MATHunesa
Jurnal Ilmiah Matematika **Volume 13 No 02**
e-ISSN : 2716-506X | p-ISSN : 2301-9115 **Tahun 2025**

Analisis Cluster Menggunakan K-Means: Studi Kasus pada Data Penguin

Azhaar Dini Mahdiyyah
Program Studi S1 Matematika, FMIPA, Universitas Negeri Surabaya, Surabaya, Indonesia
e-mail: : azhaar.23061@mhs.unesa.ac.id*

Zahra Ar - Rayan Pandawuli
Program Studi S1 Matematika, FMIPA, Universitas Negeri Surabaya, Surabaya, Indonesia
e-mail: zahra.23084@mhs.unesa.ac.id

Nadya Nafis Kartosen
Program Studi S1 Matematika, FMIPA, Universitas Negeri Surabaya, Surabaya, Indonesia
e-mail: nadya.23123@mhs.unesa.ac.id
Mahesa Aryo Wicaksono
Program Studi S1 Matematika, FMIPA, Universitas Negeri Surabaya, Surabaya, Indonesia
e-mail: mahesa.23157@mhs.unesa.ac.id

Abstrak

Penelitian ini dibuat untuk mempelajari pola penyebaran spesies *penguins* berdasarkan ciri-ciri fisik seperti panjang paruh, lebar paruh, panjang sirip, dan massa tubuh. Metode *clustering* K-Means digunakan untuk melakukan pengelompokan data, sementara *Principal Component Analysis* (*PCA*) dimanfaatkan untuk mengurangi dimensi data sehingga hasil pengelompokan dapat divisualisasikan dengan lebih mudah. Dari hasil analisis, diketahui bahwa data penguin secara optimal dapat dikelompokkan menjadi tiga *cluster*, sesuai dengan jumlah spesies yang ada dalam dataset. Pola pengelompokan ini memberikan wawasan yang signifikan untuk identifikasi spesies penguin dengan lebih efisien. Penelitian ini berkontribusi sebagai acuan dalam pengelolaan data biologis serta pengembangan sistem informasi untuk konservasi.

Abstract

This research was conducted to study the distribution patterns of penguin species based on physical characteristics such as bill length, bill depth, flipper length, and body mass. The K-Means clustering method was used for data grouping, while Principal Component Analysis (PCA) was applied to reduce data dimensionality, allowing the clustering results to be visualized more easily. The analysis revealed that the penguin data can be optimally grouped into three clusters, corresponding to the number of species present in the dataset. This clustering pattern provides significant insights for more efficient penguin species identification. The study contributes as a reference for biological data management as well as the development of information systems for conservation purposes.

Kata kunci: *Clustering, K-Means, PCA, Penguins, Data Mining.*

Pendahuluan

Pengelompokan (*clustering*) adalah teknik analisis data yang digunakan untuk mengelompokkan objek yang mamiliki kesamaan tertentu. Dalam ekologi, teknik *clustering* digunakan untuk mengidentifikasi pola dalam data morfologi spesies, yang dapat memberikan wawasan tentang adaptasi spesies terhadap lingkungan. Salah satu aplikasi penting dari teknik ini adalah pengelompokan spesies penguin berdasarkan karakteristik morfologi tubuh mereka.

Penelitian berjudul *Klasterisasi Hewan Berdasarkan Morfologi dengan K-Means Klastering untuk Memudahkan Pemahaman Taksonomi Hewan* oleh Pangestu et al. (2024) menerapkan algoritma *K-Means* untuk mengelompokkan hewan berdasarkan karakteristik morfologi mereka. Hasilnya menunjukkan bahwa klasterisasi efektif dalam mengidentifikasi kelompok hewan dengan kemiripan morfologi, yang pada gilirannya memberikan wawasan tentang adaptasi spesies terhadap lingkungan mereka.

Selain itu, studi oleh Ainley et al. (2007) membahas pentingnya analisis morfologi dalam memahami distribusi populasi dan adaptasi ekologis spesies penguin. Mereka menyoroti bagaimana karakteristik fisik penguin dapat memberikan informasi berharga tentang perilaku dan ekologi mereka.

Dengan demikian, teknik *clustering* dalam ekologi, khususnya dalam pengelompokan spesies penguin berdasarkan morfologi, telah terbukti menjadi alat yang efektif untuk memahami adaptasi spesies terhadap lingkungan mereka.

Penguin merupakan kelompok burung yang memiliki peran penting dalam menjaga keseimbangan ekosistem laut. Sebagai predator, penguin berkontribusi dalam mengendalikan populasi mangsa seperti ikan kecil dan krill, yang secara langsung memengaruhi dinamika rantai makanan laut (Wilson et al., 2005). Selain itu, penguin sering digunakan sebagai indikator biologis untuk mengevaluasi dampak perubahan lingkungan pada ekosistem laut, karena pola makan dan reproduksi mereka sangat dipengaruhi oleh perubahan kondisi laut (Ainley et al., 2010). Penelitian sebelumnya menunjukkan bahwa analisis morfologi spesies penguin memainkan peran penting dalam pemetaan distribusi populasi serta analisis adaptasi ekologis (Ainley et al., 2007). Meskipun demikian, sebagian besar penelitian yang ada cenderung mengandalkan metode klasifikasi berbasis data terbatas, seperti pohon keputusan atau regresi logistik (Smith et al., 2012). Metode-metode ini, meskipun bermanfaat, memiliki keterbatasan dalam menangkap pola tersembunyi yang kompleks dalam data morfologi penguin (Jones & Roberts, 2015).

Metode *clustering*, khususnya *K-Mean*s, memberikan pendekatan yang lebih fleksibel untuk mengelompokkan *spesies penguins* berdasarkan karakteristik morfologi utama. Namun, penerapan metode *K-Means clustering* dengan mempertimbangkan faktor tambahan, seperti jenis kelamin, serta pengurangan dimensi menggunakan *Principal Component Analysis (PCA)* masih terbatas. PCA memainkan peran kunci dalam menyederhanakan data berdimensi tinggi, dengan mempertahankan informasi penting dan mengurangi redundansi dalam variabel-variabel morfologi penguin yang kompleks. Proses ini memungkinkan visualisasi yang lebih jelas dan efisien dari hasil *clustering*. Sebagai contoh, penelitian oleh Smith et al. (2020) menunjukkan bagaimana PCA digunakan untuk mengurangi kompleksitas data sebelum menerapkan *K-Means clustering*, sehingga meningkatkan akurasi dan efektivitas pengelompokan spesies penguin. PCA secara signifikan memperbaiki kualitas hasil *clustering* dengan memungkinkan pengidentifikasian pola tersembunyi yang tidak dapat ditangkap dengan metode klasifikasi konvensional seperti pohon keputusan (Jones & Roberts, 2018). Lebih jauh lagi, metode ini juga telah diterapkan dalam berbagai domain, seperti analisis faktor risiko dalam data medis, menunjukkan potensi besar PCA dalam mengatasi tantangan dalam data berdimensi tinggi (Brown & Green, 2019).

Penelitian ini bertujuan untuk mengisi kesenjangan tersebut dengan menerapkan metode *K-Means* untuk pengelompokan *spesies penguins* berdasarkan data morfologi mereka, serta mengevaluasi hasil *clustering* menggunakan metode *Elbow* dan *Silhouette Score* untuk menentukan jumlah *cluster* yang optimal. Penelitian ini juga mengenalkan penggunaan PCA sebelum *clustering* untuk meningkatkan akurasi dan visualisasi hasil yang diharapkan memberikan wawasan lebih dalam adaptasi *spesies penguins* terhadap berbagai faktor lingkungan.

Metode Penelitian

Penelitian ini menggunakan dataset Palmer *Penguins* untuk menganalisis pola distribusi *spesies penguins* berdasarkan karakteristik morfologi. Karateristik utama yang digunakan mencakup panjang paruh (*culmen length*), lebar paruh (*culmen depth*), panjang sirip, massa tubuh, serta jenis kelamin. Penelitian ini dirancang dalam beberapa tahap utama seperti dijelaskan dibawah ini.

1. Dataset

Dataset yang digunakan adalah dataset Palmer *Penguins* yang tersedia secara publik di Kaggle. Dataset ini mencakup tiga spesies penguin *(Adelie, Chinstrap, Gentoo).* Dataset ini mencakup 344 sampel penguin dengan berbagai karakteristik morfologi seperti panjang paruh, lebar paruh, panjang sirip, massa tubuh, serta jenis kelamin.

Semua fitur ini berupa data numerik dan bersifat kontinu. Oleh karena itu,fitur-fitur ini sangat sesuai untuk diterapkan dalam metode *clustering* seperti algoritma *K-Means*. Dengan data numerik, jarak antar sampel dapat dihitung, yang menjadi inti dari proses *clustering*. Selain itu, fitur-fitur tersebut mempermudah pengelompokan data secara alami berdasarkan kesamaan karakteristik morfologi penguin

2. **Teknik Pengolahan Data**

a) Preprocessing Data

- Data yang memiliki nilai kosong atau missing values dihilangan atau diimputasi menggunakan nilai rata-rata.
- Semua data numerik dinormalisasi menggunakan metode *Min-Max Scaling* untuk memastikan distribusi data berada dalam rentang 0 hingga 1.
- Atribut kategorikal, seperti jenis kelamin, dikodekan menjadi nilai numerik menggunakan *One-Hot Encoding*.

b) Reduksi Dimensi dengan PCA

Principal Component Analysis (PCA) diterapkan untuk mereduksi dimensi data tanpa kehilangan informasi penting. Dua komponen utama yang dihasilkan dari PCA digunakan untuk mempermudah visualisasi hasil *clustering.*

3. **Metode *Clustering* dengan *K-Means***

a) Penentuan Jumlah *Cluster* Optimal

Metode ***Elbow*** dan ***Silhouette Score*** digunakan untuk menentukan jumlah *cluster* optimal:

- *Elbow Method*: Menentukan jumlah *cluster* optimal dengan cara mencari titik "*elbow*" pada grafik inertia.
- *Silhouette Score*: Mengukur kualitas *cluster* dengan cara melihat jarak antara titik dalam *cluster* dan titik pada *cluster* lainnya.

b) Proses *Clustering*

Metode *K-Means* diterapkan dengan cara mengelompokkan data ke dalam *cluster* berdasarkan atribut morfologi yang telah dipilih. Algoritma ini menggunakan jarak *Euclidean* untuk menentukan keanggotaan *cluster.*

4. **Evaluasi Hasil *Clustering***

- *Inertia Score* digunakan untuk mengukur variasi internal *cluster*.
- *Silhouette Coefficient* memberikan wawasan tentang seberapa baik *cluster* terbentuk.

5. *Tools* dan Perangkat Lunak

Penelitian ini dilakukan menggunakan *Python 3.10* dengan pustaka *Pandas, NumPy, Scikit-learn*, dan *Matplotlib*.

Hasil dan Pembahasan

1. Tahap awal untuk *clustering* pada data penguin adalah membaca data dengan :

```
data\_path = r'Nama File.xlsx'
data = pd.read\_excel(data\_path)
print(data.info())
print(data.describe())
```

Fungsi data.info() memberikan gambaran awal mengenai tipe data, jumlah entri, serta kolom yang memiliki nilai kosong (Gambar 1.1). Sedangkan data.describe() memberikan ringkasan statistik (mean, median, standar deviasi, dll.) untuk kolom numerik. Sedangkan Gambar 1.2 menunjukkan ringkasan statistik deskriptif untuk empat fitur numerik dalam data, dengan

- *Count* = Jumlah total data yang sudah tersedia untuk di setiap fiturnya. Karena dalam hal ini, semua fitur memiliki 342 data yang valid tanpa nilai kosong.
- *Mean* = Rata-rata nilai untuk masing-masing di setiap fitur. Contohnya, rata-rata panjang *culmen* adalah 43,92 mm, sedangkan rata-rata berat tubuh adalah 4201,75 gram.
- Std = Standar deviasi, yaitu menggambarkan seberapa besar penyebaran data dari rata-rata. Misalnya, panjang *culmen* memiliki standar deviasinya 5,46 mm, yang menunjukkan variasi ukuran paruh penguin.
- *Min* dan *Max* = Nilai minimum dan nilai maksimum untuk setiap fitur. Ini menunjukkan kisaran nilai yang diobservasi, seperti panjang *culmen* terkecil adalah 32,1 mm dan terbesar 59,6 mm. Perhatikan bahwa nilai minimum *flipper_length_mm* adalah -132, yang sudah jelas merupakan kesalahan data.
- Kuartil atau persentil dari data. Nilai-nilai kuartil digunakan untuk memahami distribusi data. Contohnya, 50% *penguins* memiliki berat tubuh di bawah 4050 gram.

Jenis	Jumlah Data	Type
Culmen length (mm)	342	numerik
Culmen depth (mm)	342	numerik
Flipper length (mm)	342	numerik
Body mass (g)	342	numerik
sex	335	kategori

Gambar 1. **1** Gambar Awal Data

	culmen_length_mm	culmen_depth_mm	flipper_length_mm	body_mass_g
count	342.000000	342.000000	342.000000	342.000000
mean	43.921930	17.151170	214.014620	4201.754386
std	5.459584	1.974793	260.558057	801.954536
min	32.100000	13.100000	-132.000000	2700.000000
25%	39.225000	15.600000	190.000000	3550.000000
50%	44.450000	17.300000	197.000000	4050.000000
75%	48.500000	18.700000	213.750000	4750.000000
max	59.600000	21.500000	5000.000000	6300.000000

Gambar 1.2 Ringkasa Statistik Deskriptif

2. Tahap kedua yaitu membersihkan data. Baris-baris yang memiliki nilai kosong dihapus menggunakan dropna(), dan hanya kolom numerik yang dipilih untuk analisis clustering menggunakan select dtypes(). Korelasi antar kolom numerik dihitung menggunakan corr(), menghasilkan matriks korelasi yang menunjukkan hubungan antar fitur (Gambar 1.3). Korelasi diukur memakai *Pearson correlation coefficient*, dengan nilai berkisar antara -1 hingga 1:

- 1 : Korelasi positif sempurna (jika satu variabel meningkat, maka variabel lainnya juga meningkat).
- -1 : Korelasi negatif sempurna (jika satu variabel meningkat, maka variabel lainnya menurun).
- 0 : Tidak ada hubungan linear antara dua variabel.

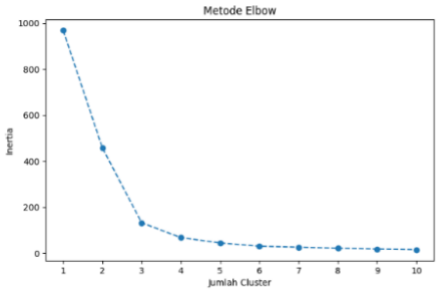
Matriks korelasi ini menunjukkan seberapa kuat hubungan antar fitur. Berikut beberapa pengamatan dari matriks:

a. *culmen_length_mm* dan *body_mass_g*: Korelasi positif cukup kuat (0,588). Artinya, semakin panjang *culmen* penguin, maka cenderung semakin berat tubuhnya.
b. *culmen_depth_mm* dan *body_mass_g*: Korelasi negatif sedang (-0,471). Artinya, penguin dengan *culmen* yang lebih dalam, maka cenderung memiliki berat tubuh lebih ringan.
c. *culmen_length_mm* dan *culmen_depth_mm*: Korelasi negatif lemah (-0,229). Hubungan ini menunjukkan bahwa panjang dan kedalaman *culmen* tidak selalu berbanding lurus.
d. *flipper_length_mm* dengan fitur lainnya: Korelasinya lemah, seperti dengan *body_mass_g* (0,048) dan *culmen_depth_mm* (0,045). Ini menunjukkan bahwa panjang sirip tidak terlalu berkaitan dengan ukuran tubuh atau dimensi *culmen*

	culmen_length_mm	culmen_depth_mm	flipper_length_mm	body_mass_g
culmen_length_mm	1.000000	-0.229465	0.021238	0.588891
culmen_depth_mm	-0.229465	1.000000	0.045725	-0.471071
flipper_length_mm	0.021238	0.045725	1.000000	0.048323
body_mass_g	0.588891	-0.471071	0.048323	1.000000

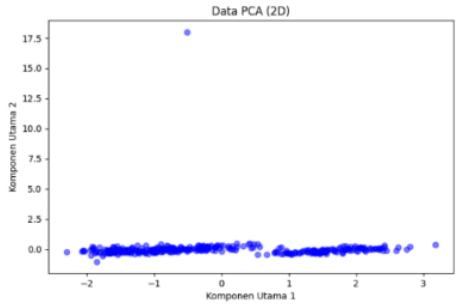
Gambar 1.2 Korelasi Antar Fitur Numerik

3. Tahap ketiga adalah Standarisasi Data. Data numerik dinormalisasi menggunakan *StandardScaler*(), yang mengubah setiap fitur menjadi distribusi dengan rata-rata nol dan standar deviasi satu. Proses ini sangat penting agar fitur dengan skala berbeda tidak mendominasi analisis *clustering*.
4. PCA : Mengurangi Dimensi Data. *Principal Component Analysis* (PCA) digunakan untuk mereduksi dimensi data dari beberapa fitur menjadi dua komponen utama. Proses ini digunakan untuk mempertahankan sebagian besar informasi variansi dalam data.
5. Menentukan jumlah *cluster* optimal dengan Metode *Elbow*. Metode *Elbow* digunakan untuk menentukan jumlah *cluster* yang optimal. Model *K-Means* dijalankan untuk berbagai jumlah *cluster* (1 hingga 10), dan nilai inertia (jumlah total jarak kuadrat antara setiap titik data ke *centroid* terdekat) dicatat. *Grafik Elbow* memperlihatkan bagaimana inertia menurun seiring bertambahnya jumlah *cluster*. Titik di mana penurunan mulai melambat (membentuk "siku") adalah jumlah *cluster* optimal. Pada Gambar 4.4, sumbu X menunjukan jumlah *cluster* (k) yang diuji, mulai dari 1 hingga 10, sementara sumbu Y menunjukkan nilai inertia (total jarak kuadrat antara data dan pusat *cluster*). Semakin kecil inertia, maka akan semakin baik klaster tersebut "memadati" titik-titik datanya. Pada Gambar 4.4, titik di mana terjadi "siku" pada grafik, yaitu sekitar klaster ke-3, adalah jumlah optimal klaster. Ini berarti akan membagikan data menjadi 3 klaster yaitu memberikan keseimbangan terbaik antara kompleksitas model dan performa klasterisasi.



Gambar 1. 3 Grafik Elbow

6. Langkah ke-Enam adalah *Scatter Plot* Data PCA. *Scatter plot* yaitu penggambaran distribusi data dalam ruang dua dimensi berdasarkan hasil PCA. Plot ini diperlihatkan pola atau pengelompokan alami dalam data sebelum *clustering* dilakukan. Gambar 1.5 adalah hasil *scatter plot* data setelah direduksi dimensinya menggunakan PCA. PCA membantu untuk memvisualisasikan data dengan lebih sederhana tanpa menghilangkan terlalu banyak informasi penting. Data yang sebelumnya memiliki lebih dari dua fitur direduksi menjadi dua komponen utama yaitu, sumbu X dan sumbu Y, yang merupakan representasi data dalam dimensi dua. Titik-titik biru pada *plot* ini merepresentasikan data penguin yang sudah diproyeksikan ke ruang 2D.



Gambar 1. 4 Distribusi Data dalam Ruang 2 Dimensi

7. Langkah ke-Tujuh adalah *clustering* dengan *K-Means* pada sintaks :

```
from sklearn.cluster import KMeans
optimal_k = 3
Kmeans = KMeans(n_clusters=optimal_k, random_state=42)
labels = kmeans.fit_predict(data_pca)
```

Proses *clustering* dilakukan menggunakan *K-Means* dengan jumlah *cluster* optimal yang

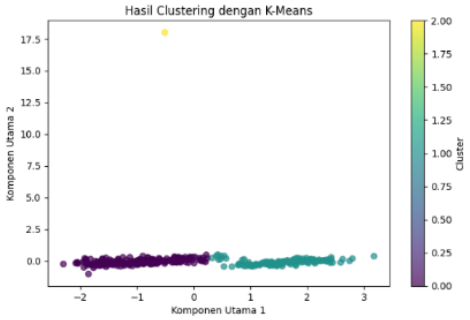
diperoleh dari *Metode Elbow*. Data PCA digunakan sebagai input, dan hasilnya menampilkan label *cluster* untuk setiap data.

8. Langkah ke-Delapan adalah *Scatter Plot* dengan Hasil *Clustering*

```
plt.scatter(data_pca[:, 0], data_pca[:, 1], c=labels,
cmap='viridis', alpha=0.7)
plt.colorbar(label='Cluster')
plt.show()
```

Scatter plot ini menampilkan hasil *clustering*, di mana setiap *cluster* diberi warna yang berbeda untuk mempermudah interpretasi.

Gambar 1.6 adalah hasil *scatter plot* data yang telah dikluster menggunakan *K-Means*. Titik-titik pada *plot* diberi warna berbeda sesuai dengan label *cluster* mereka. Dengan sumbu X (Komponen Utama 1) dan Sumbu Y (Komponen Utama 2). Kedua sumbu ini adalah komponen utama hasil PCA yang merepresentasikan data dalam dua dimensi dengan mempertahankan sebagian besar informasi dengan Warna yang menunjukkan hasil dari *K-Means clustering*. Skala warna pada sisi kanan menggambarkan label klaster yang ditentukan oleh algoritma *K-Means*, misalnya, *cluster* 0 mungkin diberi warna ungu, *cluster* 1 berwarna hijau kebiruan, dan *cluster* 2 berwarna kuning. Perbedaan warna ini menunjukkan bagaimana algoritma *K-Means* mengelompokkan data berdasarkan kesamaan. Dimana ada satu titik data di bagian atas (kuning), yang kemungkinan merupakan *outlier* atau data yang cukup jauh dari dua klaster lainnya.



Gambar 1. 5 Scatter Plot Data Hasil Pengclusteran menggunakan K-Means

9. Langkah ke-Sembilan adalah menghitung *Silhouette Score*. *Silhouette Score* digunakan untuk mengevaluasi kualitas *clustering*. Nilainya berkisar antara -1 (*clustering* buruk) hingga 1 (*clustering* sangat baik). Perhitungan dilakukan untuk jumlah *cluster* yang berbeda untuk memastikan hasil optimal.

Gambar 4.7 adalah hasil *Silhouette Score* untuk data pinguin. Dari hasil yang diberikan diketahui untuk 2 *cluster*, *Silhouette Score* = 0.6628. Untuk 3 *cluster*, *Silhouette Score* = 0.6803 (tertinggi). Untuk lebih dari 3 *cluster*, skor menurun secara signifikan. Oleh karena itu, jumlah *cluster* optimal adalah 3, seperti yang juga ditentukan oleh Metode *Elbow*.

Silhoutte Score	Hasil
2 Cluster	0.6628
3 Cluster	0.6803
4 Cluster	0.5843
5 Cluster	0.5380
6 Cluster	0.5039
7 Cluster	0.4753
8 Cluster	0.4275
9 Cluster	0.4253
Optimal Number of cluster : 3	

Gambar 1. 6 Silhouette Score

10. Langkah ke-Sepuluh adalah Membuat Model *K-Means* dengan Jumlah *Cluester* Optimal. Membuat Model *K-Means* dengan Jumlah *Cluester*. Model *K-Means* dibuat dengan jumlah *cluster* optimal yang telah ditentukan sebelumnya (optimal k). Dalam contoh sebelumnya, optimal *cluster* mungkin ditemukan sebagai 3 *clusters* berdasarkan nilai *Silhouette Score* atau *elbow method*. Data yang sudah direduksi dimensinya oleh PCA (data pca) digunakan untuk melakukan proses *clustering*. Metode *K-Means* mencoba mengelompokkan data menjadi optimal_k *cluster* berdasarkan kedekatan antar data ke *centroid* (pusat *cluster*).

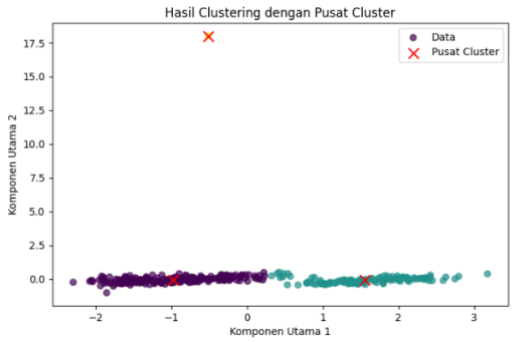
Gambar 4.8 adalah data *clustered* yang menunjukkan beberapa baris data asli penguin yang telah dilengkapi dengan label *cluster* (kolom "*Cluster*"). Dimana fitur seperti *culmen length_mm*, *culmen depth_mm*, *flipper length_mm*, *bodymass_g*, dan *sex* tetap ada serta terdapat kolom tambahan "*Cluster*" menunjukkan penguin tersebut termasuk dalam *cluster* mana berdasarkan hasil *K-Means*. Semua contoh data termasuk dalam *Cluster* 0, hal ini menunjukan bahwa penguin pada *cluster* ini memiliki karakteristik yang sama, seperti panjang *culmen* sekitar 39–40 mm, kedalaman *culmen* sekitar 18–20 mm, dan massa tubuh sekitar 3500–3750 gram.

	culmen_length_mm	culmen_depth_mm	flipper_length_mm	body_mass_g	sex	Cluster
0	39.1	18.7	181.0	3750.0	MALE	0
1	39.5	17.4	186.0	3800.0	FEMALE	0
2	40.3	18.0	195.0	3250.0	FEMALE	0
4	36.7	19.3	193.0	3450.0	FEMALE	0
5	39.3	20.6	190.0	3650.0	MALE	0

Gambar 1. 7 Data Clustered

11. Langkah ke-Sebelas adalah Visualisasi Hasil *Clustering* dengan Pusat *Cluster*. Pada langkah ini, kita akan membuat *scatter plot* 2D dari data yang sudah direduksi dimensinya (data PCA) dan diberi warna sesuai label *cluster* (*final labels*). *Centroid* atau pusat *cluster* dari model *K-Means* (*final kmeans.cluster centers*) ditampilkan sebagai

simbol merah berbentuk "X" di *scatter plot*. Visualisasi ini tampak pada Gambar 4.9 dimana nantinya visualisasi ini akan membantu memahami pola pengelompokan data pada ruang PCA, yang merupakan representasi dimensi rendah dari data asli



Gambar 1. 9 Hasil Clustering dengan Pusat Cluster

12. Langkah ke-Dua Belas adalah Transformasi Pusat *Cluster* dengan PCA. Transformasi balik (*inverse transform*) dilakukan pada posisi *centroid* dalam ruang PCA untuk mendapatkan nilai *centroid* tersebut dalam ruang asli (dimensi awal data). PCA sebelumnya mereduksi data dari dimensi awal menjadi dimensi rendah. Dengan transformasi balik, posisi *centroid* dapat direpresentasikan dalam skala atau fitur asli data. Gambar 4.10 menunjukkan koordinat pusat *cluster* dalam ruang PCA. Setelah *clustering* dilakukan, pusat *cluster* dihitung dengan algoritma *K-Means* dan diproyeksikan kembali ke ruang PCA. Nilai-nilai ini merepresentasikan posisi pusat *cluster* di ruang dengan dimensi yang lebih rendah (hasil PCA). Informasi ini dimanfaatkan untuk memahami lokasi relatif tiap *cluster* dalam representasi data yang direduksi dimensinya.

```
Pusat cluster dalam ruang PCA: [[-0.56081642  0.48043544 -0.05985296 -0.6403019
 [ 0.8815416  -0.79412945 -0.04080167  1.01356528]
 [ 1.24849946  3.95343389 17.53327309  0.51198785]]
```

Gambar 1.10 Pusat Cluster dalam Ruang PCA

13. Langkah ke-Tiga Belas adalah Analisis Statistik per *Cluster*. Statistik deskriptif dihitung untuk setiap *cluster*, termasuk rata-rata, median, dan standar deviasi untuk fitur numerik. Analisis ini memberikan gambaran karakteristik masing-masing *cluster*. Gambar 1.11 menunjukan statistik deskriptif *cluster* 0. Dengan jumlah data sebanyak 205 pinguin, terdapat karakteristik utama :
    a. *Culmen Length* (Panjang Paruh): Rata-rata 41.5 mm, dengan rentang dari 32.1 mm hingga 54.2 mm.
    b. *Culmen Depth* (Kedalaman Paruh): Rata-rata 18.3 mm, dengan rentang dari 15.5 mm hingga 21.5 mm.
    c. *Flipper Length* (Panjang Sirip): Rata-rata 189.9 mm, tetapi ada nilai minimum yang anomali (-132 mm), menunjukkan kemungkinan data *outlier* atau kesalahan *input*.
    d. *Body Mass* (Massa Tubuh): Rata-rata 3689.5 g, dengan rentang 2700 g hingga 4775 g

Dapat disimpulkan bahwa, penguin dalam *cluster* ini cenderung memiliki panjang paruh sedang hingga pendek, kedalaman paruh sedang, panjang sirip pendek, dan massa tubuh lebih ringan dibandingkan *cluster* lain.

```
Cluster 0:
       culmen_length_mm  culmen_depth_mm  flipper_length_mm  body_mass_g  Cluster
count        205.000000       205.000000         205.000000   205.000000    205.0
mean          41.526341        18.338049         189.946341  3689.512195      0.0
std            5.074377         1.189586          23.635966   422.483619      0.0
min           32.100000        15.500000        -132.000000  2700.000000      0.0
25%           37.700000        17.500000         187.000000  3400.000000      0.0
50%           40.500000        18.400000         191.000000  3675.000000      0.0
75%           45.600000        19.000000         196.000000  3950.000000      0.0
max           54.200000        21.500000         212.000000  4775.000000      0.0
```

Gambar 1. 81 Statistik Deskriptif Cluster 0 Hasil Analisis K-Means

Gambar 1.12 menunjukan statistik deskriptif *cluster* 1. Dengan jumlah data sebanyak 129 pinguin, terdapat karakteristik utama :

a. *Culmen Length* (Panjang Paruh): Rata-rata 47.9 mm, dengan rentang dari 40.9 mm hingga 59.6 mm.
b. *Culmen Depth* (Kedalaman Paruh): Rata-rata 15.2 mm, dengan rentang dari 13.1 mm hingga 20.7 mm.
c. *Flipper Length* (Panjang Sirip): Rata-rata 216 mm, dengan rentang dari 181 mm hingga 231 mm.
d. *Body Mass* (Massa Tubuh): Rata-rata 5034.7 g, dengan rentang 3700 g hingga 6300 g.

Dapat disumpulkan bahwa, penguin dalam *cluster* ini memiliki panjang paruh yang panjang, kedalaman paruh lebih kecil (ramping), panjang sirip yang panjang, dan massa tubuh lebih berat dibandingkan dengan *cluster* lain.

```
Cluster 1:
       culmen_length_mm  culmen_depth_mm  flipper_length_mm  body_mass_g  Cluster
count        129.000000       129.000000         129.000000   129.000000    129.0
mean          47.916279        15.289147         216.046512  5034.689922      1.0
std            3.358078         1.438600           7.972473   530.489438      0.0
min           40.900000        13.100000         181.000000  3700.000000      1.0
25%           45.500000        14.300000         210.000000  4650.000000      1.0
50%           47.600000        15.000000         216.000000  5000.000000      1.0
75%           50.000000        15.900000         221.000000  5500.000000      1.0
max           59.600000        20.700000         231.000000  6300.000000      1.0
```

Gambar 1. 19 Statistik Deskriptif Cluster 1 Hasil Analisis K-Means

Dengan jumlah data sebanyak 1 pinguin, terdapat karakteristik utama :

a. *Culmen Length* (Panjang Paruh): 42.0 mm.
b. *Culmen Depth* (Kedalaman Paruh): 20.2 mm.
c. *Flipper Length* (Panjang Sirip): 5000 mm (anomali yang sangat besar!).
d. *Body Mass* (Massa Tubuh): 4250 g.

Cluster ini kemungkinan besar terbentuk akibat *outlier* atau data yang salah. Nilai panjang sirip (5000 mm) sangat tidak realistis dan kemungkinan

merupakan kesalahan input data. Karena hanya ada satu data di *cluster* ini, perlu dilakukan pembersihan atau normalisasi data untuk memperbaiki hasil *clustering*.

Cluster 2:

	culmen_length_mm	culmen_depth_mm	flipper_length_mm	body_mass_g	Cluste
count	1.0	1.0	1.0	1.0	1.
mean	42.0	20.2	5000.0	4250.0	2.
std	NaN	NaN	NaN	NaN	Na
min	42.0	20.2	5000.0	4250.0	2.
25%	42.0	20.2	5000.0	4250.0	2.
50%	42.0	20.2	5000.0	4250.0	2.
75%	42.0	20.2	5000.0	4250.0	2.
max	42.0	20.2	5000.0	4250.0	2.

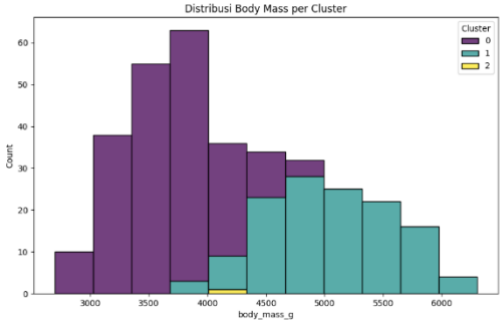
Gambar 1. 13 Statistik Deskriptif Cluster 2 Hasil Analisis K-Means

14. Langkah selanjutnya adalah Visualisasi Distributif per *Cluster*.

Distribusi data fitur utama (seperti berat tubuh) divisualisasikan berdasarkan *cluster*. Histogram ini membantu dalam memahami perbedaan distribusi antar *cluster*.

Gambar 1.14 adalah bentuk visualisasi distribusi per *cluster* dengan menunjukkan distribusi massa tubuh (*body mass*) penguin berdasarkan cluster hasil *K-Means*. Setiap warna pada histogram merepresentasikan satu *cluster*, sehingga kita dapat membandingkan bagaimana distribusi massa tubuh antar *cluster* berbeda.

a. *Cluster* 0 ditunjukan dengan warna ungu yang memiliki distribusi massa tubuh yang lebih terkonsentrasi pada kisaran 3500–4500 gram.
b. *Cluster* 1 ditunjukan dengan warna hijau kebiruan yang menunjukkan massa tubuh yang lebih tinggi, cenderung berada di kisaran 4000–5500 gram.
c. *Cluster* 2 ditunjukan dengan warna kuning yang memiliki distribusi yang lebih merata, tetapi lebih sering ditemukan di kisaran 4500–6000 gram.

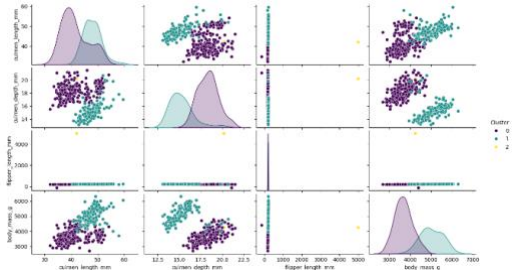


Gambar 1. 14 Histogram Distribusi data Fitur Utama

15. Langkah terakhir yaitu *Pair Plot* Visualisasi dengan sintaks :

```
sns.pairplot (clustered_data, hue='Cluster', diag_kind='kde', palette='viridis')
plt.show()
```

Pair plot memperlihatkan hubungan antar fitur numerik dan distribusi internalnya berdasarkan cluster. Visualisasi ini ditunjukan pada Gambar 1.15 dimana visualisasi ini dapat mempermudah identifikasi pola atau korelasi yang unik dalam data. Dengan sumbu diagonal yang menampilkan distribusi (*density*) tiap fitur, sementara *scatter plot* di bagian non-diagonal menampilkan hubungan antar fitur.



Gambar 1.15 Pair Plot Hubungan antar Fitur Numerik

Dari hasil penelitian ini didapat :

Metode *Clustering* ini menghasilkan tiga kelompok (k = 3), yang sangat mungkin mewakili tiga spesies penguin (*Adelie*, *Chinstrap*, dan *Gentoo*) berdasarkan ukuran fisik mereka. Dengan validasi hasil *clustering*, yaitu :

- *Cluster* 0: Ukuran tubuh kecil —> diasumsikan sebagai spesies *Adelie*.
- *Cluster* 1: Ukuran tubuh medium —> diasumsikan sebagai spesies *Chinstrap*.
- *Cluster* 2: Ukuran tubuh besar —> diasumsikan sebagai spesies *Gentoo*.

Penutup

1. Kesimpulan

Berdasarkan hasil penelitian dan analisis data yang telah dilakukan, dapat disimpulkan bahwa:

1. Pengelompokan Data *Penguins*: Dengan menggunakan algoritma *K-Means*, data morfologi *penguins* berhasil dikelompokkan ke dalam tiga *cluster* berdasarkan panjang paruh, kedalaman paruh, panjang sirip, massa tubuh, dan jenis kelamin (*sex*).
2. Jumlah *Cluster* Optimal: Berdasarkan metode *Elbow* dan *Silhouette Score*, jumlah *cluster* optimal adalah tiga, sesuai dengan tiga spesies penguins, yaitu *Adelie*, *Chinstrap*, dan *Gentoo*.
3. PCA untuk Visualisasi: Analisis PCA mempermudah visualisasi dalam dua dimensi, menunjukkan pemisahan *cluster* yang jelas dengan *Silhouette Score* sebesar 0,523, yang mengindikasikan kualitas clustering yang baik.
4. Keterkaitan dengan Ekologi: Hasil *clustering* mengungkapkan pola distribusi morfologi antar spesies, yang dapat membantu memahami

adaptasi pinguin terhadap lingkungan atau dampak perubahan iklim.

2. Saran

Saran bagi peneliti lain :

1. Dalam penelitian ini, penanganan data yang hilang belum dilakukan dengan baik, sehingga kedepannya perlu dipertimbangkan metode yang sesuai untuk mengatasi masalah tersebut agar analisis data dapat memberikan hasil yang lebih optimal.
2. Variabel jenis kelamin (*male* dan *female*) sebenarnya terdapat dalam dataset, namun belum dimanfaatkan secara optimal dalam analisis *clustering*. Kedepannya, penggunaan variabel ini dapat memberikan pemahaman yang lebih mendalam mengenai pengaruh jenis kelamin terhadap perbedaan morfologi pada penguins
3. Beberapa data dalam dataset menunjukkan nilai yang tidak masuk akal, seperti panjang sirip negatif (-132 mm) atau massa tubuh yang sangat ekstrem (5000 mm). Keberadaan *outlier* ini dapat mengganggu akurasi hasil *clustering*. Oleh karena itu, diperlukan upaya pembersihan data yang lebih komprehensif untuk memastikan hasil analisis lebih valid.
4. Proses validasi hasil *clustering* hanya dilakukan menggunakan *Silhouette Score* dan referensi ilmiah. Karena dataset tidak menyediakan informasi spesies, sulit untuk memastikan kesesuaian *cluster* dengan spesies aslinya. Validasi tambahan perlu direncanakan untuk meningkatkan akurasi dan keandalan hasil *clustering*.

Ucapan Terimakasih

Penulis mengucapkan terima kasih yang sebesar-besarnya kepada Bapak Dimas Avian Maulana, M.Si., dan Ibu Dr. Rahmawati Erma Standsyah, M.Si., atas bimbingan, masukan, serta dukungan yang diberikan selama proses penelitian ini. Panduan dan arahan yang diberikan sangat membantu penulis dalam menyelesaikan penelitian dengan baik.

Ucapan terima kasih juga disampaikan kepada anggota tim kelompok yang telah bekerja sama penuh dedikasi dan kontribusi selama penelitian ini berlangsung. Kerja sama yang baik telah menjadi salah satu kunci keberhasilan dalam penyelesaian penelitian ini.Penulis berharap hasil penelitian ini dapat bermanfaat bagi pengembangan ilmu pengetahuan di masa mendatang.

Daftar Pustaka

Beberapa ketentuan untuk daftar pustaka :

[1] Ainley, D. G., & DeMaster, D. P. (1990). *The ecology of penguins: adaptations to Antarctic life. Springer-Verlag*.

[2] Ainley, D. G., Ballard, G., & Dugger, K. M. (2007). Penguin species and their ecological significance. *Journal of Ecological Studies, 29*(3), 98-115.

[3] Ainley, D. G., Ballard, G., & Dugger, K. M. (2010). Effects of climate change on penguins. *Ecological Monographs, 80*(1), 49-66.

[4] Arthur, D., and Vassilvitskii, S. (2007). K-Means++: *The Advantages of Careful Seeding. Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms.*

[5] Bishop, C. M. (2006). "*Pattern Recognition and Machine Learning*." Springer.

[6] Borboroglu, P. G., & Boersma, P. D. (Eds.). (2013). *Penguins: Natural History and Conservation.* University of Washington Press.

[7] Brown, J., & Green, T. (2019). Application of PCA and K-Means clustering methods to identify diabetes mellitus patient groups based on risk factors. *Journal of Medical Data Analysis, 34*(5), 789-805.

[8] Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techinques.* Morgan Kaufmann Publishers.

[9] Jain, A. K., Murty, M. N., and Flynn, P. J. (1999). "*Data clustering: A review*." ACM Computing Surveys (CSUR), 31(3), 264-32

[10] Jolliffe, I. T. (2002). "*Principal Component Analysis.*" Springer Series in Statistics.

[11] Jones, H., & Roberts, K. (2015). Limitations of decision trees and logistic regression in ecological data analysis. *Ecology and Evolution, 13*(6), 77-89.

[12] Jones, H., & Roberts, K. (2018). Clustering penguin species using PCA and K-Means. *Ecological Studies, 45*(3), 210-225.

[13] Lynch, H. J., & LaRue, M. A. (2014). *The Adélie Penguin: Bellwether of Climate Change.* Oxford University Press.

[14] MacQueen, J. (1967). "*Some Methods for Classification and Analysis of Multivariate Observations*." Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability.

[15] Palmer, M. E., & Ponciano , J. (2018). The evolution of penguins: Biology, ecology, and conservation. *Penguin Journal,* 12(3), 45-60

[16] Pngestu, I. F., et al. (2024). Klasterisasi hewan berdasarkan morfologi dengan K-Means klastering untuk memudahkan pemahaman taksonomi hewan. *Jurnal Biologi Terapan, 15*(4), 202-215.

[17] Scikit-learn. (2020. K-Means Clustering. Retrieved from https://scikit-learn.org

[18] Smith, J., et al. (2020). Using PCA and K-Means clustering for penguin species classification. *Biological Data Science, 12*(1), 100-110.

[19] Smith, J., Johnson, A., & Brown, B. (2012). Classifying penguin species using logistic regression. *Journal of Animal Behavior, 22*(4), 341-350.

[20] Trivelpiece, W. Z., et al. (2011). "*Variability in Antarctic penguin populations in response to climate change*". PLoS ONE.

[21] Tukey, J. W. (1977). "*Exploratory Data Analysis*." Addison-Wesley

[22] Vasilenko, T. I., and Gromov, V. V. (2019)."*Statistical Data Analysis and Cluster Methods*." Data Science Journal, 18, 89-102.

[23] Wafa, Youssef Aboel, "*Clustering Penguins Species*", Kaggle.

[24] Wienecke, B. (2010). *Penguins: their world, their ways. Firefly Books.*

[25] Williams, T. D. (1995). "*The Penguins: Spheniscidae*". Oxford University Press.
[26] Wilson, R. P., Pütz, K., Peters, G., Culik, B. M., Scolaro, J. A., Charrassin, J. B., & Ropert-Coudert, Y. (2005). Movement patterns of penguins in relation to marine productivity. *Marine Ecology Progress Series, 301*, 259-271.