Tugas 1: Praktikum Mandiri 3

Aisyah Hanani - 0110222286

Teknik Informatika, STT Terpadu Nurul Fikri, Depok

*E-mail: aisyahhanani82@gmail.com

Abstract. Praktikum ini bertujuan untuk memprediksi jumlah penyewaan sepeda (*bike sharing count*) berdasarkan berbagai faktor cuaca dan waktu menggunakan model *Linear Regression*. Data yang digunakan berasal dari dataset *Bike Sharing System* yang mencakup variabel seperti suhu, kelembapan, kecepatan angin, kondisi cuaca, hari kerja, bulan, dan musim. Proses analisis meliputi tahap pra-pemrosesan data, pembagian data menjadi data latih dan uji, penerapan *one-hot encoding* untuk variabel kategorikal, serta *standardization* pada variabel numerik.

1. Menyambungkan Google Colab dengan Drive

```
from google.colab import drive
drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

Baris ini mengimpor modul drive dari library google colab, yang berisi fungsi untuk mengakses Google drive dari colab, kemudian memasang google drive ke direktori di Colab.

2. Import Library dan Membaca Data

2.1 Import data

Pandas digunakan untuk membaca dan mengolah data dalam bentuk tabel
Numpy digunakan untuk operasi numerik (angka)
Train_test_split membagi data menjadi data latih (data train) dan data uji (data test)
Onehotencoder mengubah data kategorikal menjadi numerik (biner)
StandardScaler menstandarkan data numerik agara memiliki skala yang seragan
Linearregression model regresi linear dari scikitlearn
Mean_absolute_error_mean_squared_error_r2_score metrik evalusi untuk mengukur permorma model

2.2 Membaca Data

Membaca file csv yang ada di folder drive dan menyimpannya dalam variabel df sebagai DataFrame

3. Memisahkan target dan fitur

```
# Variabel dependen (target)
y = df['cnt']

# Variabel independen
X = df[['season','yr','mnth','holiday','weekday',
        'workingday','weathersit','temp','atemp','hum','windspeed']]
```

Memisahkan variabel (Y) dan fitur independen (X) . cnt = jumlah total peminjaman sepeda (gabungan antara penyewa casual dan registered), kemudian model akan mencoba memprediksi nilai cnt berdasarkan faktor faktor lain seperti cuaca, musim, dan hari.

4. **Split data train dan data test**

```
# Split data untuk training & testing
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, random_state=42
)
```

Membagi dataset menjadi data latih (training) dan data uji (testing). Data training (latih) digunakan untuk melatih model, data testing digunakan untuk mengevaluasi seberapa baik model memprediksi data baru yang belum pernah dilihat sebelumnya.

5. **Prepocessing Data**

```
cat_features = ['season','mnth','weekday','weathersit']
num_features = ['yr','holiday','workingday','temp','atemp','hum','windspeed']

# One-hot encoding untuk fitur kategori
encoder = OneHotEncoder(drop='first')
X_train_cat = encoder.fit_transform(X_train[cat_features])
X_test_cat = encoder.transform(X_test[cat_features])

# Gabungkan dengan fitur numerik
X_train_num = X_train[num_features].values
X_test_num = X_test[num_features].values

X_train_proc = np.hstack([X_train_num, X_train_cat.toarray()])
X_test_proc = np.hstack([X_test_num, X_test_cat.toarray()])
```

Ini adalah langkah prepocessing data sebelum model Linear Regresion dilatih, yaitu mengubah data kategorikal menjadi bentuk numerik dan menggabungkannya dengan data numerik agar model bisa memproses semuanya.

6. **Normalisasi/standarisasi data**

```
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train_proc)
X_test_scaled = scaler.transform(X_test_proc)
```

Digunakan untuk menstandarkan fitur numerik supaya setiap kolom memiliki rata-rata (mean) = 0, standar deviasi =1,

7. **Melatih model linear Regression**

```
model_lr = LinearRegression()
model_lr.fit(X_train_scaled, y_train)
```

Langkah utama melatih model linear Regression menggunakan data training yang sudah diproses dan distandarisasi.

8. **Tahap evaluasi model Linear Regression**

```
y_pred = model_lr.predict(X_test_scaled)

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

print(f"MAE  : {mae:.2f}")
print(f"RMSE : {rmse:.2f}")
print(f"R²   : {r2:.3f}")
```

```
MAE  : 583.02
RMSE : 796.46
R²   : 0.842
```

Untuk mengukur seberapa baik model memprediksi jumlah penyewa sepeda (cnt) berdasarkan data uji.

8.1 Prediksi data uji

Menggunakam model yang sudah dilatih untuk memprediksi nilai (cnt) dari data uji (x_tset_scaled), hasilnya (y_pred) berisi niali prediksi jumlah sepeda yang disewa untuk setiap baris di data uji.

8.2 Menghitunh metrik evaluasi

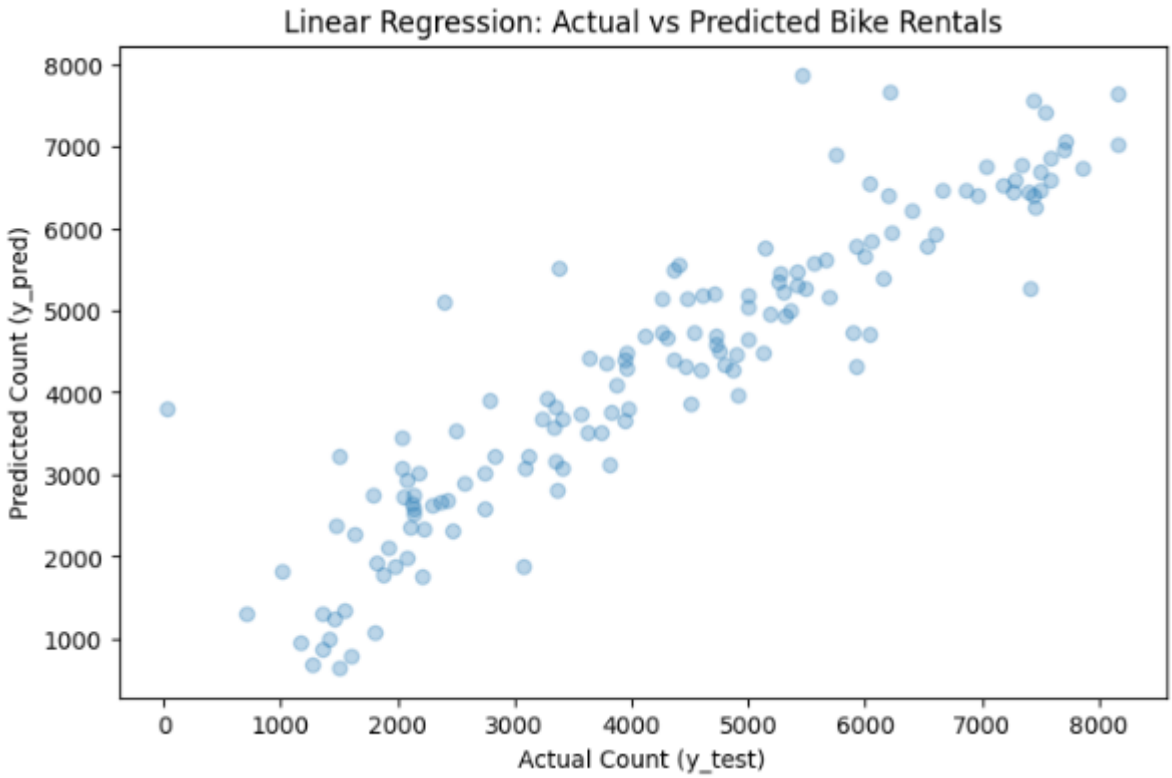
a. Mae (mean Absolute Error)
   Mengukur rata rata selisih absolut antara nilai aktual dan prediksi. Nilai semakin kecil maka model semakin akurat. Rata rata kesalahan prediksi model adalah sekitar 583 penyewaan sepeda per hari.

b. RMSE (Root Mean Squared Error)
   Mengukur akar dari rata rata kuadrat selisih antara prediksi dan nilaai aktual. Kesalahan prediksi yang mempertimbangkan outlier adalah sekitar 796 sepeda per hari.

c. R-squared / koefisien determinasi
   Menunjukkan seberapa besar variasi data target (cnt) yang bisa dijelaskan oleh model. Nilai mendekati 1 berarti model sangat baik, nilai mendekati 0 berarti model kurang menjelaskan variasi data. Model menjelaskan 84,2% variasi jumlah penyewaan sepeda berdasarkan cuaca, musim, dan hari kerja.

9. Visualisasi hubungan antara nilai aktual dan nilai prediksi

```
plt.figure(figsize=(8,5))
plt.scatter(y_test, y_pred, alpha=0.3)
plt.xlabel("Actual Count (y_test)")
plt.ylabel("Predicted Count (y_pred)")
plt.title("Linear Regression: Actual vs Predicted Bike Rentals")
plt.show()
```



Dari scatter plot yang dihasilkan terlihat bahwa:

a. Titik-titiknya tersebar cukup rapat di sekitar garis diagonal imajiner (y = x), yang artinya model *Linear Regression* sudah memprediksi jumlah penyewaan sepeda dengan cukup baik.
b. Tidak terlihat ada pola penyimpangan besar. Penyebaran relatif konsisten dari nilai kecil hingga besar.
c. Namun, ada beberapa titik yang agak jauh dari pola utama (terutama di nilai prediksi tinggi), menandakan adanya sedikit underprediction atau overprediction di kasus tertentu.

10. Melihat fitur paling berpengaruh

```
coef = pd.Series(model_lr.coef_, name="Coefficient")

# Ambil nama fitur setelah one-hot encoding
encoded_features = encoder.get_feature_names_out(cat_features)
all_features = num_features + list(encoded_features)

coef_df = pd.DataFrame({'Feature': all_features, 'Coefficient': coef})
coef_df.sort_values('Coefficient', ascending=False).head(10)
```

Table 1. Koefisien Fitur dengan Pengaruh Positif Terbesar terhadap Jumlah Penyewaan Sepeda

Feature	Produk Pertanian
yr (990.34)	Sangat tinggi → menandakan bahwa penyewaan meningkat dari tahun ke tahun. Artinya, jumlah pengguna sepeda bertambah signifikan di tahun berikutnya (2012 dibanding 2011).
season_4 (707.94)	Musim ke-4 (kemungkinan *fall/autumn*) meningkatkan jumlah penyewaan sepeda, artinya musim ini paling populer untuk bersepeda.
temp (686.30)	Semakin hangat suhu udara, semakin banyak orang menyewa sepeda. Cuaca dingin biasanya menurunkan minat bersepeda.
season_2 (416.77)	Musim ke-2 (mungkin *spring*) juga meningkatkan penyewaan, tapi tidak sebanyak musim ke-4.
season_3 (358.63)	Musim ke-3 (mungkin *summer*) juga berpengaruh positif, berarti cuaca cerah dan liburan bisa mendorong penyewaan.
mnth_9 (237.18)	Bulan ke-9 (September) punya tren penyewaan tinggi. Bisa jadi karena cuaca masih nyaman dan aktivitas luar ruang meningkat.
atemp (195.47)	Suhu yang terasa (*apparent temperature*) juga punya pengaruh positif, tapi lebih kecil dibanding suhu aktual.
weekday_6 (173.29)	Hari ke-6 (kemungkinan Sabtu), akhir pekan cenderung meningkatkan penyewaan karena banyak orang rekreasi.
mnth_5 (157.98)	Bulan Mei juga termasuk periode penyewaan tinggi. Cuaca cenderung stabil dan cocok untuk aktivitas luar ruangan.
mnth_3 (146.10)	Bulan Maret juga mulai meningkat (mungkin karena transisi dari musim dingin ke musim hangat).

Tabel ini menunjukkan sepuluh fitur dengan nilai koefisien regresi linear tertinggi yang berpengaruh positif terhadap jumlah penyewaan sepeda (*cnt*). Nilai positif menunjukkan bahwa peningkatan pada fitur tersebut cenderung diikuti oleh peningkatan jumlah penyewaan sepeda.

LINK GITHUB : https://github.com/aisyahhana/Machine-Learning.git