Nama	: Muhammad Hisyam Kamil
NIM	: 202210370311060
Mata Kuliah	: Data, Informasi, dan Pengetahuan
Kelas	: 6B
Source Code	: https://github.com/hisyam99/dip-task5
Google Colab (IPYNB)	: https://mil.kamil.my.id/dip-task5-hisyam99
Google Drive (Result Files)	: https://mil.kamil.my.id/dip-task5-files-hisyam99
Datasets	: https://mil.kamil.my.id/dip-task5-dataset-hisyam99

Investigasi Seleksi Fitur Melalui Korelasi Pearson dan Information Gain

1. Latar Belakang

Koefisien Korelasi Pearson adalah metrik statistik yang digunakan untuk mengukur kekuatan dan arah hubungan linier antara dua variabel numerik. Nilainya berkisar dari -1 (korelasi negatif sempurna) hingga +1 (korelasi positif sempurna), dengan nilai 0 menunjukkan tidak adanya hubungan linier. Metode ini membantu mengidentifikasi fitur numerik yang berkorelasi erat dengan variabel target, yang sangat penting dalam membangun model prediktif.

Information Gain, berdasarkan teori informasi, mengukur penurunan entropi (ketidakpastian) pada variabel target akibat adanya fitur tertentu. Entropi mencerminkan tingkat ketidakpastian dalam klasifikasi; semakin tinggi nilai Information Gain, semakin relevan fitur tersebut untuk prediksi.

Laporan ini menganalisis dataset Iris untuk mengidentifikasi fitur paling relevan terhadap variabel target, yaitu spesies bunga Iris (Iris-setosa, Iris-versicolor, Iris-virginica), yang dikonversi menjadi nilai numerik (0, 1, 2). Dataset ini memiliki empat fitur numerik: panjang sepal, lebar sepal, panjang petal, dan lebar petal.

2. Prosedur Analisis

Dataset Iris, yang terdiri dari 150 sampel dan 5 kolom (4 fitur numerik dan 1 target kategorikal), diolah melalui tahapan berikut:

a. Pra-pemrosesan Data

1. Dataset diunduh dari repositori UCI Machine Learning.
2. Kolom dinamai sebagai: `sepal_length`, `sepal_width`, `petal_length`, `petal_width`, dan `species`.
3. Nilai hilang (jika ada) dihapus menggunakan fungsi `dropna()`, meskipun dataset Iris sudah bersih.
4. Kolom `species` (kategorikal) diubah menjadi numerik (setosa: 0, versicolor: 1, virginica: 2) dan disimpan sebagai kolom `target`.

b. Kalkulasi Korelasi Pearson

Korelasi dihitung untuk setiap fitur numerik terhadap kolom `target` menggunakan metode `corrwith()` dari pustaka pandas.

c. Kalkulasi Information Gain

Information Gain dihitung menggunakan fungsi `mutual_info_classif` dari scikit-learn, dengan parameter `random_state=42` untuk konsistensi hasil.

d. Pembuatan Visualisasi

Heatmap korelasi dibuat menggunakan `seaborn.heatmap` dengan skema warna `coolwarm` (merah untuk korelasi positif, biru untuk negatif). Visualisasi disimpan sebagai `correlation_heatmap.png`.

3. Ringkasan Hasil Kalkulasi

a. Nilai Koefisien Korelasi Pearson

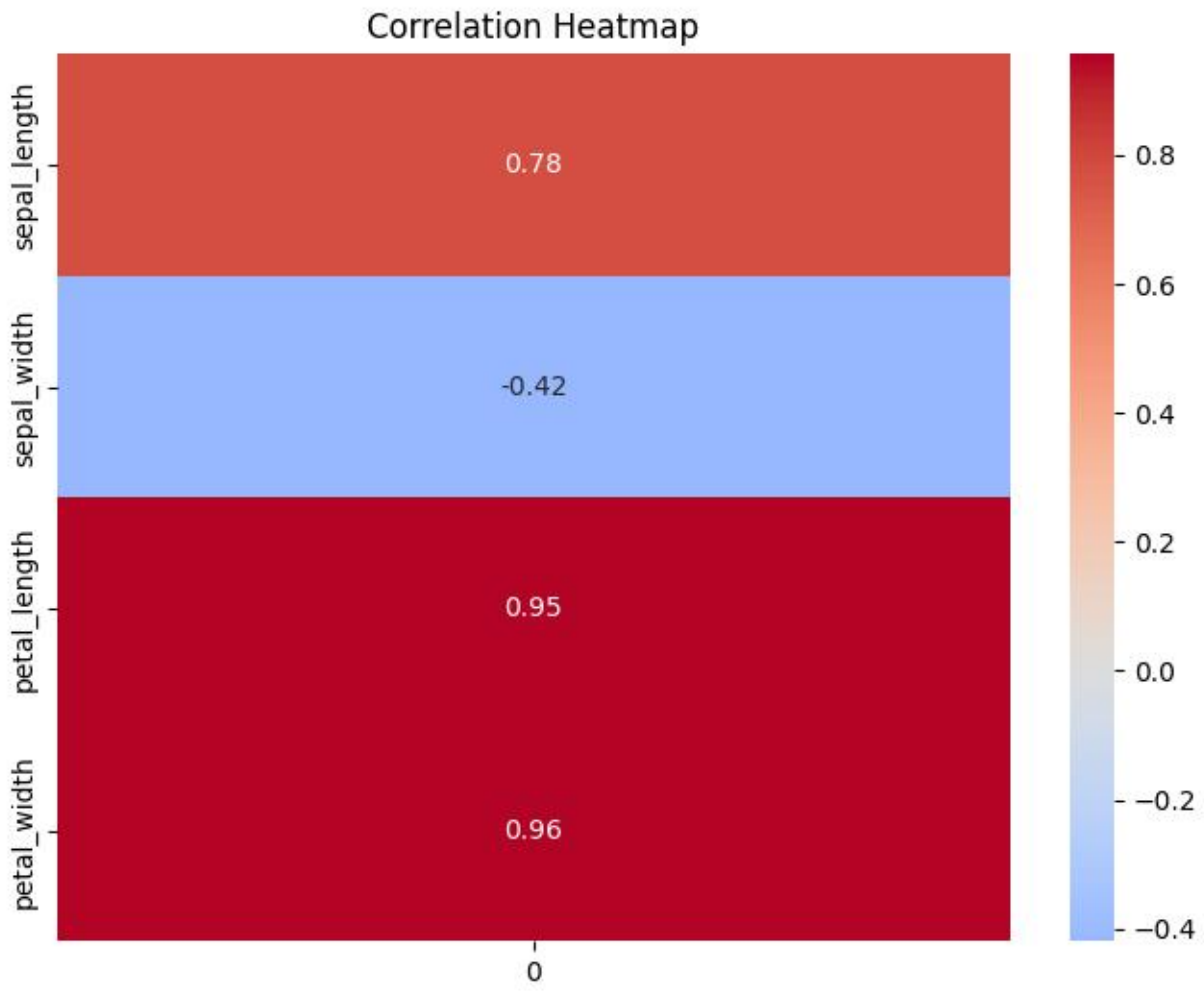
- Panjang Sepal: 0.782561
- Lebar Sepal: -0.419446
- Panjang Petal: 0.949043
- Lebar Petal: 0.956464

b. Nilai Information Gain

- Panjang Sepal: 0.511365
- Lebar Sepal: 0.289759
- Panjang Petal: 0.992573
- Lebar Petal: 0.985643

c. Deskripsi Visualisasi

Heatmap korelasi menunjukkan bahwa `panjang petal` dan `lebar petal` memiliki warna merah pekat, mengindikasikan korelasi positif kuat dengan variabel target. Sebaliknya, `lebar sepal` berwarna biru, menunjukkan korelasi negatif. `Panjang sepal` memiliki warna merah terang, menandakan korelasi positif sedang.



Gambar 1. Heatmap Korelasi

4. Implementasi Kode Python

Berikut adalah kode Python yang digunakan untuk analisis:

```python
import pandas as pd
import numpy as np
from sklearn.preprocessing import LabelEncoder
from sklearn.feature_selection import mutual_info_classif
import seaborn as sns
import matplotlib.pyplot as plt

# Load dataset
url = "https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data"
columns = ['sepal_length', 'sepal_width', 'petal_length', 'petal_width', 'species']
df = pd.read_csv(url, header=None, names=columns)

# Preprocessing
# Remove null values (if any)
df = df.dropna()
```

```python
# Encode species to numeric
le = LabelEncoder()
df['target'] = le.fit_transform(df['species'])

# Select features and target
X = df[['sepal_length', 'sepal_width', 'petal_length', 'petal_width']]
y = df['target']

# Calculate Pearson Correlation
correlations = X.corrwith(y)
print("Pearson Correlation Coefficients:")
print(correlations)

# Calculate Information Gain
info_gain = mutual_info_classif(X, y, random_state=42)
info_gain = pd.Series(info_gain, index=X.columns)
print("\nInformation Gain:")
print(info_gain)

# Create heatmap for correlation
plt.figure(figsize=(8, 6))
sns.heatmap(correlations.to_frame(), annot=True, cmap='coolwarm', center=0)
plt.title('Correlation Heatmap')
plt.savefig('correlation_heatmap.png')
plt.show()
```

5. Pembahasan Hasil

a. Interpretasi Korelasi Pearson

- **Lebar Petal** (0.956464) dan **Panjang Petal** (0.949043) memiliki korelasi linier positif yang sangat kuat dengan variabel target, menunjukkan bahwa spesies dengan ukuran petal lebih besar cenderung termasuk dalam kelas 2 (Iris-virginica).
- **Panjang Sepal** (0.782561) menunjukkan korelasi positif sedang.
- **Lebar Sepal** (-0.419446) memiliki korelasi negatif yang relatif lemah.

b. Interpretasi Information Gain

- **Panjang Petal** (0.992573) dan **Lebar Petal** (0.985643) memberikan informasi tertinggi untuk prediksi variabel target, menegaskan peran penting petal dalam membedakan spesies Iris.
- **Panjang Sepal** (0.511365) cukup informatif, tetapi kontribusinya lebih rendah.
- **Lebar Sepal** (0.289759) memberikan kontribusi informasi paling kecil.

c. Kajian Visualisasi

Heatmap secara visual mengkonfirmasi korelasi positif kuat pada `panjang petal` dan `lebar petal` (warna merah pekat), korelasi negatif pada `lebar sepal` (warna biru), dan korelasi sedang pada `panjang sepal` (warna merah terang).

d. Komparasi Metode

Korelasi Pearson dan Information Gain menghasilkan kesimpulan serupa: `panjang petal` dan `lebar petal` adalah fitur paling relevan. Namun, `lebar petal` unggul menurut Korelasi Pearson, sedangkan `panjang petal` lebih tinggi berdasarkan Information Gain. Kedua metode setuju bahwa `lebar sepal` adalah fitur paling kurang relevan.

6. Simpulan Akhir

1. **Fitur Utama**: `Panjang petal` dan `lebar petal` adalah fitur paling signifikan untuk prediksi kelas spesies Iris. Kedua fitur ini harus diprioritaskan dalam model prediktif.
2. **Fitur Pendukung**: `Panjang sepal` relevan sebagai fitur sekunder, sedangkan `lebar sepal` dapat diabaikan karena relevansinya rendah dan korelasi negatifnya.
3. **Temuan Botani**: Ukuran petal merupakan karakteristik utama yang membedakan spesies Iris, konsisten dengan pengetahuan botani.