

Tugas 3: Praktikum Machine Learning (Tugas/Praktikum Mandiri)

Shapiere Januar Rafiansyah - 0110224191

¹ Teknik Informatika, STT Terpadu Nurul Fikri, Depok

*E-mail: shapierejanuarr@gmail.com

1. Tugas Praktikum Mandiri 3

```
# import library
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score

# baca dataset
df = pd.read_csv(path + "/data/day.csv", sep=',')
df.head()
```

	instant	dteday	season	yr	mnth	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered	cnt
0	1	2011-01-01	1	0	1	0	6	0	2	0.344167	0.363625	0.805833	0.160446	331	654	985
1	2	2011-01-02	1	0	1	0	0	0	2	0.363478	0.353739	0.696087	0.248539	131	670	801
2	3	2011-01-03	1	0	1	0	1	1	1	0.196364	0.189405	0.437273	0.248309	120	1229	1349
3	4	2011-01-04	1	0	1	0	2	1	1	0.200000	0.212122	0.590435	0.160296	108	1454	1562
4	5	2011-01-05	1	0	1	0	3	1	1	0.226957	0.229270	0.436957	0.186900	82	1518	1600

Next steps: [Generate code with df](#) [New interactive sheet](#)

Bagian ini buat ngimpor semua library yang dibutuhkan.

- pandas dipakai buat ngelola dataset (baca CSV, ubah dataframe, dll).
- matplotlib.pyplot buat visualisasi hasil model.
- train_test_split dari sklearn buat bagi data jadi training dan testing.
- LinearRegression buat bikin model regresinya.
- mean_absolute_error, mean_squared_error, dan r2_score buat evaluasi hasil prediksi model.

Dan buat baca dataset day.csv dan nampilin 5 baris pertama. Tujuannya buat ngecek apakah data udah berhasil dimuat dan kolomnya sesuai.

Output:

Tampil 5 baris awal dari dataset, dengan kolom seperti: instant, dteday, season, yr, mnth, holiday, weekday, workingday, weathersit, temp, hum, windspeed, casual, registered, cnt. Totalnya ada 731 baris dan 16 kolom.

```
# ambil kolom yang dibutuhkan
# Y = cnt (jumlah peminjaman)
# X = suhu (temp), kelembapan (hum), kecepatan angin (windspeed)
X = df[['temp', 'hum', 'windspeed']]
y = df['cnt']
```

Di sini X adalah variabel independen (faktor yang mempengaruhi), yaitu suhu, kelembapan, dan kecepatan angin. y adalah variabel dependen (target) yaitu jumlah peminjaman sepeda (cnt). Karena ada lebih dari satu variabel X, model ini jadi **Multiple Linear Regression**.

```
# bagi data jadi data training dan testing
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

print("Jumlah data training:", len(X_train))
print("Jumlah data testing:", len(X_test))
```

➡ Jumlah data training: 584
Jumlah data testing: 147

Dataset dibagi jadi **80% data training** dan **20% data testing**. `random_state=42` biar hasilnya konsisten setiap dijalankan. Data training dipakai buat ngelatih model, sedangkan data testing buat nguji seberapa akurat modelnya.

Output:

- Jumlah data training: 584
- Jumlah data testing: 147

```
# buat model regresi linear
model = LinearRegression()
model.fit(X_train, y_train)

# prediksi data testing
y_pred = model.predict(X_test)
```

`LinearRegression()` bikin objek model regresi linear. `fit()` buat ngelatih model berdasarkan data training (`X_train`, `y_train`). `predict()` dipakai buat menghasilkan prediksi jumlah sepeda (`y_pred`) berdasarkan data testing.

```
# evaluasi model
mae = mean_absolute_error(y_test, y_pred)
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)

print("Mean Absolute Error (MAE):", mae)
print("Mean Squared Error (MSE):", mse)
print("R2 Score:", r2)
```

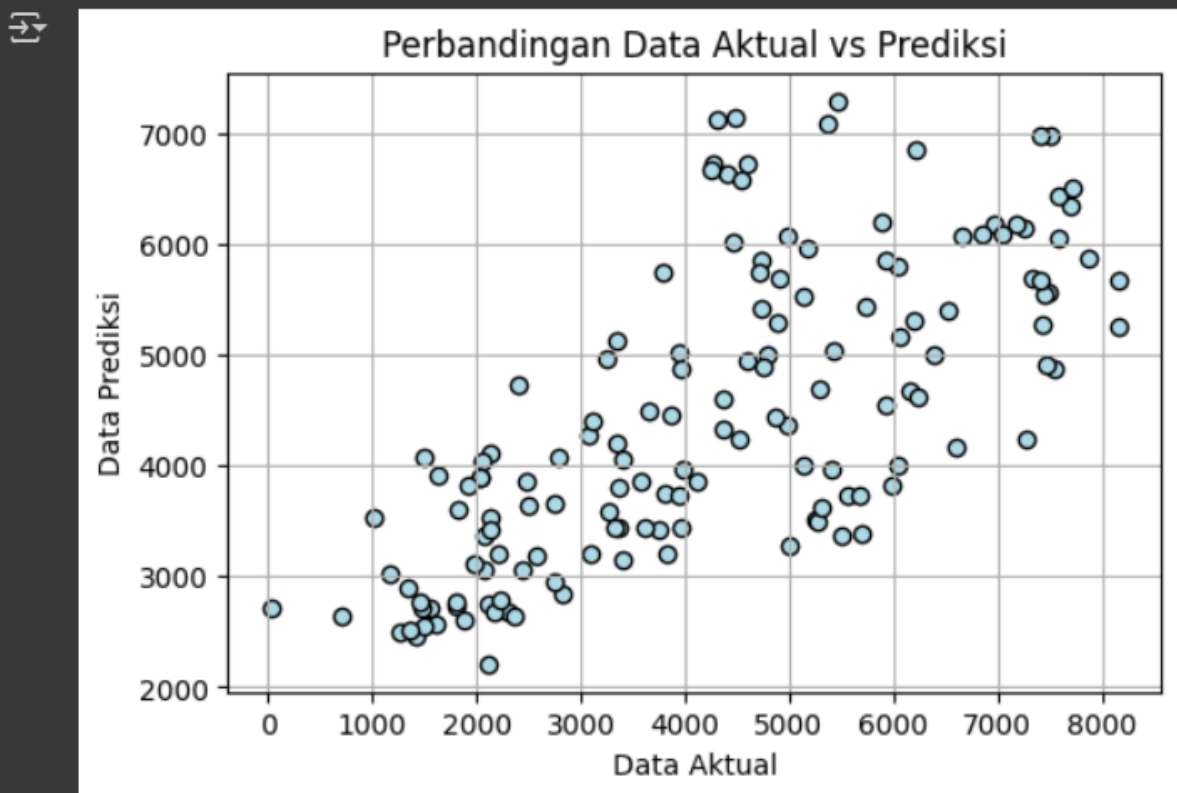
⇒ Mean Absolute Error (MAE): 1196.956055531855
Mean Squared Error (MSE): 2020722.2320442528
R² Score: 0.49606445111931263

Penjelasan:

Bagian ini buat ngukur performa model:

- **MAE** → rata-rata selisih antara prediksi dan data aktual.
- **MSE** → selisih kuadrat (semakin kecil semakin bagus).
- **R² Score** → seberapa baik model menjelaskan data (semakin mendekati 1 semakin bagus).

```
# visualisasi hasil prediksi
plt.figure(figsize=(6,4))
plt.scatter(y_test, y_pred, color='lightblue', edgecolor='black')
plt.xlabel('Data Aktual')
plt.ylabel('Data Prediksi')
plt.title('Perbandingan Data Aktual vs Prediksi')
plt.grid(True)
plt.show()
```



Scatter plot ini nampilin hubungan antara nilai aktual (y_{test}) dan nilai prediksi (y_{pred}). Kalau titik-titiknya mendekati garis diagonal, berarti model punya akurasi bagus.

Output:

Grafik sebaran hijau dengan pola yang cenderung naik — artinya prediksi model masih lumayan sesuai sama pola data aslinya.

2. Kesimpulan

Dari hasil implementasi algoritma **Multiple Linear Regression** pada dataset *day.csv*, bisa disimpulkan kalau model ini cukup efektif buat lihat hubungan antara jumlah peminjam sepeda harian (*cnt*) dengan faktor-faktor kayak suhu, kelembapan, dan kecepatan angin. Hasilnya nunjukin kalau tiap variabel punya pengaruh yang beda-beda terhadap jumlah pengguna. Tapi secara kritis, model ini masih bisa dikembangkan lagi — misalnya dengan eksplorasi fitur tambahan atau optimasi data biar hasil prediksinya lebih akurat dan realistis.

3. GitHub

<https://github.com/Shapiere/TugasMandiri-Praktikum-MachineLearning.git>