Tugas 3: Praktikum & Praktikum Mandiri 3

Pandu Linggar Kumara - 0110221277, Link GitHub - https://github.com/PanduLgg/M_Learning.git

¹ Teknik Informatika, STT Terpadu Nurul Fikri, Depok

*E-mail: pandulinggar1@gmail.com

Abstract. Eksperimen dimulai dari studi kasus sederhana, yaitu prediksi berat badan berdasarkan tinggi dan umur, kemudian dilanjutkan dengan penerapan pada dataset nyata yaitu Bike Sharing Dataset dari Kaggle. Analisis dilakukan menggunakan metode Ordinary Least Squares (OLS) dengan bantuan pustaka statsmodels pada Python untuk memodelkan hubungan antara variabel dependen (cnt, total penyewaan sepeda) dan beberapa variabel independen seperti suhu, kelembapan, musim, serta kecepatan angin. Proses pemodelan mencakup tahap persiapan data, pemilihan fitur, pembuatan model, serta evaluasi menggunakan metrik statistik seperti koefisien determinasi (R²), koefisien regresi, dan nilai signifikansi (p-value). Hasil analisis menunjukkan bahwa regresi linear berganda mampu menjelaskan variasi jumlah penyewaan sepeda berdasarkan faktor lingkungan dan musiman secara efektif.

1. Connecting Google Colab & Drive

1.1 Menghubungkan lingkungan Google Colab dengan akun Google Drive Sel ini berfungsi untuk menghubungkan lingkungan Google Colab dengan akun Google Drive



Gambar 1.1. Proses ini hanya perlu dilakukan satu kali per sesi.

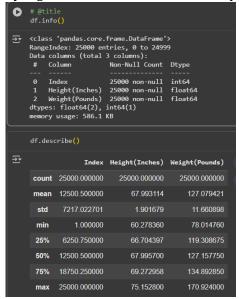
1.2 Memanggil Data set dari Gdrive dan Membaca file .CSV menggunakan Pandas Sel ini menggunakan library Pandas untuk membaca file data, yang diinginkan

	<pre>import pandas as pd df = pd.read_csv('/content/drive/MyDrive/praktikum_ml/praktikum@3/data/socr.csv') df.head()</pre>										
₹		Index	Height(Inches)	Weight(Pounds)	⊞						
	0		65.78331	112.9925	1						
			71.51521	136.4873							
	2		69.39874	153.0269							
	3		68.21660	142.3354							
	4		67.78781	144.2971							
Lanç	gkah	berikut	nya: Buat kode o	dengan df Nev	v interactive sheet						

Gambar 1.2. Proses ini hanya perlu dilakukan satu kali per sesi.

1.3 Mencari informasi data yang ada pada file

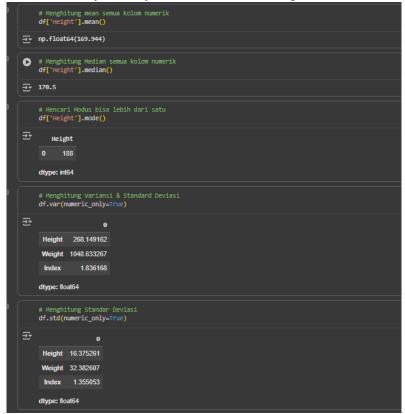
Sel ini menampilkan informasi yang ada di dalam file dari mulai tipe data nama kolom, dsb.



Gambar 1.3. Mencari info data pada file

1.4 Data Pre-processing

kita membuat DataFrame baru bernama df1 yang berisi data tinggi dan berat badan yang telah dikonversi dari satuan inci dan pon menjadi sentimeter dan kilogram



Gambar 1.4. Tabel tinggi dan berat dalam satuan metrik yang lebih mudah dibaca

1.5 Membagi dataset untuk Training dan Test

Sel ini membagi variabel dependen dan independen dibagi menjadi 2 yaitu train dan test

```
from sklearn.model_selection import train_test_split
x = df1[["tinggi_cm"]]
y = df1[["berat_kg"]]
x_train, x_test, y_train, y_test = train_test_split(
    x, y, test_size=0.2, random_state=7
)
```

Gambar 1.5. Variabel dependen dan independen dibagi menjadi 2

1.6 Melatih model dengan menggunakan regresi linear

Untuk menerapkan regresi linear kita perlu menggunakan LinearRegression milik scikit-learn

```
from sklearn.linear_model import LinearRegression
model = LinearRegression()
model.fit(x_train, y_train)

LinearRegression ① ②
LinearRegression()
```

Gambar 1.6. Linear Regression

1.7 Evaluasi Model dan Menghubungkan 2 Variabel

Menilai seberapa baik model Multiple Linear Regression bekerja dan membentuk persamaan regresi

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

y_pred = model.predict(x_test)
r2 = r2_score(y_test, y_pred)

print("Koefisien (kg per cm):", model.coef_[0])
print("Koefisien (kg per cm): ", model.intercept_)
print("R2 (test): ", r2)
print("R2 (test): ", r2)
print("M2 (kg): ", mean_absolute_error(y_test, y_pred))

mse = mean_absolute_error(y_test, y_pred)
rmse = np.sqrt(mse)

**Toefisien (kg per cm): [0.5518218]
Intersep (kg): [-37.6568823]
R2 (test): 0.24989238901493693

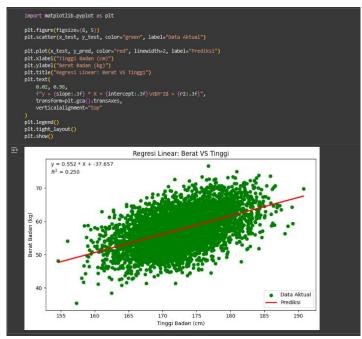
MAE (kg): 3.6704108331736673

**Slope = model.coef_[0][0]
intercept = model.intercept_[0]
print(f*Persamaan: y = (slope:.3f) * X + (intercept:.3f)")

**Persamaan: y = 0.552 * X + -37.657
```

Gambar 1.7. Evaluasi Model menunjukkan Koef, Intercept, R2, MAE

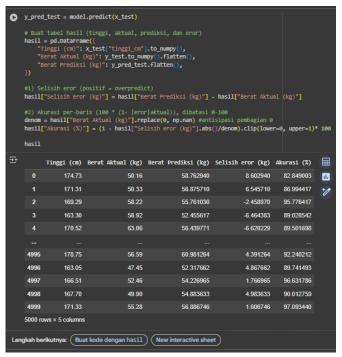
1.8 Visualisasi Regresi Menghubungkan Berat dan TInggi badan



Gambar 1.8. grafik menunjukkan bahwa sekitar 25% variasi berat badan dapat dijelaskan oleh tinggi badan

1.9 Model Prediksi

Menghitung prediksi dan akurasi data hasil prediksi per-baris

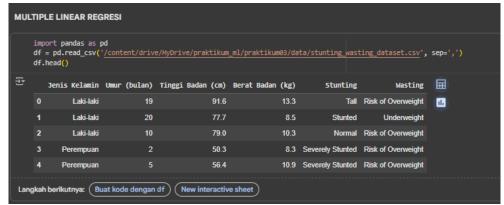


Gambar 1.8. Tabel hasil menampilkan 4 kolom utama

2. MULTIPLE LINEAR REGRESI

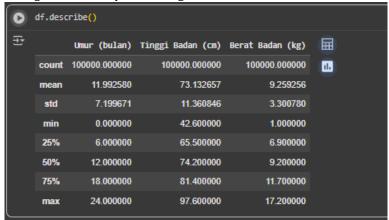
2.1 Membaca Data

Membaca data stunting_wasting_dataset.csv menggunakan library pandas dan menggunakan sep untuk memisahkan koma



Gambar 2.1. Menampilkan Kolom data

2.2 Menghitung statistik deskriptif pada kolom numeric dengan describe Metode .describe() secara otomatis menghitung statistik deskriptif dasar untuk semua kolom numerik.Ini memberikan gambaran cepat tentang distribusi data



Gambar 2.2. Tabel yang berisi metrik-metrik

2.3 Data Pre-processing

mengcopy variabel df dan hanya memakai pada variabel/kolom Berat Badan (kg), Jenis Kelamin, Umur (bulan), Tinggi Badan (cm) dengan menamai dengan df1. Namun dilanjutkan dengan mengubah nama kolom dengan function rename.

Gambar 2.3. Kolom hasil filter data yang diinginkan

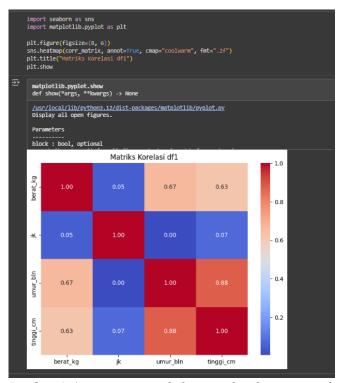
2.4 Analisis Korelasi

Tahapan ini melakukan evaluasi terhadap beberapa variabel indepen (x) untuk memprediksi variable dependen (y).

Gambar 2.4. Evaluasi variable

2.5 Heatmap

Memilih variabel yang dominan untuk dimasukan dalam model prediksi dengan menggunakan fungsi koefisien korelasi .corr(), untuk selanjutnya divisulisasikan dalam bentuk grafik Heatmap.



Gambar 2.5. Heatmap Variabel yang paling berperngaruh

Hasil analisis koefisien korelasi menunjukkan bahwa variabel yang paling berpengaruh dalam prediksi berat badan balita adalah sebagai berikut:

1. Umur: 0.67 → berpengaruh dominan

2. Tinggi: 0.63 → berpengaruh dominan

3. Jenis Kelamin: $0.05 \rightarrow tidak$ berpengaruh signifikan

Berdasarkan hasil korelasi tersebut, variabel yang digunakan untuk membangun model regresi adalah: Variabel independen (X):

o $X_1 = Umur$

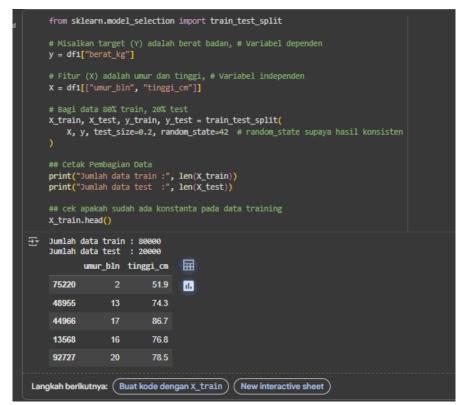
o X_2 = Tinggi

Variabel dependen (Y):

o Y = Berat

2.6 Membagi dataset untuk Training dan Test

Pada tahapan ini membagi dataset menjadi 80% data training dan 20% data testing



Gambar 2.6. Membagi dataset menjadi 80% data training dan 20% data testing

2.7 Pemodelan

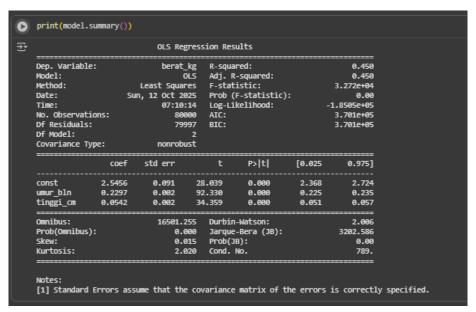
Melakukan pemodelan dengan Pustaka program OLS, dan jalankan training data, kemudian cetak parameter constan, x1 dan x2 dan tampilkan persamaan regresi nya. Cek apakah data training telah memiliki nilai konstan, jika belum ada tambahkan variabel konstan bernilai 1.0

0	▶ import statsmodels.api as sm										
	<pre>X_train_const = sm.add_constant(X_train) X_train_const.head()</pre>										
		const	umur_bln	tinggi_cm							
	75220	1.0	2	51.9	11.						
	48955	1.0	13	74.3							
	44966	1.0	17	86.7							
	13568	1.0	16	76.8							
	92727	1.0	20	78.5							

Gambar 2.7.1 Menambahkan nilai konstan

Gambar 2.7.2 Persamaan Regrsi X1 dan X2

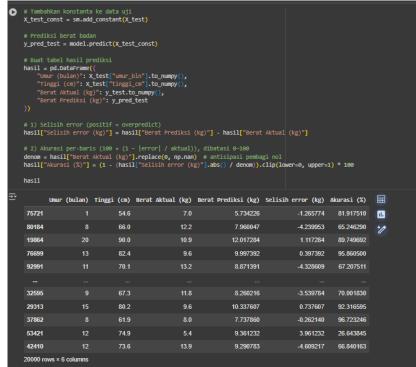
2.8 Model Regresi OLS Cetak informasi model regresi OLS



Gambar 2.8. Model OLS

2.9 Pengujian model dengan data testing

Ordinary Least Squares (OLS) adalah metode statistik yang banyak digunakan untuk memperkirakan parameter model regresi linear. Hasil regresi OLS menunjukkan bahwa model memiliki R-squared sebesar 0.450, artinya 45,0% variasi pada variabel berat badan (berat_kg) dapat dijelaskan oleh variabel umur (umur_bln) dan tinggi (tinggi_cm), sisanya 55% dijelaskan oleh faktor lain yang tidak termasuk dalam model.



Gambar 2.9. Hasil Uji Model dan Data

3. Praktikum Mandiri

Buat model prediksi dari kasus dataset berikut ini: https://www.kaggle.com/datasets/lakshmi25npathi/bike-sharing-dataset

```
⊕ ↑ ↓ å

[1]: import pandas as pd
    # Read the CSV file with a comma delimiter
    df = pd.read_csv('../data/day.csv', sep=',')
    # cetak header data (5 baris data) dari file
   df.head()
[1]: instant dteday season yr mnth holiday weekday workingday weathersit temp atemp hum windspeed casual registered cnt
         1 2011-01-01
                      1 0
                                  0
                                        6
                                              0
                                                        2 0.344167 0.363625 0.805833 0.160446
   1 2 2011-01-02 1 0 1 0 0 0 2 0.363478 0.353739 0.696087 0.248539
                                                                                      131 670 801
      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 0.20000 0.212122 0.590435 0.160296 108 1454 1562
         5 2011-01-05 1 0 1 0 3 1 1 0.226957 0.229270 0.436957 0.186900 82 1518 1600
```

dengan variable dependen (Y) kolom cnt, tentukan variabel independent (x) dari kolom2 yang tersedia !!!

3.1 Baca Dataset

Sel ini membaca dataset yang diinginkan yaitu day.csv

```
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 r2_score, mean_squared_error
import numpy as np

# Baca dataset
df = pd.read_csv('/content/drive/MyDrive/praktikum_ml/praktikum@3/data/day.csv')
df.head()
```

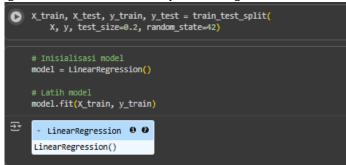
Gambar 3.1. Proses ini hanya perlu dilakukan satu kali per sesi.

3.2 Variabel Independen (X) dan Dependen (Y)

Menentukan Variabel X dan Y

Gambar 3.2. Hasil Variabel

3.3 Membuat Data Training dan Testing serta melatih Model Linear Regrresion Membagi data training dan melatih model Multiple Line Regression



Gambar 3.3. Linear Reression

3.4 Evaluasi Model

Evaluasi model mulai dari hasil prediksi dan skor R2 RMSE

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

# Hitung skor R² dan RMSE
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
```

Gambar 3.4. Hasil Model Evaluasi

3.5 Visualisasi Hasil Prediksi dan Aktual Visualisasi menggunakan matplotlib

plt.figure(figsize=(8,6))
plt.scatter(y_test, y_pred, alpha=0.5, color='royalblue')
plt.xlabel("Milai Aktual (cnt)")
plt.ylabel("Milai Prediksi (cnt)")
plt.plabel("Milai Prediksi (cnt)")
plt.plabel("Milai Prediksi (cnt)")
plt.plot([y_test.min(), y_test.max()], [y_test.min(), y_test.max()], color='red')
plt.grid(True)
plt.show()

Perbandingan Nilai Aktual vs Prediksi Jumlah Penyewaan Sepeda

8000

Perbandingan Nilai Aktual vs Prediksi Jumlah Penyewaan Sepeda

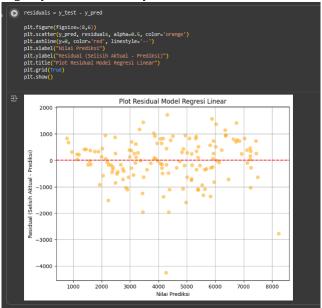
8000

Nilai Aktual vs Prediksi Jumlah Penyewaan Sepeda

Gambar 3.5. Hasil Visualisasi

3.6 Visualisasi Residual (Kesalahan Prediksi)

Visualisasi kesalahan yang dapat muncul saat prediksi



Gambar 3.6. Hasil Visualisasi Residual

3.7 Koefisien Setiap Variabel Melihat Koefisien Setiap Variabel

```
coef_df = pd.DataFrame({
     'Variabel': X.columns,
    'Koefisien': model.coef_
print(coef_df)
      Variabel
                 524.722536
        season
                2023.997547
                 -38.444658
          mnth
       holiday
                 -391.550766
       weekday
                  72.937003
                 160.804892
                 -632.856284
    weathersit
                2097.247836
                3488.042179
                -865.439419
     windspeed -2080.540395
```

Gambar 3.7. Nilai Koefisien variable X

Hasil:

- 1. Nilai R² menunjukkan seberapa baik model menjelaskan variasi data (semakin mendekati 1 semakin baik).
- 2. Grafik pertama menunjukkan seberapa dekat prediksi dengan data aktual.
- 3. Grafik kedua menunjukkan distribusi error (residual), idealnya tersebar acak di sekitar 0.

Referensi:

- Munir, S., Seminar, K. B., Sudradjat, Sukoco, H., & Buono, A. (2022). The Use of Random Forest Regression for Estimating Leaf Nitrogen Content of Oil Palm Based on Sentinel 1-A Imagery. *Information*, *14*(1), 10. https://doi.org/10.3390/info14010010
- Seminar, K. B., Imantho, H., Sudradjat, Yahya, S., Munir, S., Kaliana, I., Mei Haryadi, F., Noor Baroroh, A., Supriyanto, Handoyo, G. C., Kurnia Wijayanto, A., Ijang Wahyudin, C., Liyantono, Budiman, R., Bakir Pasaman, A., Rusiawan, D., & Sulastri. (2024). PreciPalm: An Intelligent System for Calculating Macronutrient Status and Fertilizer Recommendations for Oil Palm on Mineral Soils Based on a Precision Agriculture Approach. *Scientific World Journal*, 2024(1). https://doi.org/10.1155/2024/1788726