Tugas 3: Membuat Model Prediksi Jumlah Penyewaan Sepeda (Bike Sharing Dataset)

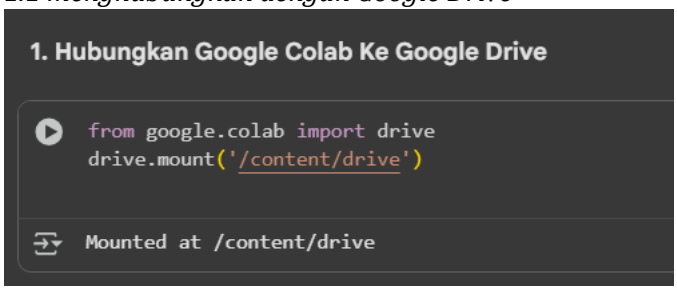
Ahmad Azka Ridha - 0110222062 [1]

[1] Teknik Informatika, STT Terpadu Nurul Fikri, Depok

*E-mail: ahma22062ti@student.nurulfikri.ac.id

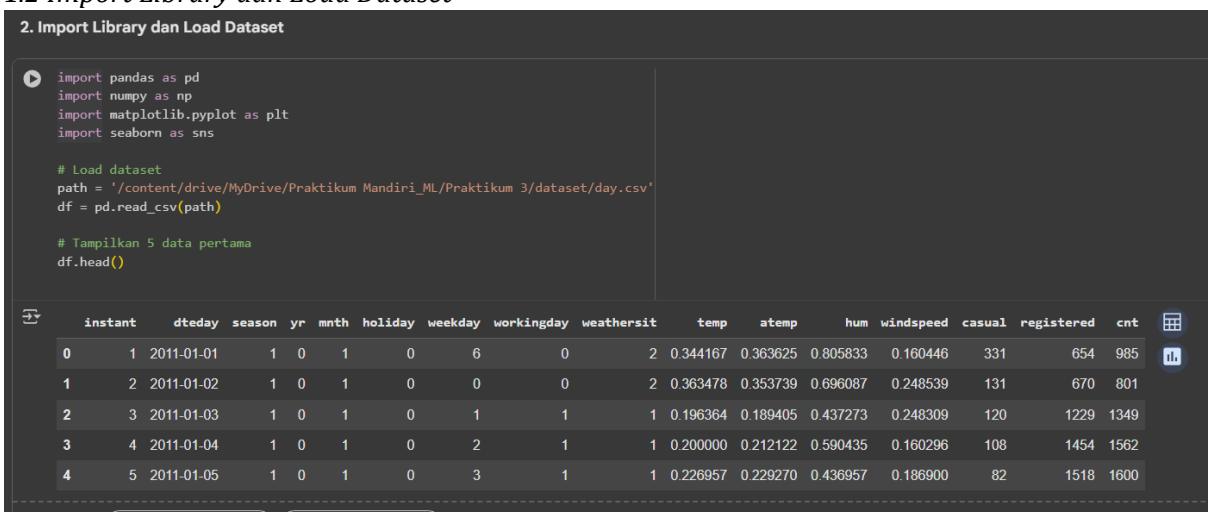
1. Praktikum Mandiri

1.1 menghubungkan dengan Google Drive



Langkah ini digunakan agar Colab dapat mengakses file yang tersimpan di Google Drive. Dataset yang telah diunduh dari Kaggle harus disimpan di dalam Drive.

1.2 Import Library dan Load Dataset



	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

- Library seperti **pandas** dan **numpy** digunakan untuk membaca serta memproses data.
- **matplotlib** dan **seaborn** digunakan untuk visualisasi grafik.
- Dataset “day.csv” dibaca dan ditampilkan 5 baris pertama untuk melihat struktur datanya.

1.3 Explorasi Awal Dataset

3. Explorasi Awal Dataset

[3] ✓ 0s

```
# Informasi data
df.info()

# Cek missing value
print(df.isnull().sum())

# Statistik ringkas
df.describe()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 731 entries, 0 to 730
Data columns (total 16 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   instant     731 non-null    int64
 1   dteday      731 non-null    object
 2   season      731 non-null    int64
 3   yr          731 non-null    int64
 4   mnth        731 non-null    int64
 5   holiday     731 non-null    int64
 6   weekday     731 non-null    int64
 7   workingday  731 non-null    int64
 8   weathersit  731 non-null    int64
 9   temp        731 non-null    float64
 10  atemp       731 non-null    float64
 11  hum         731 non-null    float64
 12  windspeed   731 non-null    float64
 13  casual      731 non-null    int64
 14  registered  731 non-null    int64
 15  cnt         731 non-null    int64
```

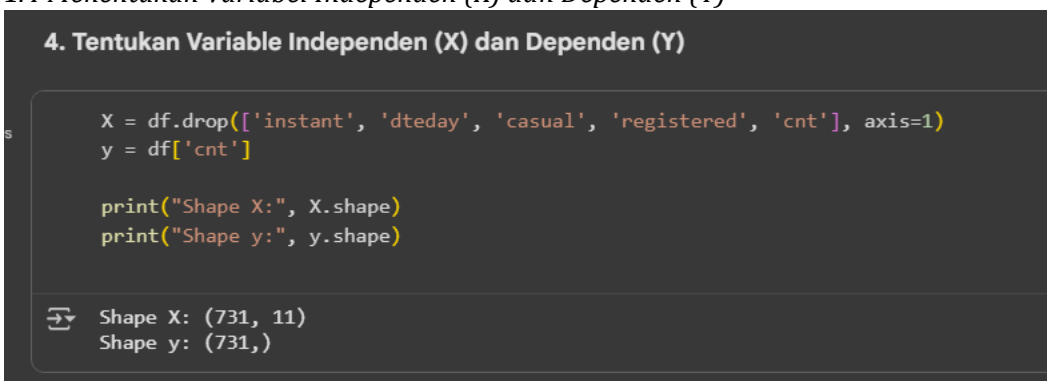
```
memory usage: 91.5+ KB
instant       0
dteday        0
season        0
yr            0
mnth          0
holiday       0
weekday       0
workingday    0
weathersit    0
temp          0
atemp         0
hum           0
windspeed     0
casual        0
registered    0
cnt           0
dtype: int64
```

	instant	season	yr	mnth	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered	cnt
count	731.000000	731.000000	731.000000	731.000000	731.000000	731.000000	731.000000	731.000000	731.000000	731.000000	731.000000	731.000000	731.000000	731.000000	731.000000
mean	366.000000	2.496580	0.500684	6.519836	0.028728	2.997264	0.683995	1.395349	0.495385	0.474354	0.627894	0.190486	848.176471	3656.172367	4504.348837
std	211.165812	1.110807	0.500342	3.451913	0.167155	2.004787	0.465233	0.544894	0.183051	0.162961	0.142429	0.077498	686.622488	1560.256377	1937.211452
min	1.000000	1.000000	0.000000	1.000000	0.000000	0.000000	0.000000	1.000000	0.059130	0.079070	0.000000	0.022392	2.000000	20.000000	22.000000
25%	183.500000	2.000000	0.000000	4.000000	0.000000	1.000000	0.000000	1.000000	0.337083	0.337842	0.520000	0.134950	315.500000	2497.000000	3152.000000
50%	366.000000	3.000000	1.000000	7.000000	0.000000	3.000000	1.000000	1.000000	0.498333	0.486733	0.626667	0.180975	713.000000	3662.000000	4548.000000
75%	548.500000	3.000000	1.000000	10.000000	0.000000	5.000000	1.000000	2.000000	0.655417	0.608602	0.730209	0.233214	1096.000000	4776.500000	5956.000000
max	731.000000	4.000000	1.000000	12.000000	1.000000	6.000000	1.000000	3.000000	0.861667	0.840896	0.972500	0.507463	3410.000000	6946.000000	8714.000000

- df.info menampilkan tipe data tiap kolom (numerik, kategorikal, dsb).
- df.isnull.sum memeriksa apakah terdapat nilai kosong (missing value).
- df.describe menampilkan statistik deskriptif seperti rata-rata, nilai minimum, dan maksimum.

Hasilnya menunjukkan bahwa dataset bersih dan tidak memiliki nilai kosong.

1.4 Menentukan Variabel Independen (X) dan Dependen (Y)



- **Variabel dependen (Y)** adalah cnt, yaitu jumlah total penyewaan sepeda.
- **Variabel independen (X)** adalah kolom-kolom lain yang memengaruhi cnt, **kecuali:**
  - instant → hanya nomor urut
  - dteday → tanggal
  - casual dan registered → sudah merupakan komponen dari cnt

Jadivariabel X yang digunakan adalah: season, yr, mnth, holiday, weekday, workingday, weathersit, temp, atemp, hum, windspeed

1.5 Membagi Data Menjadi Training dan Testing

5. Bagi Data untuk Testing dan Training

```
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

print("Training data:", X_train.shape)
print("Testing data:", X_test.shape)
```

```
Training data: (584, 11)
Testing data: (147, 11)
```

- Dataset dibagi menjadi dua bagian:
    - **Data Training (80%)** → untuk melatih model.
    - **Data Testing (20%)** → untuk menguji model.

- Parameter `random_state=42` memastikan pembagian data selalu sama setiap kali dijalankan (reproducible).

1.6 Membuat Model Prediksi (Linear Regression)

6. Buat Model Prediksi

```
from sklearn.linear_model import LinearRegression

# Inisialisasi dan latih model
model = LinearRegression()
model.fit(X_train, y_train)

# Prediksi
y_pred = model.predict(X_test)
```

- Digunakan algoritma **Linear Regression**, yaitu model sederhana yang memprediksi hubungan linier antara variabel input dan output.

- `model.fit()` → melatih model dengan data training.

- `model.predict()` → menghasilkan prediksi pada data testing.

1.7 Evaluasi Model

7. Evaluasi Model

```
from sklearn.metrics import r2_score, mean_squared_error
import numpy as np

r2 = r2_score(y_test, y_pred)
rmse = np.sqrt(mean_squared_error(y_test, y_pred))

print("R² Score:", r2)
print("RMSE:", rmse)
```

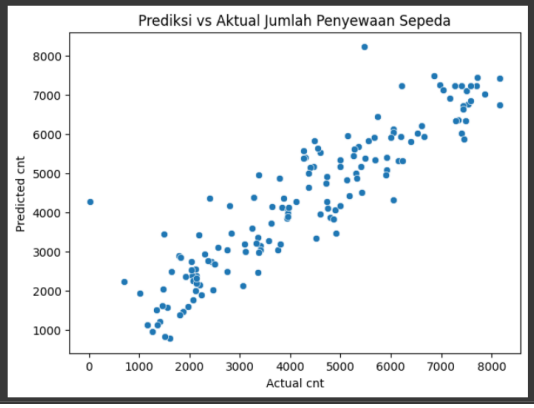
```
R² Score: 0.8276670090367212
RMSE: 831.2851545662686
```

- **R² Score** menunjukkan seberapa besar variasi data target dapat dijelaskan oleh model. Nilai mendekati **1** berarti model bagus.

- **RMSE (Root Mean Square Error)** menunjukkan rata-rata error prediksi. Semakin kecil nilai RMSE, semakin akurat model.

1.8 Visualisasi Hasil

```
plt.figure(figsize=(7,5))
sns.scatterplot(x=y_test, y=y_pred)
plt.xlabel("Actual cnt")
plt.ylabel("Predicted cnt")
plt.title("Prediksi vs Aktual Jumlah Penyewaan Sepeda")
plt.show()
```



Grafik ini memperlihatkan perbandingan antara nilai aktual (y_test) dan nilai prediksi (y_pred). Semakin dekat titik-titik pada garis diagonal, semakin akurat model dalam memprediksi data.