Tugas 7 Praktikum Mandiri

SYAHRI GHIFARI MAULIDI 0110222217

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

1. Import Library dan Membaca Dataset

```
import pandas as pd
from sklearn import tree
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import accuracy_score, classification_report, confusion_matrix
```

```
df = pd.read_csv('../Data/Iris.csv')
print(df.head())
```

```
   Id  SepalLengthCm  SepalWidthCm  PetalLengthCm  PetalWidthCm      Species
0   1            5.1           3.5            1.4           0.2  Iris-setosa
1   2            4.9           3.0            1.4           0.2  Iris-setosa
2   3            4.7           3.2            1.3           0.2  Iris-setosa
3   4            4.6           3.1            1.5           0.2  Iris-setosa
4   5            5.0           3.6            1.4           0.2  Iris-setosa
```

Bagian ini digunakan untuk memuat seluruh library yang dibutuhkan dalam analisis.

- pandas (pd) Digunakan untuk memanipulasi dan menganalisis data dalam bentuk tabel (DataFrame).
- numpy (np) Digunakan untuk operasi matematika dan komputasi numerik.
- train_test_split Fungsi dari sklearn.model_selection untuk membagi dataset menjadi dua bagian: data latih (training) dan data uji (testing).
- LinearRegression Model regresi linear dari sklearn.linear_model, digunakan untuk memprediksi hubungan linear antara variabel input dan output.
- mean_squared_error & r2_score Metrik evaluasi dari sklearn.metrics:
- mean_squared_error: Mengukur rata-rata kesalahan kuadrat antara nilai prediksi dan nilai sebenarnya.

- r2_score: Mengukur seberapa baik model menjelaskan variasi data (nilai 1 = sempurna, nilai 0 = tidak ada hubungan linear).

Selanjutnya, dataset satelit dibaca menggunakan pd.read_csv(). pd.read_csv() digunakan untuk membaca file CSV dan mengubahnya menjadi DataFrame (df). Argumen "../Data/dataset_satelit.csv" menunjukkan lokasi file CSV:

- berarti naik satu folder dari direktori kerja saat ini.
- Folder Data berisi file bernama dataset_satelit.csv.

Jadi, kode ini memuat data dari file dataset_satelit.csv ke dalam variabel df.

Fungsi head() menampilkan **5 baris pertama** dari DataFrame. Tujuannya untuk melihat struktur awal dataset, seperti:

- Nama kolom
- Tipe data
- Contoh nilai pada setiap kolom

1.1 Menentukan Fitur (X) dan Target (y)

```
df.columns
```

```
Index(['No', 'Longitude', 'Lattitude', 'N', 'P', 'K', 'Ca', 'Mg', 'Fe', 'Mn',
       'Cu', 'Zn', 'B', 'b12', 'b11', 'b9', 'b8a', 'b8', 'b7', 'b6', 'b5',
       'b4', 'b3', 'b2', 'b1', 'Sigma_VV', 'Sigma_VH', 'plia', 'lia', 'iafe',
       'gamma0_vv', 'gamma0_vh', 'beta0_vv', 'beta0_vh'],
      dtype='object')
```

```
df['Mg'] = pd.to_numeric(df['Mg'], errors='coerce')
df = df.dropna()
```

```
X_3 = df[['b2', 'b3', 'b4', 'b8', 'b11', 'Sigma_VV', 'Sigma_VH']]
y_3 = df['N']
print(df.columns.tolist())
```

Kode ini digunakan untuk menampilkan daftar nama kolom dalam DataFrame df. Output berupa objek Index([...], dtype='object'), yang menunjukkan semua nama kolom dan tipe datanya (object di sini berarti tipe data string atau campuran).

- Kolom 'Mg' kemungkinan berisi nilai angka yang disimpan sebagai teks (string), misalnya '12.5' atau 'N/A'.

- Fungsi pd.to_numeric() mengubah nilai di kolom 'Mg' menjadi tipe numerik (float atau int).
- Parameter errors='coerce' berarti:
- Jika ada nilai yang tidak bisa dikonversi ke angka (misalnya teks “abc”), nilainya akan diganti dengan NaN (Not a Number).

Tujuannya adalah memastikan kolom 'Mg' hanya berisi nilai numerik agar bisa digunakan dalam perhitungan atau analisis model.

- Fungsi dropna() menghapus baris yang memiliki nilai kosong (NaN) di DataFrame.
- Ini penting setelah konversi di atas, karena ada kemungkinan beberapa nilai 'Mg' menjadi NaN akibat gagal dikonversi.
- X_3 → berisi kolom-kolom fitur (variabel independen) yang digunakan untuk memprediksi.
- Kolom yang dipilih: 'b2', 'b3', 'b4', 'b8', 'b11', 'Sigma_VV', 'Sigma_VH'
- Kolom ini biasanya adalah data reflektansi atau sinyal radar dari satelit (misalnya Sentinel).
- y_3 → kolom target (variabel dependen) yang akan diprediksi, yaitu 'N' (kemungkinan kadar nitrogen pada tanah atau tanaman).
- df.columns.tolist() mengubah daftar kolom (Index) menjadi list Python biasa.
- Fungsinya sama seperti df.columns, hanya beda format output (lebih mudah dibaca atau digunakan untuk manipulasi lanjutan).

1.2 Membuat dan Melatih Model Regresi Linear

```
X_train, X_test, y_train, y_test = train_test_split(X_3, y_3, test_size=0.2, random_state=42)
```

```
model = LinearRegression()
model.fit(X_train, y_train)
```

```
▾ LinearRegression
▶ Parameters
```

Fungsi train_test_split() digunakan untuk membagi dataset menjadi dua bagian:

- Data latih (train set) → digunakan untuk melatih model.
- Data uji (test set) → digunakan untuk menguji performa model.
- LinearRegression() adalah model regresi linear dari scikit-learn.
- fit(X_train, y_train) berfungsi untuk melatih model agar menemukan hubungan linear antara variabel input (X_train) dan output (y_train).

Setelah proses ini, model memiliki koefisien (slope) dan intercept (titik potong) yang menjelaskan hubungan antar variabel.

2. Melakukan Prediksi dan Evaluasi Model

```
y_pred = model.predict(X_test)
r2 = r2_score(y_test, y_pred)
mse = mean_squared_error(y_test, y_pred)
rmse = np.sqrt(mse)
print(f'R2: {r2}')
print(f'RMSE: {rmse}')
```

- model.predict(X_test) → menghasilkan prediksi nilai N berdasarkan fitur pada data uji.
- r2_score(y_test, y_pred) → menghitung koefisien determinasi (R^2):
- Nilai R^2 mendekati 1 → model sangat baik menjelaskan data.
- Nilai R^2 mendekati 0 → model buruk (tidak mampu menjelaskan variasi data).
- mean_squared_error() → menghitung Mean Squared Error (MSE), yaitu rata-rata kuadrat selisih antara nilai prediksi dan nilai aktual.
- np.sqrt(mse) → menghitung Root Mean Squared Error (RMSE), yaitu akar dari MSE.
- Semakin kecil RMSE, semakin akurat model.

2.1 Melihat Koefisien Setiap Fitur

```
coeff = pd.DataFrame({
    'Fitur': X_3.columns,
    'Koefisien': model.coef_
})
print(coeff)
```

- model.coef_ menyimpan nilai koefisien regresi untuk setiap variabel independen.
- Kode ini membuat DataFrame baru berisi:
- Kolom Fitur: nama fitur dari dataset.
- Kolom Koefisien: bobot/kontribusi tiap fitur terhadap target (N).
- Nilai koefisien positif → fitur tersebut meningkatkan nilai prediksi N.
- Nilai koefisien negatif → fitur tersebut menurunkan nilai prediksi N

2.2 Analisis Regresi Menggunakan statsmodels

```
import statsmodels.api as sm

X_train_sm = sm.add_constant(X_train)
model_sm = sm.OLS(y_train, X_train_sm).fit()
print(model_sm.summary())
```

- statsmodels digunakan untuk analisis statistik yang lebih mendalam dibanding sklearn.
- sm.add_constant(X_train) → menambahkan kolom konstanta (intercept) ke data fitur.
- sm.OLS() → membuat model Ordinary Least Squares (OLS), yaitu regresi linear klasik.
- .fit() → melatih model.
- .summary() → menampilkan ringkasan statistik lengkap seperti:
- Koefisien dan p-value tiap variabel.
- R-squared dan Adjusted R-squared.
- Statistik uji F, t, Durbin-Watson, dll.

📊 Hasil dari summary() memberikan gambaran apakah setiap fitur signifikan terhadap variabel target (N).