Laporan Mata Kuliah Pembelajaran Mesin

TUGAS 2
LINEAR DAN POLYNOMIAL REGRESSION

disusun untuk memenuhi
tugas mata kuliah Pembelajaran Mesin A

oleh :

Sadinal Mufti	(2208107010007)
M. Agradika Ridhal Eljatin	(2208107010020)
Jihan Nabilah	(2208107010035)
Firjatullah Afny Abus	(2208107010059)
Athar Rayyan Muhammad	(2208107010074)

DEPARTEMEN INFORMATIKA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS SYIAH KUALA

TAHUN 2025

1. Pendahuluan

1.1. Latar Belakang

Dalam era digital, analisis data menjadi aspek krusial dalam pengambilan keputusan. Salah satu pendekatan yang sering digunakan dalam analisis data adalah regresi, baik regresi linear maupun regresi polinomial. Regresi digunakan untuk memahami hubungan antara variabel independen dengan variabel dependen dan membuat prediksi berdasarkan pola yang ditemukan.

Pada tugas ini, kami menggunakan dataset yang berkaitan dengan performa akademik siswa. Dataset ini terdiri dari data siswa yang mencakup informasi tentang kebiasaan belajar, tingkat kehadiran, serta nilai akademik sebelumnya. Selain itu, dataset ini juga berisi faktor tambahan seperti kondisi sosial ekonomi siswa dan akses mereka terhadap sumber belajar. Dengan menggunakan teknik regresi, kami bertujuan untuk memahami pola hubungan antar variabel serta memprediksi nilai akademik siswa berdasarkan faktor-faktor yang tersedia. Analisis ini dapat membantu dalam mengidentifikasi faktor utama yang mempengaruhi keberhasilan akademik serta memberikan wawasan bagi pendidik dalam merancang strategi pembelajaran yang lebih efektif.

1.2. Tujuan Tugas

- Memahami karakteristik dataset yang digunakan.
- Melakukan eksplorasi dan pra-pemrosesan data.
- Mengimplementasikan model regresi linear dan regresi polinomial.
- Mengevaluasi kinerja model berdasarkan metrik evalua
- Menyimpulkan apakah model yang dibuat cukup baik dalam memprediksi target.

2. Data Description

2.1. Informasi Dataset

- Nama Dataset: Student Performance (Multiple Linear Regression)
- Sumber:https://www.kaggle.com/datasets/nikhil7280/student-performance-multiple-linear-regression
- Deskripsi Singkat: Dataset "Student Performance (Multiple Linear Regression)" dari Kaggle berisi 10.000 catatan siswa, masing-masing mencakup berbagai prediktor dan indeks kinerja akademik. Dataset ini dirancang untuk menganalisis faktor-faktor yang mempengaruhi performa akademik siswa.

2.2. Struktur Dataset

- Jumlah Sampel: 10.000 baris
- Jumlah Fitur: 6 kolom

- Label: Kolom Performance Index (Indeks kinerja akademik siswa)
- Format Data: CSV (Comma-Separated Values)
- Deskripsi Kolom:
    - **Hours Studied** – Jumlah jam yang dihabiskan siswa untuk belajar
    - **Previous Scores** – Nilai ujian sebelumnya
    - **Extracurricular Activities** – Apakah siswa mengikuti kegiatan ekstrakurikuler
    - **Sleep Hours** – Rata-rata jumlah jam tidur siswa per hari
    - **Sample Question Papers Practiced** – Jumlah latihan soal yang dikerjakan siswa
    - **Performance Index** (target) – Index kinerja akademik siswa (bernilai antara 10 - 100)

	Hours Studied	Previous Scores	Extracurricular Activities	Sleep Hours	Sample Question Papers Practiced	Performance Index
0	7	99	Yes	9	1	91.0
1	4	82	No	4	2	65.0
2	8	51	Yes	7	2	45.0
3	5	52	Yes	5	2	36.0
4	7	75	No	8	5	66.0

Gambar 1. Struktur Awal Data

3. Pemahaman Dataset

3.1. Distribusi Data

	Hours Studied	Previous Scores	Sleep Hours	Sample Question Papers Practiced	Performance Index
count	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000
mean	4.992900	69.445700	6.530600	4.583300	55.224800
std	2.589309	17.343152	1.695863	2.867348	19.212558
min	1.000000	40.000000	4.000000	0.000000	10.000000
25%	3.000000	54.000000	5.000000	2.000000	40.000000
50%	5.000000	69.000000	7.000000	5.000000	55.000000
75%	7.000000	85.000000	8.000000	7.000000	71.000000
max	9.000000	99.000000	9.000000	9.000000	100.000000

Gambar 2. Distribusi Data

Dapat dilihat diatas Dataset terlihat seimbang dengan nilai yang wajar. Rata-rata Hours Studied adalah 5 jam, Sleep Hours sekitar 6,5 jam, dan Performance Index berkisar di 55. Previous Scores menunjukkan variasi yang cukup besar (std = 17.34), sedangkan Sample Question Papers memiliki distribusi yang cukup merata dari 0 hingga 9. Nilai minimum dan maksimum setiap fitur juga berada dalam rentang yang masuk akal.

3.2. Mendeteksi Outlier pada Data

```python
def detect_outliers(df, column):
    Q1 = df[column].quantile(0.25)
    Q3 = df[column].quantile(0.75)
    IQR = Q3 - Q1
    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR
    outliers = df[(df[column] < lower_bound) | (df[column] > upper_bound)]
    return outliers

# Mendeteksi outlier untuk setiap kolom numerik
numeric_columns = ['Hours Studied', 'Previous Scores', 'Sleep Hours',
                   'Sample Question Papers Practiced', 'Performance Index']

for column in numeric_columns:
    outliers = detect_outliers(df, column)
    print(f"Jumlah outlier di kolom {column}: {len(outliers)}")
    if len(outliers) > 0:
        print(f"Range nilai normal untuk {column}: "
              f"{df[column].quantile(0.25) - 1.5 * (df[column].quantile(0.75) - df[column].quantile(0.25)):.2f} - "
              f"{df[column].quantile(0.75) + 1.5 * (df[column].quantile(0.75) - df[column].quantile(0.25)):.2f}")
```

Gambar 3. Kode Deteksi Outlier pada Data

```
Jumlah outlier di kolom Hours Studied: 0
Jumlah outlier di kolom Previous Scores: 0
Jumlah outlier di kolom Sleep Hours: 0
Jumlah outlier di kolom Sample Question Papers Practiced: 0
Jumlah outlier di kolom Performance Index: 0
```

Gambar 4. Hasil Outlier Data

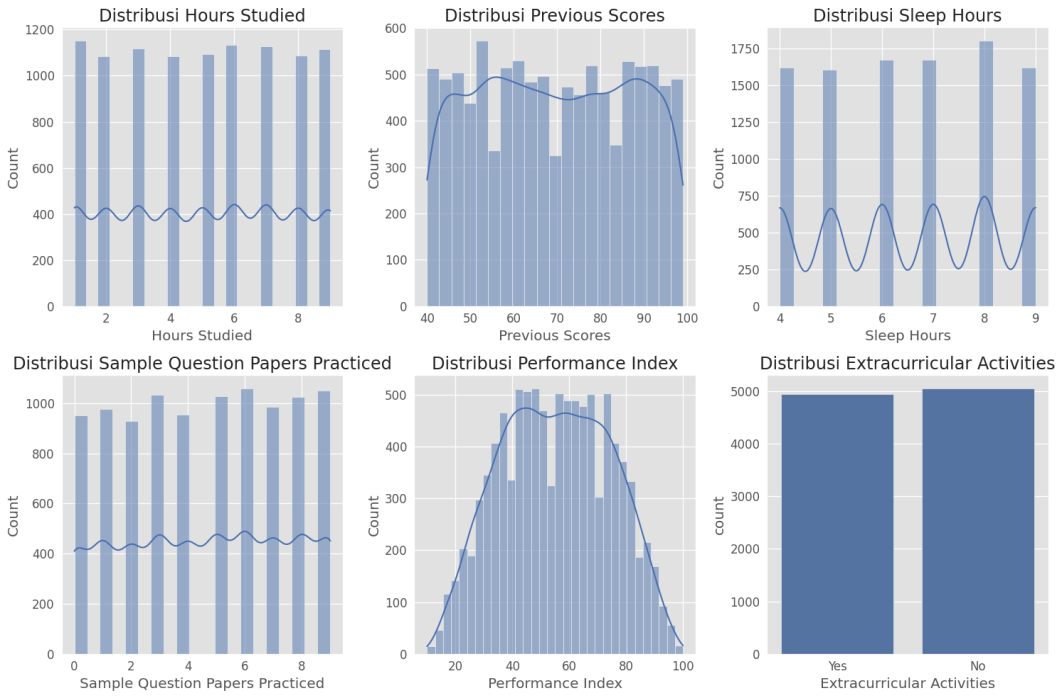
Dari hasil yang didapat dapat ditarik kesimpulan bahwa dataset ini tergolong bersih.

4. Analisis Awal Data

4.1. Distribusi Fitur Numerik dan Analisis

```
plt.figure(figsize=(15, 10))

plt.subplot(2, 3, 1)
sns.histplot(df['Hours Studied'], kde=True)
plt.title('Distribusi Hours Studied')

plt.subplot(2, 3, 2)
sns.histplot(df['Previous Scores'], kde=True)
plt.title('Distribusi Previous Scores')

plt.subplot(2, 3, 3)
sns.histplot(df['Sleep Hours'], kde=True)
plt.title('Distribusi Sleep Hours')

plt.subplot(2, 3, 4)
sns.histplot(df['Sample Question Papers Practiced'], kde=True)
plt.title('Distribusi Sample Question Papers Practiced')

plt.subplot(2, 3, 5)
sns.histplot(df['Performance Index'], kde=True)
plt.title('Distribusi Performance Index')

plt.subplot(2, 3, 6)
sns.countplot(x='Extracurricular Activities', data=df)
plt.title('Distribusi Extracurricular Activities')

plt.tight_layout()
plt.show()
plt.close()
```

Gambar 5. Kode untuk eksplorasi distribusi



Gambar 6. Output Distribusi Fitur Numerik Data

Dapat dilihat bahwa sebagian besar data terdistribusi normal atau seragam, dengan sedikit kecenderungan ke kanan pada **Previous Scores**. Tidak ada outlier yang signifikan dalam grafik ini, menunjukkan data relatif bersih.

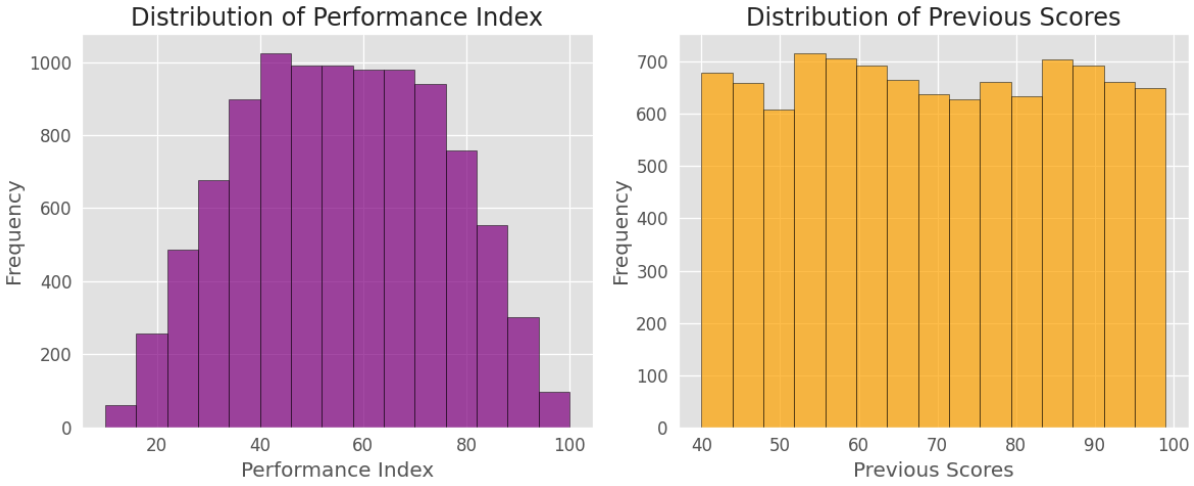
4.2. Perbandingan Variabel Dependen dan Independen

Variabel target adalah 'Performance Index' akan dibandingkan dengan total dari 'Previous Score', yang mana ini akan dapat memperlihatkan tren dari perfoma murid seiring waktu.

	Previous Scores	Performance Index
0	694457.0	552248.0

Gambar 7. Hasil Perbandingan Kedua Variabel

Dapat dilihat bahwa total nilai menurun, dimana total nilai ujian sebelumnya lebih tinggi daripada total nilai ujian saat ini. Ini menandakan bahwa mayoritas murid mengalami penurunan. Selanjutnya akan dicari tahu faktor yang menyebabkan penurunan performa siswa berdasarkan data.



Gambar 8. Histogram Distribusi Performance Index dan Previous Score

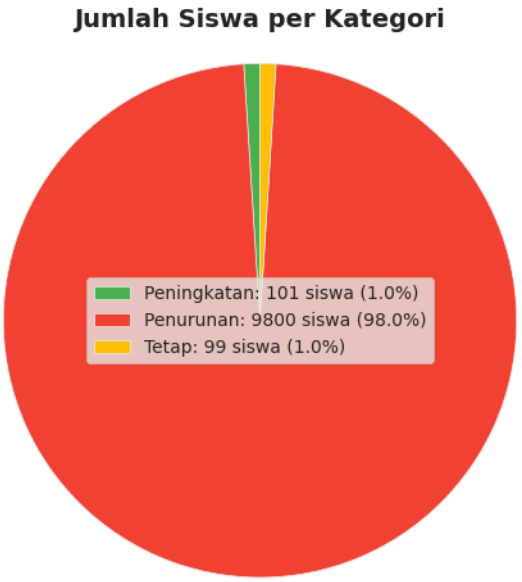
Hasil yang didapatkan adalah sebagai berikut:

- Skor terendah di ujian sebelumnya adalah 40, sedangkan skor tertinggi mencapai 99.
- Di ujian saat ini, terjadi penurunan drastis, dengan skor terendah turun hingga 10.
- Namun, berbeda dari ujian sebelumnya, kini ada yang mencapai skor sempurna (100).
- Distribusi nilai di ujian sebelumnya cukup merata di semua kategori.

- Sementara itu, di ujian sekarang, skor paling banyak berada di rentang 37-73, dan jumlahnya berkurang di bawah 37 maupun di atas 73.
- Kesimpulan: Ada perubahan signifikan dalam distribusi nilai, dengan lebih banyak siswa mendapatkan skor menengah, sementara skor ekstrem (baik rendah maupun tinggi) lebih sedikit dibandingkan sebelumnya.

4.3. Proses Pengelompokan

Setelah mendapatkan hasil dari perbandingan variabel, dapat ditentukan berapa jumlah murid yang memiliki peningkatan nilai, nilai tetap, dan penurunan nilai.



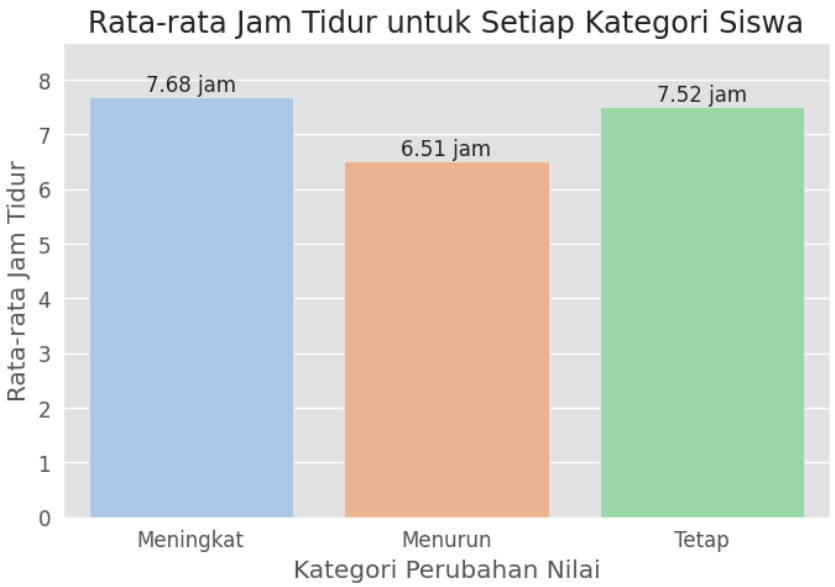
Gambar 9. Hasil Visualisasi Jumlah Siswa per Kategori

Hasil yang didapatkan adalah sebagai berikut:

- Setelah menampilkan hasil dan memvisualisasikannya dengan diagram lingkaran, terlihat jelas bahwa 101 siswa mengalami peningkatan nilai, yang setara dengan sekitar 1% dari total siswa.
- Sebaliknya, 9.800 siswa mengalami penurunan nilai, angka yang sangat besar, mencapai 98% dari keseluruhan data.
- Terakhir, 99 siswa memiliki nilai yang tetap, yang juga sekitar 1% dari total populasi.

4.4. Analisis Deskriptif berdasarkan Kategori Murid

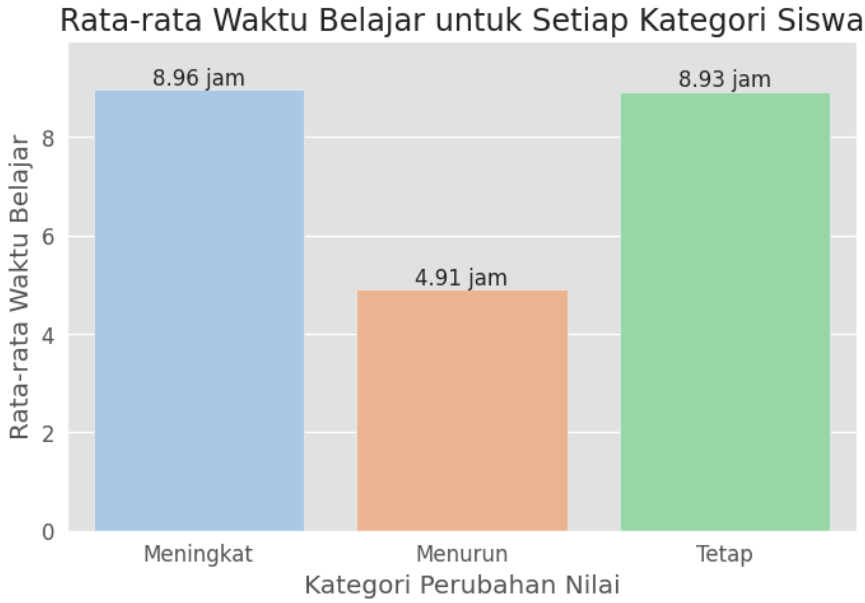
4.4.1. Analisis ‘Sleep Hours’ dari Ketiga Kelompok



Gambar 10. Hasil Visualisasi Rata-rata Jam Tidur

Siswa yang memiliki nilai meningkat atau tetap cenderung memiliki pola tidur yang lebih stabil, dengan rata-rata sekitar 7 jam per malam. Sementara itu, siswa yang nilainya menurun memiliki rata-rata jam tidur lebih rendah (sekitar 6,5 jam), yang mungkin berdampak pada performa akademik mereka.

4.4.2. Analisis ‘Hours Studied’ dari Ketiga Kelompok

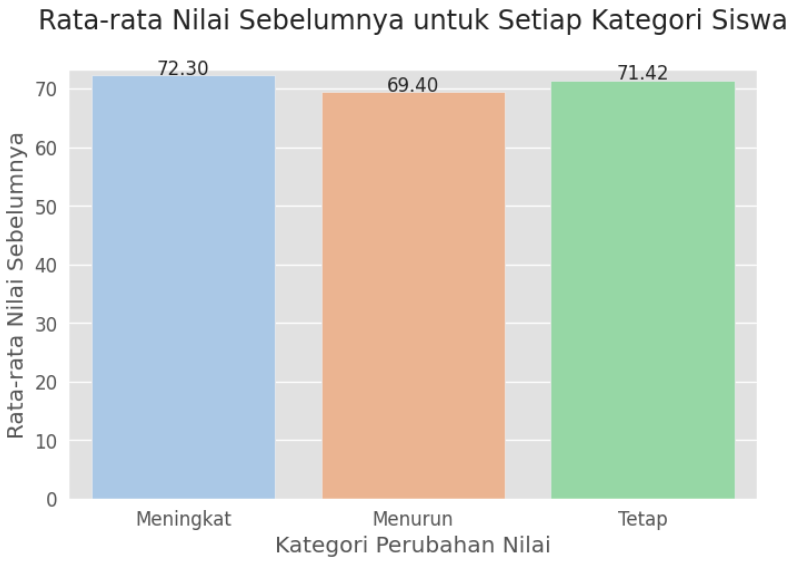


Gambar 11. Hasil Visualisasi Rata-rata Waktu Belajar

Siswa yang mempertahankan atau meningkatkan nilai mereka cenderung memiliki waktu belajar yang lebih panjang, sekitar 8-9 jam per hari. Sementara itu, siswa yang mengalami penurunan nilai memiliki waktu belajar jauh lebih sedikit, kurang dari 5 jam per hari,

yang mungkin menjadi faktor dalam penurunan performa akademik mereka

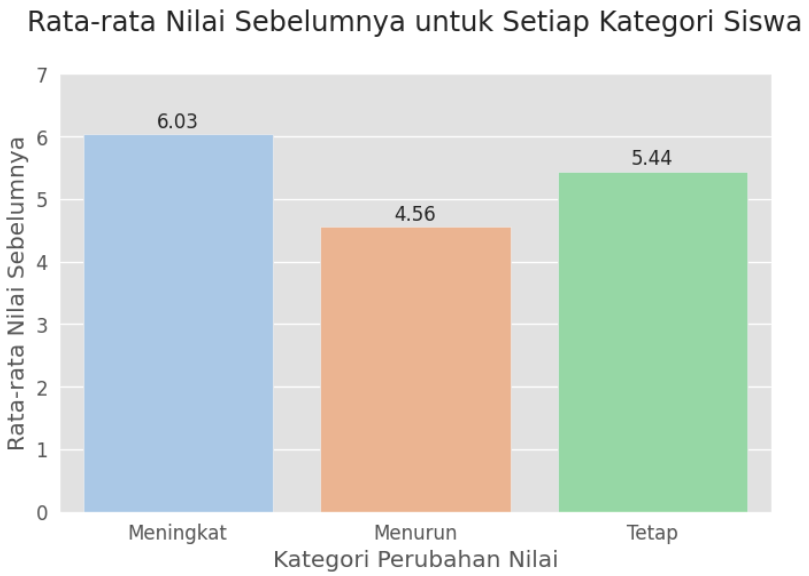
4.4.3. Analisis ‘Previous Scores’ dari Ketiga Kelompok



Gambar 12. Hasil Visualisasi Rata-rata Nilai Sebelumnya

Siswa yang mengalami penurunan nilai awalnya sudah memiliki nilai lebih rendah dibandingkan kelompok lainnya. Ini menunjukkan bahwa mereka mungkin sudah menghadapi kesulitan akademik sejak awal. Perbedaan kecil antara kategori tetap dan meningkat juga bisa menunjukkan bahwa faktor lain, seperti strategi belajar, dapat memengaruhi perubahan nilai.

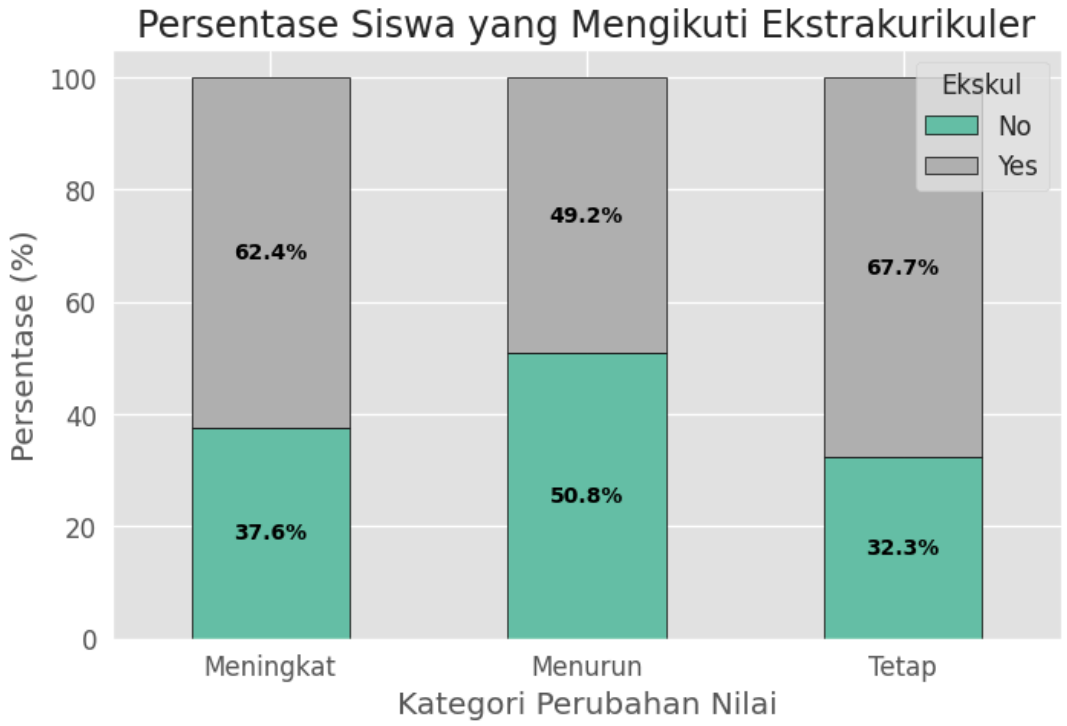
4.4.4. Analisis ‘Sample Question Paper Practiced’ dari Ketiga Kelompok



Gambar 13. Hasil Visualisasi Rata-rata Jumlah Kertas Pertanyaan yang Dilatih

Siswa yang mengerjakan lebih banyak kertas pertanyaan cenderung mengalami peningkatan nilai, sedangkan mereka yang mengerjakan lebih sedikit cenderung mengalami penurunan. Ini bisa menjadi indikasi bahwa latihan soal yang lebih banyak berkontribusi pada peningkatan performa akademik.

4.4.5. Analisis ‘Extracurricular Activities’ dari Ketiga Kelompok



Gambar 14. Hasil Visualisasi Persentase yang Mengikuti Ekstrakurikuler

Siswa yang nilai meningkat atau tetap cenderung lebih banyak mengikuti ekstrakurikuler dibandingkan siswa dengan nilai menurun. Hal ini bisa mengindikasikan bahwa ekstrakurikuler memiliki dampak positif terhadap stabilitas atau peningkatan nilai akademik. Namun, siswa dengan nilai menurun memiliki distribusi yang lebih seimbang, sehingga kemungkinan ada faktor lain yang memengaruhi penurunan nilai mereka.

4.4.6. Kesimpulan Awal dari Analisis Deskriptif

Eksplorasi awal menunjukkan bahwa mayoritas siswa mengalami penurunan nilai (98%), sementara hanya 1% yang meningkat atau tetap. Pola belajar berperan penting, di mana siswa dengan nilai meningkat atau stabil belajar lebih lama (8-9 jam/hari) dibandingkan mereka yang mengalami penurunan (4-5 jam/hari). Pola tidur juga mempengaruhi, dengan siswa yang tidur 7-7.5 jam cenderung memiliki nilai lebih baik dibandingkan yang tidur kurang dari 6.5 jam. Selain itu,

latihan soal berkontribusi terhadap performa akademik—siswa dengan nilai meningkat rata-rata mengerjakan 6 lembar, sedangkan yang menurun hanya 4.5 lembar. Ekstrakurikuler juga tampak berdampak, dengan 62-67% siswa yang nilai meningkat atau stabil aktif dalam kegiatan ini.

Kesimpulan sementara, faktor-faktor seperti jam belajar, pola tidur, latihan soal, dan ekstrakurikuler berperan dalam kinerja akademik, namun diperlukan analisis lebih lanjut untuk memahami pengaruhnya secara kuantitatif.

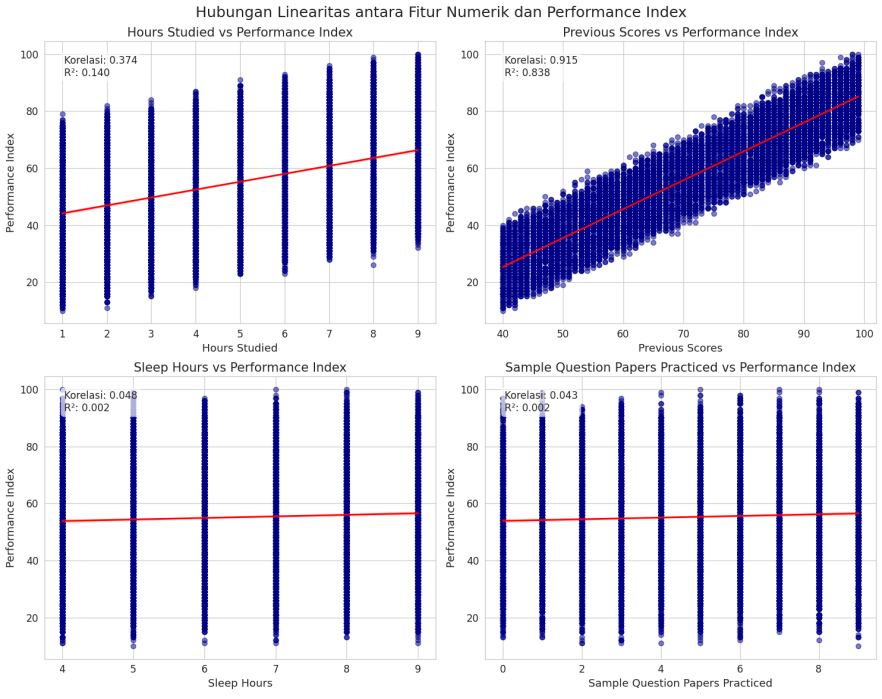
5. Eksplorasi data dan Pra-pemrosesan

Tahap eksplorasi data bertujuan untuk memahami karakteristik dataset secara lebih mendalam sebelum dilakukan pemodelan. Pada tahap ini, dilakukan pemeriksaan terhadap data yang hilang (missing values), distribusi data, serta hubungan antar variabel untuk mengidentifikasi pola atau insight awal yang berguna.

Setelah dilakukannya pemahaman terhadap data, hasil yang kami temukah bahwa data sudah bersih, dan selanjutnya adalah melihat hubungan antar variabel untuk mengidentifikasi pola atau insight awal.

Pada tahap ini kita akan mengeksplor apakah ada sebuah hubungan linearitas antara setiap fitur dengan target. Kita juga akan melihat korelasi antara fitur dengan target variabel

5.1. Analisis Hubungan Linear Fitur Numerik dengan Performance Index



Gambar 15. Hubungan Linearitas antara Fitur Numerik dan Performance Index

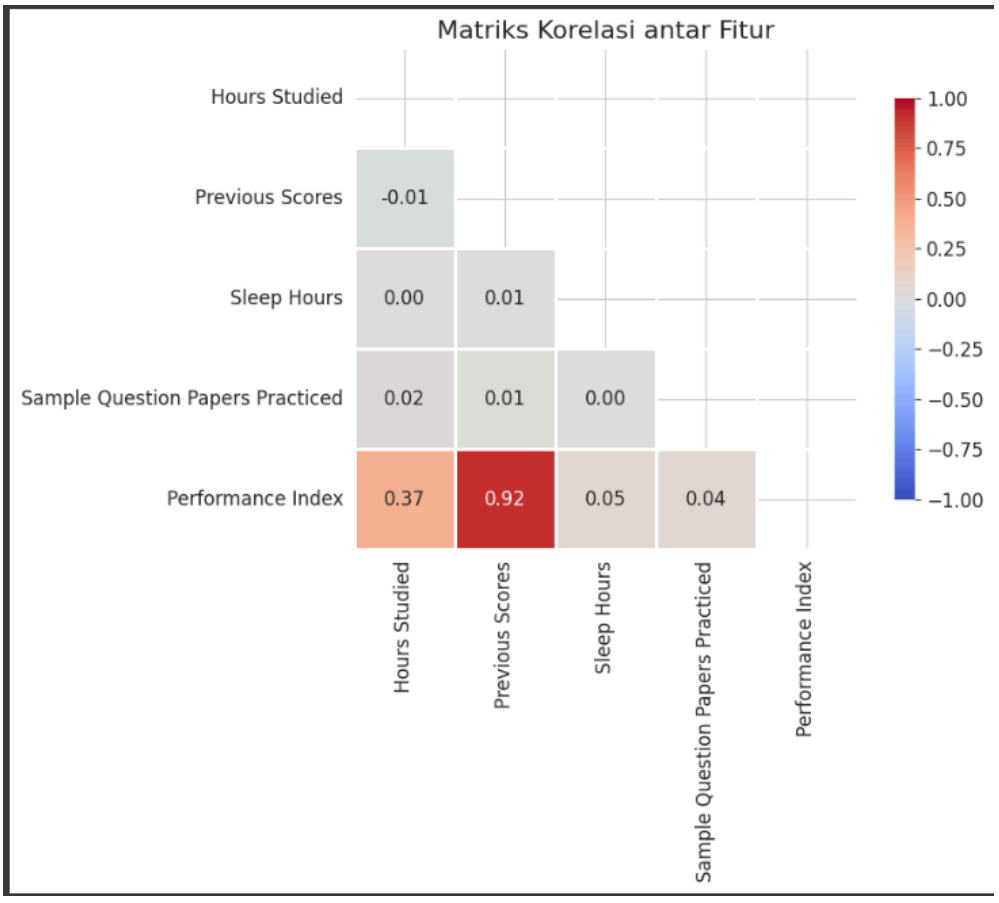
Hasil eksplorasi menunjukkan bahwa **Previous Scores** memiliki hubungan linear paling kuat dengan **Performance Index**, ditunjukkan oleh korelasi tinggi (**0.915**) dan nilai R^2 sebesar **0.838**. Ini menandakan bahwa nilai sebelumnya merupakan prediktor terbaik terhadap performa siswa.

Sementara itu, **Hours Studied** juga menunjukkan pengaruh positif dengan korelasi **0.374** dan R^2 **0.140**, meskipun tidak sekuat previous scores. Hal ini mengindikasikan bahwa peningkatan jam belajar cenderung meningkatkan performa.

Di sisi lain, **Sleep Hours** dan **Sample Question Papers Practiced** memiliki korelasi sangat lemah (masing-masing **0.048** dan **0.043**) serta R^2 yang hampir nol (**0.002**), menunjukkan bahwa keduanya tidak memiliki hubungan linear signifikan terhadap performa.

Kesimpulannya, nilai akademik sebelumnya dan jam belajar adalah faktor yang paling berkontribusi terhadap performa siswa, sedangkan jam tidur dan jumlah latihan soal tidak menunjukkan pengaruh yang berarti dalam konteks dataset ini.

5.2. Matriks Korelasi antar Fitur



Gambar 16. Matriks Korelasi antar Fitur

Matriks korelasi menunjukkan bahwa **Previous Scores** memiliki hubungan paling kuat dengan **Performance Index** (nilai korelasi **0.92**), memperkuat temuan

sebelumnya bahwa nilai akademik terdahulu merupakan indikator terbaik untuk memprediksi performa siswa. Sementara itu, **Hours Studied** juga menunjukkan korelasi positif yang cukup berarti (**0.37**), menandakan bahwa durasi belajar masih berpengaruh terhadap performa.

Menariknya, hubungan antar fitur lainnya sangat lemah atau mendekati nol, seperti antara **Sleep Hours**, **Sample Question Papers**, dan fitur lainnya. Hal ini menunjukkan bahwa masing-masing variabel mengukur aspek yang berbeda dari perilaku belajar siswa.

Meski korelasi memberikan gambaran awal yang berguna, langkah selanjutnya adalah melakukan pengujian **p-value** untuk memastikan bahwa hubungan yang ditemukan signifikan secara statistik dan bukan kebetulan semata.

5.3. Pemilihan Fitur

Berdasarkan hasil analisis korelasi dan uji signifikansi p-value, dapat disimpulkan bahwa dua fitur yang paling berpengaruh terhadap *Performance Index* adalah Previous Scores dan Hours Studied. Kedua fitur ini menunjukkan hubungan yang cukup kuat dan signifikan secara statistik dengan target variabel, menjadikannya kandidat yang relevan untuk dimasukkan ke dalam model prediksi. Dengan hanya memilih fitur-fitur yang benar-benar relevan, model yang dibangun akan menjadi lebih sederhana, efisien, dan tetap mampu menghasilkan prediksi yang akurat. Pendekatan ini juga membantu mengurangi risiko overfitting serta meningkatkan interpretabilitas model.

6. Implementasi dan Evaluasi Model

Pada tahap ini ada dua implementasi model yang akan dibangun yaitu Linear Regression dan Polynomial Regression.

6.1. Implementasi Model Linear Regression

Berdasarkan hasil analisis awal dan korelasi antar fitur, metode Regresi Linear dipilih karena terdapat hubungan linear yang cukup kuat antara beberapa variabel independen, terutama Previous Scores dan Hours Studied, dengan variabel target yaitu Performance Index. Oleh karena itu, digunakan Regresi Linear Berganda (Multiple Linear Regression) untuk membangun model prediksi.

Sebelum model diterapkan, dilakukan preprocessing data agar model dapat bekerja secara maksimal:

a. One-Hot Encoding, Variabel kategorikal diubah menjadi format numerik (biner) agar bisa digunakan dalam model regresi.

b. Standard Scaler, Fitur numerik distandarisasi agar memiliki skala yang seragam, menghindari bias pada fitur dengan nilai yang lebih besar.

Setelah preprocessing, data dibagi menjadi 80% data latih (training) dan 20% data uji (testing). Untuk menyederhanakan dan mengefisienkan proses, digunakan sebuah Pipeline yang menggabungkan seluruh tahapan dalam satu alur otomatis, terdiri dari:

- One-Hot Encoding
- Standard Scaler
- Linear Regression

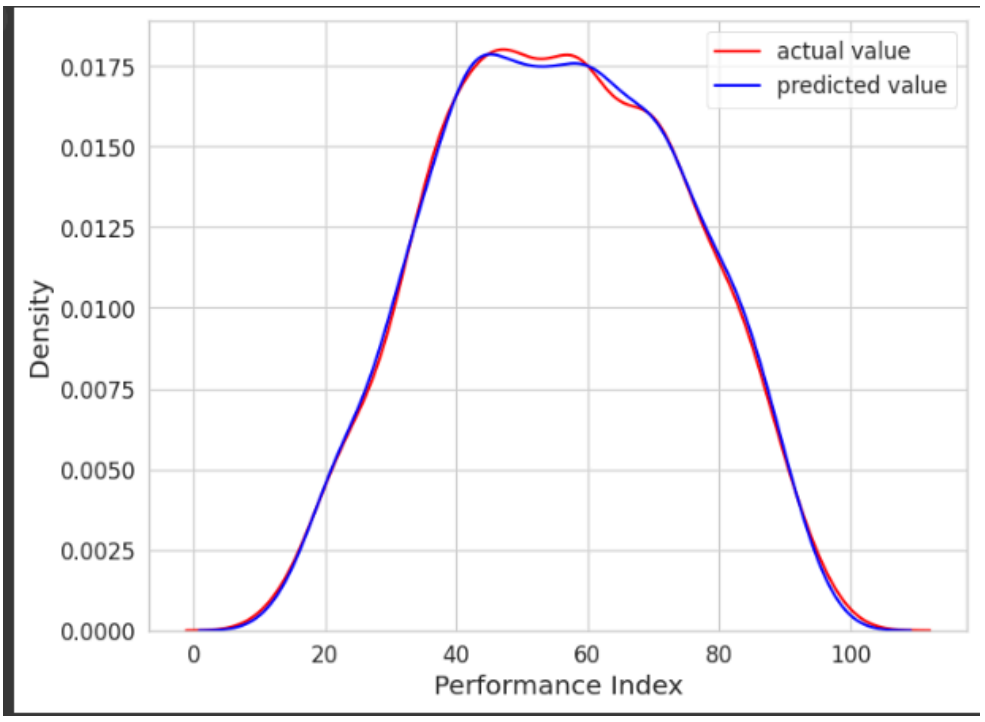
Evaluasi Model Linear Regression

Setelah model dilatih, dilakukan evaluasi dengan menggunakan data uji. Berikut adalah hasil metrik evaluasi model regresi linear:

- Mean Absolute Error (MAE): 1.7512
- Mean Squared Error (MSE): 4.8326
- Root Mean Squared Error (RMSE): 2.1983
- R-squared (R^2): 0.9860

Nilai R^2 sebesar 0.9860 menunjukkan bahwa model mampu menjelaskan sekitar 98.60% variabilitas data. Artinya, model sangat baik dalam memprediksi Performance Index berdasarkan fitur yang digunakan.

Sebagai pelengkap, divisualisasikan distribusi nilai aktual dan prediksi. Hasilnya menunjukkan bahwa kurva prediksi (garis biru) hampir menyamai kurva aktual (garis merah), yang menandakan bahwa model telah menangkap pola data dengan sangat akurat.



Gambar 17. Evaluasi Model Linear Regression

Grafik di atas menunjukkan distribusi nilai aktual (garis merah) dan prediksi (garis biru) dari model regresi. Dari grafik ini, terlihat bahwa kedua distribusi hampir berimpit, menandakan bahwa model mampu menangkap pola data dengan baik.

6.2. Implementasi Model Polynomial Regression

Untuk melihat apakah model yang lebih kompleks dapat memberikan hasil prediksi yang lebih baik, dilakukan perbandingan performa antara Regresi Linear (baseline) dan Regresi Polinomial dengan derajat 2 hingga 5 dalam memprediksi Performance Index siswa.

Alur Implementasi

a. Import Libraries & Persiapan Dataset
   Data yang telah dibersihkan digunakan untuk membangun dan menguji berbagai model regresi.
b. Fungsi Evaluasi Model
   Fungsi evaluasi dibuat untuk menghitung MAE, MSE, RMSE, dan R^2 sehingga hasil dari masing-masing model dapat dibandingkan secara objektif.
c. Train-Test Split
   Data dibagi menjadi 80% data latih dan 20% data uji.
d. Preprocessing Pipeline
   Digunakan Pipeline yang mengintegrasikan preprocessing dan pemodelan:
   StandardScaler untuk menstandarisasi fitur numerik.
   LinearRegression atau PolynomialFeatures + LinearRegression sesuai derajat yang diuji.

Evaluasi Model Polynomial Regression

Regresi Linear (Baseline)

Model pertama yang diuji adalah regresi linear standar. Pipeline terdiri dari dua tahap utama:

- Preprocessor: Melakukan standarisasi pada fitur numerik.
- Regresi Linear: Memprediksi Performance Index berdasarkan data yang telah diproses.

Hasil Evaluasi Regresi Linear:

MAE: 1.7512

MSE: 4.8326

RMSE: 2.1983

R^2: 0.9860

Nilai R^2 yang tinggi dan kesalahan yang relatif rendah menunjukkan bahwa model regresi linear sudah sangat baik dalam memprediksi.

Regresi Polinomial: Derajat 2 hingga 5

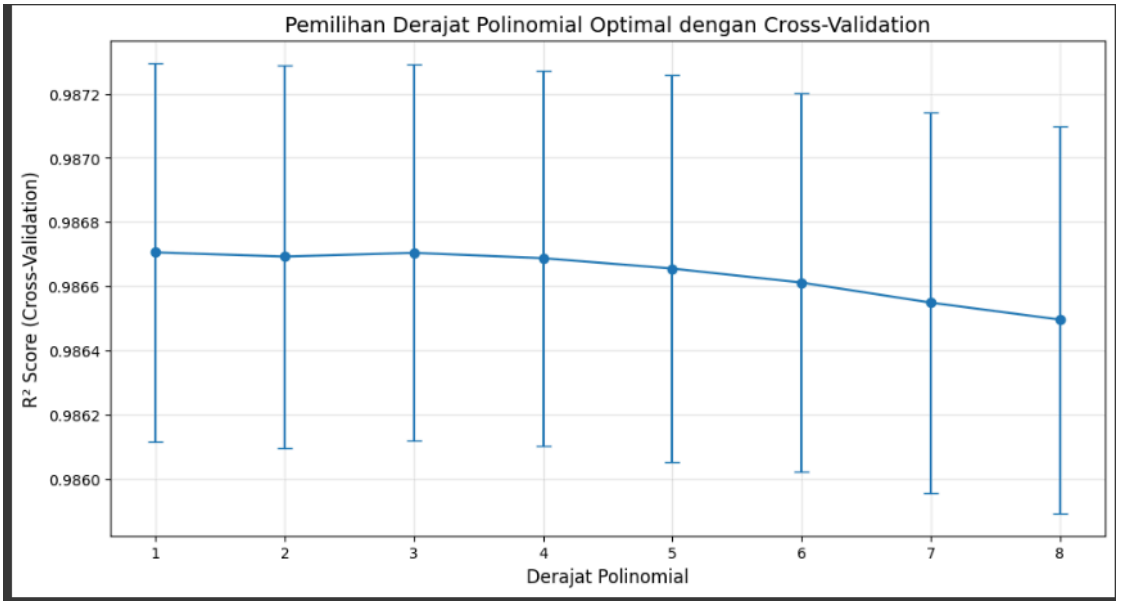
Model polinomial diuji dengan menambahkan fitur pangkat dua hingga lima dari variabel input untuk menangkap hubungan non-linear yang mungkin tidak ditangkap oleh regresi linear biasa.

Model	MAE	MSE	RMSE	R2
Linear	1.7512	4.8326	2.1983	0.9860
Derajat 2	1.7541	4.8490	2.2020	0.9859
Derajat 3	1.7551	4.8526	2.2029	0.9859
Derajat 4	1.7590	2.2069	2.2069	0.9858
Derajat 5	1.7634	4.8926	2.2119	0.9858

Interpretasi Tabel:

- Seiring bertambahnya derajat polinomial, kesalahan model (MAE, MSE, RMSE) justru sedikit meningkat, bukan menurun.
- Nilai R^2 hampir tidak berubah dan tetap sangat tinggi (> 0.985), menunjukkan bahwa semua model mampu menjelaskan variabilitas data dengan baik.
- Namun, kompleksitas model meningkat tanpa adanya perbaikan signifikan dalam performa prediksi.

Pemilihan Derajat Polinomial Optimal dengan Cross-Validation



Gambar 18. Visualisasi Derajat Polinomial Optimal dengan Cross-Validation

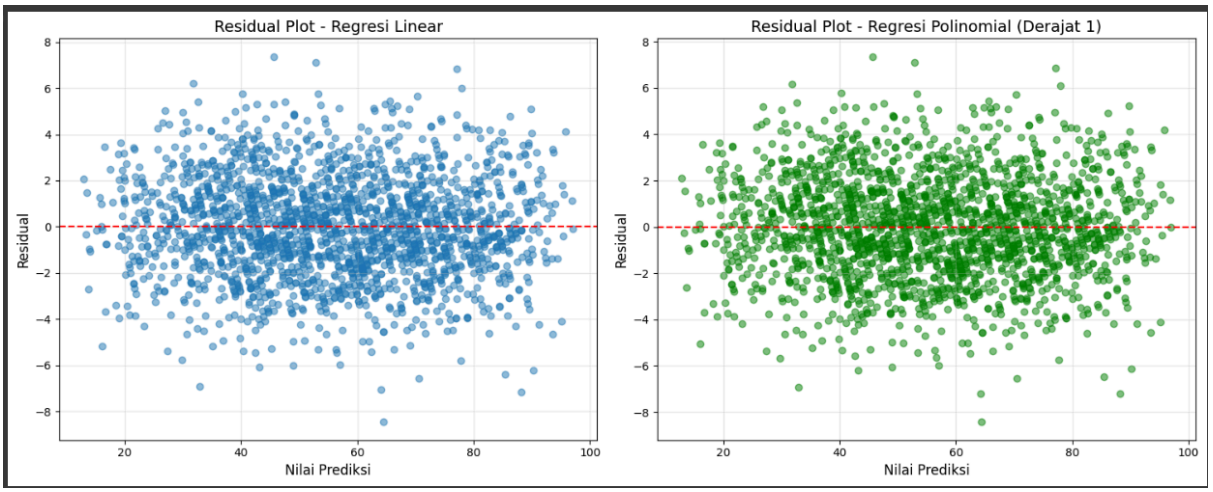
Dari visualisasi tersebut, kita bisa menyimpulkant:

- Derajat 1 (model linear) memiliki skor R^2 yang tinggi dan stabil.
- Seiring peningkatan derajat, skor R^2 cenderung menurun sedikit dan standar deviasi tetap tinggi.
- Tidak ada peningkatan signifikan pada R^2 untuk derajat lebih tinggi, bahkan performa cenderung menurun.

Derajat polinomial optimal adalah derajat 1, karena memberikan hasil terbaik dengan kompleksitas paling rendah.

6.3. Analisis Perbandingan Model Regresi Linear dan Polinomial

Visualisasi Residual Plot



Gambar 19. Visualisasi Residual Plot

Regresi Linear (kiri):
Plot residual menunjukkan penyebaran acak di sekitar garis horizontal merah (residual=0), tanpa pola khusus. Ini menunjukkan bahwa model linear cukup baik dalam menangkap pola data tanpa overfitting atau underfitting yang signifikan.

Regresi Polinomial Derajat 2 (kanan):
Meskipun model ini lebih fleksibel, distribusi residual cenderung membentuk pola tertentu, terutama pada nilai prediksi yang tinggi. Hal ini bisa mengindikasikan overfitting, yaitu model terlalu menyesuaikan data pelatihan, sehingga bisa berdampak buruk pada generalisasi.

Evaluasi Perbandingan Kinerja Model

- R^2 Regresi Linear: 0.9860
- R^2 Regresi Polinomial Derajat 2: 0.9859
- Peningkatan performa: -0.0048%

Model polinomial tidak memberikan peningkatan yang berarti dari segi akurasi (bahkan sedikit lebih buruk). Oleh karena itu, model regresi linear direkomendasikan karena lebih sederhana namun tetap akurat.

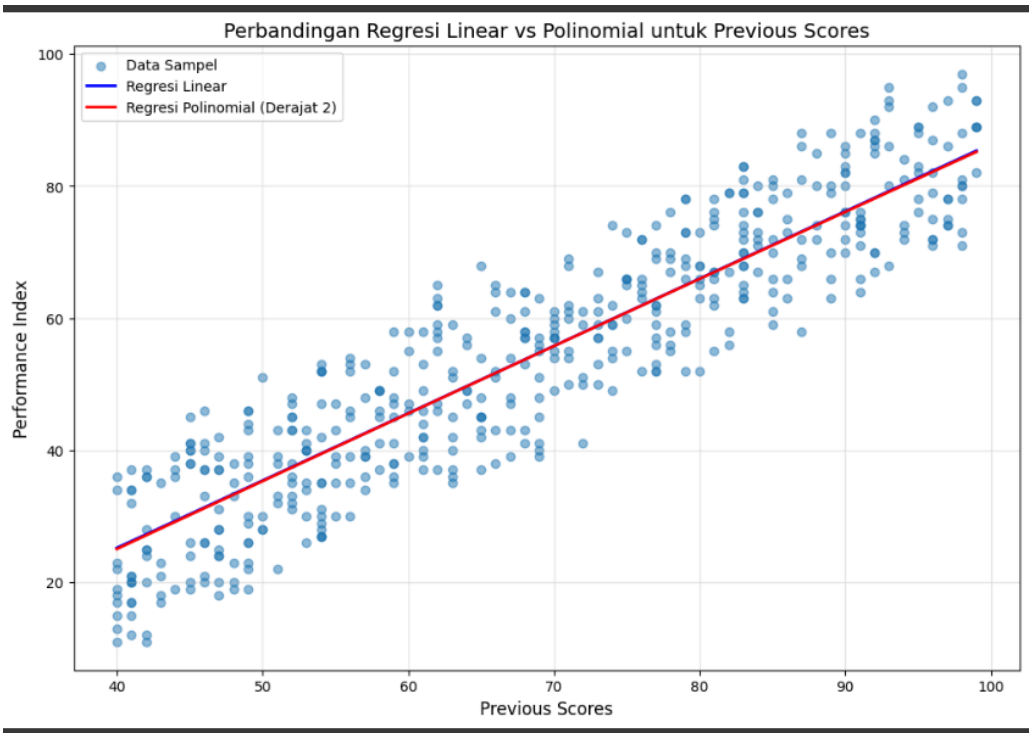
Model regresi linear memberikan hasil yang sangat baik tanpa memerlukan kompleksitas tambahan dari regresi polinomial. Oleh karena itu, **regresi linear** direkomendasikan sebagai model terbaik untuk kasus ini.

7. Analisis Hasil

Koefisien Model Regresi Linear

Fitur	Koefisien
Previous Scores	17.7924
Hours Studied	7.3928
Sample Question Papers Practiced	0.5431

- Previous Scores → Kontributor terbesar terhadap performa akademik.
- Hours Studied → Berpengaruh positif, tapi lebih kecil dibanding nilai sebelumnya.
- Sample Questions → Pengaruh sangat kecil.



Gambar 20. Perbandingan Regresi Linear vs Polinomial untuk Previous Scores

Gambar tersebut menunjukkan perbandingan antara model Regresi Linear (garis biru) dan Regresi Polinomial derajat 2 (garis merah) dalam memprediksi Performance Index berdasarkan Previous Scores. Titik-titik biru merepresentasikan data sampel asli.

Dari visualisasi ini terlihat bahwa kedua model mengikuti pola sebaran data dengan cukup baik, namun model polinomial sedikit lebih melengkung mengikuti tren naik yang sedikit non-linear pada data, terutama di ujung kiri dan kanan. Meski demikian, perbedaan antara kedua garis regresi tidak terlalu signifikan, menandakan bahwa hubungan antara Previous Scores dan Performance Index hampir linier. Artinya, model regresi linear saja sudah cukup akurat dan efisien untuk digunakan dalam konteks ini tanpa perlu menambah kompleksitas dengan model polinomial.

Kesimpulan

Berdasarkan hasil analisis regresi yang dilakukan, model yang dibangun dapat disimpulkan memiliki kinerja yang sangat baik dalam memprediksi Performance Index siswa. Hal ini ditunjukkan oleh nilai koefisien determinasi (R^2) sebesar 0.9860, yang berarti model mampu menjelaskan sekitar 98,6% variasi data target. Selain itu, nilai-nilai error seperti MAE (1.7512), MSE (4.8326), dan RMSE (2.1983) tergolong rendah, menandakan bahwa prediksi yang dihasilkan cukup akurat dan memiliki tingkat kesalahan yang kecil. Menariknya, model regresi linear sederhana menunjukkan performa yang hampir setara dengan model regresi polinomial derajat dua, tanpa peningkatan yang signifikan. Hal ini menandakan bahwa model linear sudah cukup untuk menangkap hubungan antar variabel dalam data tanpa perlu menambah kompleksitas. Dari sisi interpretasi, model juga menunjukkan hasil yang logis, di mana nilai akademik sebelumnya (Previous Scores) menjadi faktor paling dominan dalam memengaruhi performa siswa, diikuti oleh Hours Studied. Sementara itu, variabel lain seperti Sleep Hours, Sample Question Papers, dan Extracurricular Activities tidak memberikan pengaruh signifikan. Dengan demikian, dapat disimpulkan bahwa model regresi linear yang dikembangkan cukup andal dan layak digunakan untuk memprediksi kinerja akademik siswa serta dapat menjadi dasar dalam merancang strategi intervensi pendidikan yang lebih tepat sasaran.