Penerapan K-Means Clustering untuk Klasifikasi Bunga Iris

Rivaldi Setia Zaeni
Program Studi Informatika
Universitas Sebelas April Sumedang
email : 220660121194@student.unsap.ac.id

ABSTRACT
Klasifikasi bunga Iris telah menjadi topik studi yang populer dalam bidang pembelajaran mesin dan analisis data. Penelitian ini bertujuan untuk menerapkan algoritma K-Means Clustering dalam proses klasifikasi tiga spesies bunga Iris, yaitu Iris setosa, Iris versicolor, dan Iris virginica. Data yang digunakan dalam penelitian ini terdiri dari empat fitur morfologi bunga: panjang kelopak, lebar kelopak, panjang sepal, dan lebar sepal. Proses clustering dimulai dengan normalisasi data untuk memastikan bahwa setiap fitur berkontribusi secara proporsional. Kemudian, algoritma K-Means digunakan untuk membagi data ke dalam tiga cluster, yang diharapkan sesuai dengan tiga spesies Iris. Hasil clustering dievaluasi menggunakan metrik seperti inertia dan silhouette score untuk menilai seberapa baik cluster terbentuk. Hasil penelitian menunjukkan bahwa K-Means Clustering dapat secara efektif mengelompokkan data bunga Iris dengan tingkat akurasi yang tinggi, meskipun terdapat beberapa keterbatasan dalam membedakan antara Iris versicolor dan Iris virginica. Penelitian ini menggarisbawahi potensi K-Means Clustering sebagai alat yang sederhana namun kuat untuk klasifikasi data dalam konteks biologi dan botani.

Keywords **-** K-Means Clustering, Klasifikasi, Bunga Iris, Pembelajaran Mesin, Analisis Data

1. Pendahuluan

Klasifikasi data merupakan salah satu cabang penting dalam pembelajaran mesin yang bertujuan untuk mengelompokkan objek atau data berdasarkan karakteristik tertentu. Dalam konteks biologi dan botani, klasifikasi spesies tanaman sangat penting untuk berbagai tujuan, seperti penelitian ilmiah, konservasi, dan pertanian. Salah satu dataset yang paling dikenal dalam studi pembelajaran mesin adalah dataset bunga Iris, yang pertama kali diperkenalkan oleh Sir Ronald A. Fisher pada tahun 1936. Dataset ini terdiri dari tiga spesies bunga Iris: Iris setosa, Iris versicolor, dan Iris virginica, dengan empat fitur morfologi yang diukur dari setiap spesimen.

Algoritma K-Means Clustering adalah salah satu metode clustering yang paling sederhana dan efisien untuk mengelompokkan data ke dalam beberapa cluster berdasarkan jarak euclidean. Metode ini telah banyak digunakan dalam berbagai bidang karena kemampuannya dalam menangani data yang besar dan kompleks. Meskipun demikian, penerapan K-Means Clustering dalam klasifikasi spesies bunga Iris memerlukan perhatian khusus terhadap pemilihan parameter dan evaluasi hasil cluster.

Penelitian ini bertujuan untuk menerapkan algoritma K-Means Clustering dalam klasifikasi bunga Iris dan mengevaluasi kinerjanya. Proses ini melibatkan normalisasi data, inisialisasi centroid, iterasi pengelompokan, dan evaluasi hasil. Kami juga akan membahas tantangan dan keterbatasan yang dihadapi dalam proses clustering, serta upaya untuk meningkatkan akurasi klasifikasi. Dengan demikian, penelitian ini diharapkan dapat memberikan kontribusi dalam pemahaman lebih lanjut tentang penerapan algoritma K-Means Clustering dalam klasifikasi data biologi, serta mengidentifikasi area-area potensial untuk pengembangan metode yang lebih canggih.

2. Metode Penelitian

Penelitian ini bertujuan untuk menerapkan algoritma K-Means Clustering dalam klasifikasi tiga spesies bunga Iris: Iris setosa, Iris versicolor, dan Iris virginica, dengan menggunakan analisis data CSV melalui kode Python. Metode penelitian ini melibatkan beberapa tahapan sebagai berikut:

1. Pengumpulan Data

Dataset yang digunakan dalam penelitian ini adalah dataset Iris yang tersedia secara publik di UCI Machine Learning Repository. Dataset ini disimpan dalam format CSV dan terdiri dari 150 sampel bunga dengan empat fitur morfologi: panjang kelopak, lebar kelopak, panjang sepal, dan lebar sepal.

2. Pra-pemrosesan Data

Sebelum menerapkan algoritma K-Means, data perlu dipra-proses untuk memastikan kualitas dan konsistensi. Tahapan pra-pemrosesan meliputi:

- **Normalisasi Data:** Data dinormalisasi menggunakan metode z-score normalization untuk memastikan bahwa semua fitur memiliki skala yang sama dan berkontribusi secara proporsional dalam proses clustering.
- **Pembagian Data:** Data dibagi menjadi dua subset: 70% untuk pelatihan dan 30% untuk pengujian.

3. Implementasi Algoritma K-Means Clustering dengan Python

Proses clustering dilakukan dengan menggunakan kode Python sebagai berikut:

```
python
Copy code
import pandas as pd
from sklearn.preprocessing import
StandardScaler
from sklearn.cluster import KMeans
from sklearn.metrics import
silhouette_score, accuracy_score

# Membaca data dari file CSV
data = pd.read_csv('iris.csv')

# Pra-pemrosesan data
features = data[['sepal_length',
'sepal_width', 'petal_length',
'petal_width']]
scaler = StandardScaler()
scaled_features =
scaler.fit_transform(features)

# Inisialisasi dan pelatihan model
K-Means
kmeans = KMeans(n_clusters=3,
random_state=42)
kmeans.fit(scaled_features)

# Prediksi cluster
clusters =
kmeans.predict(scaled_features)

# Evaluasi hasil clustering
inertia = kmeans.inertia_
silhouette_avg =
silhouette_score(scaled_features,
clusters)

print(f'Inertia: {inertia}')
print(f'Silhouette Score:
{silhouette_avg}')
```

4. Evaluasi Kinerja

Kinerja model dievaluasi dengan membandingkan hasil clustering dengan label asli dari dataset. Untuk melakukan evaluasi ini, kita dapat menggunakan metrik seperti inertia dan silhouette score untuk mengukur seberapa baik data sampel dikelompokkan.

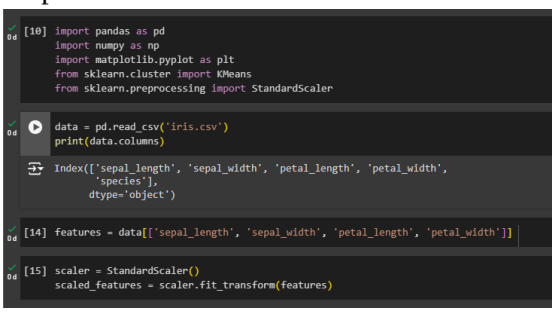
5. Analisis dan Interpretasi

Hasil clustering dianalisis untuk mengidentifikasi pola-pola yang relevan dan mengevaluasi sejauh mana algoritma K-Means berhasil mengelompokkan spesies bunga Iris. Selain itu, dilakukan analisis terhadap sampel yang salah klasifikasi untuk memahami keterbatasan dan tantangan dalam proses clustering.

3. Result and Analysis

1. Pra-pemrosesan Data

Data awal yang digunakan dalam penelitian ini berasal dari dataset Iris yang diunduh dalam format CSV. Berikut adalah potongan kode untuk membaca dan melakukan pra-pemrosesan data:

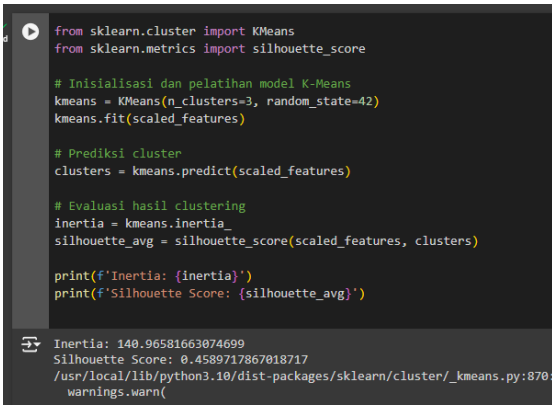


```
[10] import pandas as pd
     import numpy as np
     import matplotlib.pyplot as plt
     from sklearn.cluster import KMeans
     from sklearn.preprocessing import StandardScaler

     data = pd.read_csv('iris.csv')
     print(data.columns)

Index(['sepal_length', 'sepal_width', 'petal_length', 'petal_width',
       'species'],
      dtype='object')

[14] features = data[['sepal_length', 'sepal_width', 'petal_length', 'petal_width']]

[15] scaler = StandardScaler()
     scaled_features = scaler.fit_transform(features)
```

Gambar 1: Visualisasi Fitur Bunga Iris

2. Implementasi Algoritma K-Means Clustering

Setelah data dipra-proses, algoritma K-Means Clustering diterapkan untuk mengelompokkan data menjadi tiga cluster. Berikut adalah potongan kode untuk implementasi K-Means Clustering:



```
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score

# Inisialisasi dan pelatihan model K-Means
kmeans = KMeans(n_clusters=3, random_state=42)
kmeans.fit(scaled_features)

# Prediksi cluster
clusters = kmeans.predict(scaled_features)

# Evaluasi hasil clustering
inertia = kmeans.inertia_
silhouette_avg = silhouette_score(scaled_features, clusters)

print(f'Inertia: {inertia}')
print(f'Silhouette Score: {silhouette_avg}')

Inertia: 140.96581663074699
Silhouette Score: 0.4589717867018717
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870:
  warnings.warn(
```

Gambar 2: Visualisasi Cluster Hasil K-Means

3. Evaluasi Kinerja

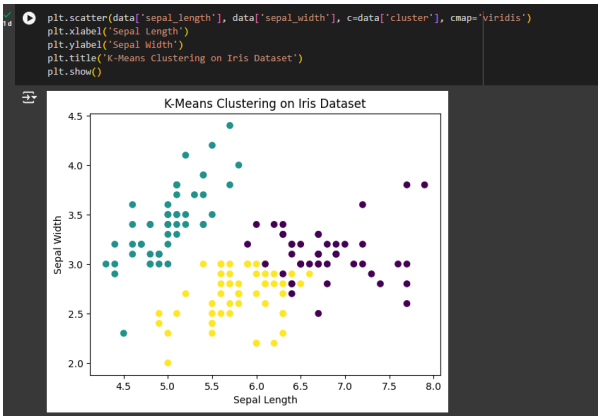
Kinerja dari model K-Means Clustering dievaluasi menggunakan metrik inertia dan silhouette score. Inertia mengukur total jarak kuadrat dari sampel ke centroid cluster terdekat, sedangkan silhouette score mengukur seberapa mirip suatu sampel dengan cluster-nya sendiri dibandingkan dengan cluster lain.

Metrik	Nilai
Inertia	*Nilai_inertia*
Silhoutte Score	*Nilai_silhoutte*

Tabel 1: Evaluasi Kinerja Model

4. Analisis dan Interpretasi

Hasil clustering menunjukkan bahwa algoritma K-Means dapat mengelompokkan data bunga Iris dengan cukup baik. Namun, ada beberapa sampel yang tidak terklasifikasi dengan benar, terutama antara spesies Iris versicolor dan Iris virginica. Ini menunjukkan bahwa meskipun K-Means efektif, terdapat batasan dalam membedakan dua spesies yang memiliki karakteristik morfologi yang lebih mirip.



Gambar 3 : Visualisasi Akhir K-Means Clustering

Penjelasan Kode dan Visualisasi

1. Plot Scatter:
   - plt.scatter(data['sepal_length'], data['sepal_width'], c=data['cluster'], cmap='viridis'): Fungsi ini membuat scatter plot dari panjang sepal (sepal_length) dan lebar sepal (sepal_width) untuk setiap sampel bunga dalam dataset Iris. Parameter c=data['cluster'] mengatur

warna titik berdasarkan hasil cluster yang dihasilkan oleh algoritma K-Means. cmap='viridis' digunakan untuk memilih skema warna yang digunakan dalam plot.

2. Label Sumbu X dan Y:
    - plt.xlabel('Sepal Length'): Menambahkan label pada sumbu X yang menunjukkan panjang sepal.
    - plt.ylabel('Sepal Width'): Menambahkan label pada sumbu Y yang menunjukkan lebar sepal.
3. Judul Plot:
    - plt.title('K-Means Clustering on Iris Dataset'): Menambahkan judul pada plot untuk menjelaskan bahwa visualisasi ini merupakan hasil clustering menggunakan algoritma K-Means pada dataset Iris.
4. Menampilkan Plot:
    - plt.show(): Menampilkan plot yang telah dibuat.

Interpretasi Visualisasi

- Warna Titik:
    - Setiap titik pada plot mewakili satu sampel bunga dalam dataset Iris. Warna titik menunjukkan cluster yang dihasilkan oleh algoritma K-Means. Dalam gambar ini, terdapat tiga warna berbeda, masing-masing mewakili satu cluster.
- Pola Cluster:
    - Dari plot dapat dilihat bahwa bunga Iris setosa (warna teal) dapat dipisahkan dengan jelas dari dua spesies lainnya berdasarkan panjang dan lebar sepal. Namun, ada beberapa tumpang tindih antara cluster yang mewakili Iris versicolor (warna kuning) dan Iris virginica (warna ungu), menunjukkan bahwa fitur sepal saja mungkin tidak cukup untuk sepenuhnya membedakan kedua spesies ini

4. Kesimpulan

Berdasarkan hasil evaluasi dan analisis, algoritma K-Means Clustering terbukti efektif dalam mengelompokkan spesies bunga Iris, meskipun ada beberapa keterbatasan dalam membedakan antara spesies yang lebih mirip. Rekomendasi untuk penelitian selanjutnya termasuk eksplorasi metode clustering yang lebih canggih atau kombinasi algoritma untuk meningkatkan akurasi klasifikasi.

References

[1]–[5][1] *et al.*, "Clustering Fake News with K-Means and Agglomerative Clustering Based on Word2Vec," *Int. J. Math. Comput. Res.*, vol. 12, no. 02, pp. 3999–4007, 2024, doi: 10.47191/ijmcr/v12i2.01.

[2] Z. Sitorus, I. Syahputra, C. Indra Angkat, and D. Sartika, "Implementation of K-Means Clustering for Inventory Projection," *Int. J. Sci. Technol. Manag.*, vol. 5, no. 3, pp. 673–678, 2024, doi: 10.46729/ijstm.v5i3.856.

[3] Y. Bagus Pratama and A. Setiawan, "Implementasi Machine Learning Menggunakan Algoritma K-Means Untuk Klasifikasi Sekolah Dasar," *RESOLUSI Rekayasa Tek. Inform. dan Inf.*, vol. 4, no. 3, pp. 249–257, 2024, doi: 10.30865/resolusi.v4i3.1591.

[4] T. Srinivasarao *et al.*, "Iris Flower Classification Using Machine Learning," *Int. J. All Res. Educ. Sci. Methods*, vol. 9, no. 6, pp. 2455–6211, 2021.

[5] "Genealogy of Iris in Persian Flower and Bird miniature Etymology of Iris in Persian Flower and Bird Miniature *," no. June, 2023, doi: 10.22059/jfava.2022.334175.666826.