

Standardisasi dilakukan untuk menghilangkan inkonsistensi penulisan seperti "PhD" dan "phD", sehingga data lebih seragam.

Persiapan Data untuk Pemodelan

memisahkan fitur dan target, serta melakukan one-hot encoding untuk fitur kategorikal.

Kolom kategorikal: ['Education Level', 'Job Title']

Kolom numerik: ['Years of Experience']

Ukuran data setelah splitting:

Training set: (5358, 3) sampel

Test set: (1340, 3) sampel

IMPLEMENTASI MODEL

Pemodelan dengan Linear Regression

Membangun pipeline untuk preprocessing dan linear regression.

```
● ● ●  
# Linear Regression Pipeline  
lr_pipeline = Pipeline([  
    ('preprocessor', preprocessor),  
    ('regressor', LinearRegression())  
])  
  
# Training Linear Regression  
lr_pipeline.fit(X_train, y_train)  
y_pred_lr = lr_pipeline.predict(X_test)
```

IMPLEMENTASI MODEL

Pemodelan dengan Polynomial Regression

Mencoba beberapa derajat polynomial untuk menemukan model terbaik.

```
● ● ●  
1 # Polynomial Regression Pipelines  
2 poly_degrees = [2, 3]  
3 poly_predictions = {}  
4  
5 for degree in poly_degrees:  
6     poly_pipeline = Pipeline([  
7         ('preprocessor', preprocessor),  
8         ('poly_features', PolynomialFeatures(degree=degree)),  
9         ('regressor', LinearRegression())  
10    ])  
11  
12    # Training Polynomial Regression  
13    poly_pipeline.fit(X_train, y_train)  
14    poly_predictions[degree] = poly_pipeline.predict(X_test)
```

EVALUASI MODEL

Evaluasi Regresi Linear

Linear Regression Performance:
MAE: 15487.83
MSE: 458246826.30
RMSE: 21406.70
R2: 0.84

Evaluasi Regresi Polinomial

Polynomial Regression (Degree 2) Performance:
MAE: 10155.95
MSE: 247653840.17
RMSE: 15737.02
R2: 0.91

Polynomial Regression (Degree 3) Performance:
MAE: 9940.47
MSE: 740203043.98
RMSE: 27206.67
R2: 0.74

EVALUASI MODEL

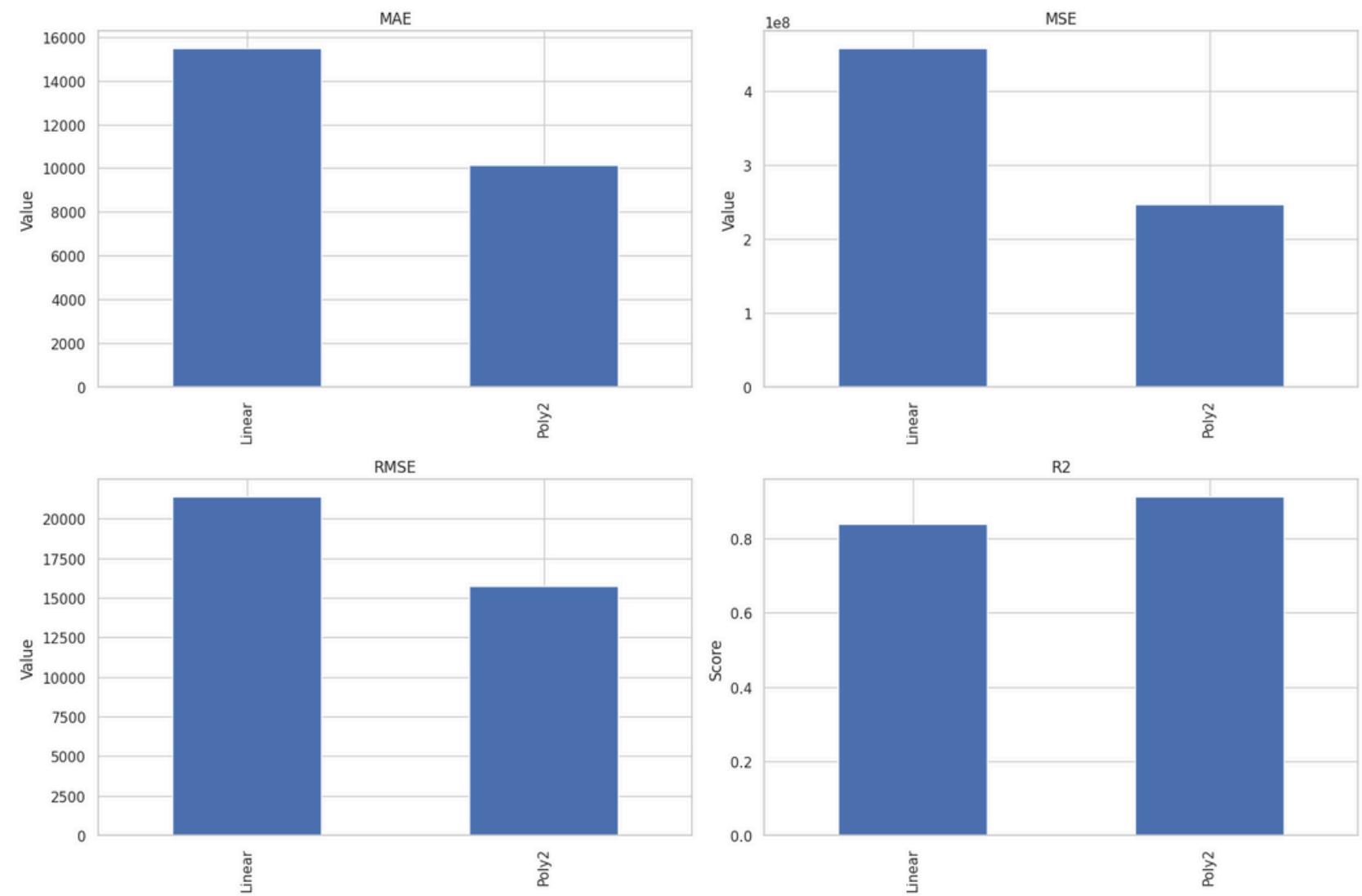
- Hasil evaluasi model disajikan dalam tabel

Model	MAE	MSE	RMSE	R ²
Regresi Linear	15487.83	458246826.30	21406.70	0.84
Regresi Polinomial (Derajat 2)	10155.95	247653840.17	15737.02	0.91
Regresi Polinomial (Derajat 3)	9940.47	740203043.98	27206.67	0.74

Interpretasi:

- Regresi Linear: R² sebesar 0.84 menunjukkan model cukup baik, tetapi MSE dan MAE yang tinggi mengindikasikan adanya error besar pada beberapa prediksi.
- Regresi Polinomial Derajat 2: R² tertinggi (0.91) dan MSE serta MAE terendah menunjukkan performa terbaik.
- Regresi Polinomial Derajat 3: R² menurun (0.74) dan MSE meningkat, menunjukkan tanda-tanda overfitting.

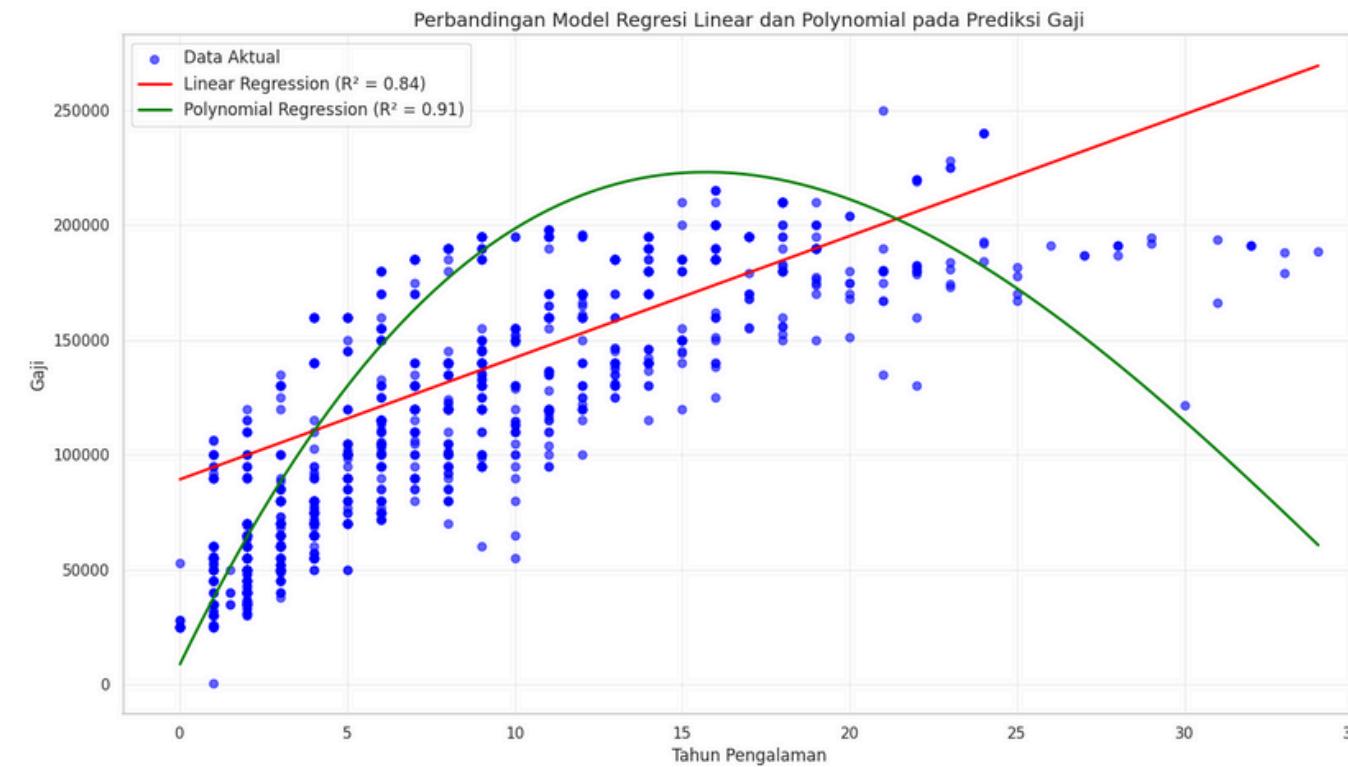
Visualisasi Performa Model



Visualisasi menunjukkan bahwa regresi polinomial derajat 2 secara konsisten memiliki nilai MAE, MSE, dan RMSE yang lebih rendah, serta R² yang lebih tinggi dibandingkan regresi linear.

ANALISIS HASIL

Visualisas Regresi



Interpretasi:

- Regresi Linear: Garis regresi linear menunjukkan hubungan positif yang kuat antara pengalaman kerja dan gaji, dengan $R^2 = 0.84$.
- Regresi Polinomial: Kurva polinomial derajat 2 lebih sesuai dengan data, dengan $R^2 = 0.91$, menangkap pola non-linier seperti perlambatan kenaikan gaji pada pengalaman tinggi.

Regresi polinomial derajat 2 menunjukkan performa terbaik dengan R^2 sebesar 0.91 dan MAE sebesar 10.155. Model ini mengungguli regresi linear (R^2 : 0.84, MAE: 15.487) karena mampu menangkap hubungan non-linier antara pengalaman kerja dan gaji.

Meskipun derajat 3 memiliki MAE lebih rendah (9.940), model tersebut menunjukkan overfitting dengan R^2 lebih rendah dan MSE jauh lebih tinggi. Visualisasi juga mendukung bahwa hubungan non-linier lebih baik ditangkap oleh model polinomial derajat 2.

KESIMPULAN

- Regresi Linear: $R^2 = 0.84$, MSE = 458246826.30, RMSE = 21406.70, MAE = 15487.83.
- Regresi Polinomial Derajat 2: $R^2 = 0.91$, MSE = 247653840.17, RMSE = 15737.02, MAE = 10155.95.
- Regresi Polinomial Derajat 3: $R^2 = 0.74$, MSE = 740203043.98, RMSE = 27206.67, MAE = 9940.47.

Regresi polinomial derajat 2 terbukti sebagai model terbaik karena memberikan performa yang lebih baik dengan R^2 tertinggi dan error terendah. Pengalaman kerja merupakan prediktor utama gaji, dengan hubungan non-linier yang jelas terlihat.

Rekomendasi:

- Gunakan model polinomial derajat 2 untuk prediksi gaji yang lebih akurat, tetapi perhatikan potensi overfitting pada data baru.
- Untuk pengembangan lebih lanjut, eksplorasi fitur tambahan seperti Age dan Gender, atau gunakan model lain seperti regresi ridge untuk mengatasi overfitting.



TERIMA KASIH

