Evaluasi Model Clustering

# Konsep Dasar:

Evaluasi Model Clustering lebih sulit daripada evaluasi supervised learning karena tidak ada label kebenaran (Ground Truth) yang bersedia. Namun, ada beberapa metrik yang dapat digunakan untuk mengukur kualitas clustering:

* **Silhouette Score:** Mengukur seberapa mirip suatu titik data dengan clustering sendiri dibandingkan dengan cluster lain. Nilai yang lebih tinggi menunjukkan clustering yang lebih baik.
* **Calinski-Harabasz Index:** Mengukur rasio antara varians antar cluster dan varians dalam cluster. Nilai yang lebih tinggi menunjukkan clustering yang lebih baik.
* **Davies-Bouldin Index:** Mengukur rata-rata kemiripan antara setiap cluster dengan tetangga terdekatnya. Nilai yang lebih rendah menunjukkan clustering yang lebih baik.

**Contoh Kode (Scikit-learn)**

from sklearn.metrics import silhouette\_score, calinski\_harabasz\_score, davies\_bouldin\_score

# ... (kode untuk membaca, mempersiapkan dataset, dan melakukan clustering)

# Evaluasi

silhouette = silhouette\_score(X, labels)

calinski\_harabasz = calinski\_harabasz\_score(X, labels)

davies\_bouldin = davies\_bouldin\_score(X, labels)

print(f'Silhouette Score: {silhouette:.2f}')

print(f'Calinski-Harabasz Index: {calinski\_harabasz:.2f}')

print(f'Davies-Bouldin Index: {davies\_bouldin:.2f}')

**Tugas:**

1. Evaluasi hasil clustering yang Anda dapatkan pada Hari 36, 37, dan 38 menggunakan Silhouette Score, Calinski-Harabasz Index, dan Davies-Bouldin Index.
2. Bandingkan kinerja ketiga algoritma clustering (K-Means, Hierarchical Clustering, DBSCAN) pada dataset yang berbeda.

**Selamat Mengerjakan Tugas! 😀**

**Tugas Hari 39:**

1. **Evaluasi Hasil Clustering:**

**import numpy as np**

**import pandas as pd**

**import matplotlib.pyplot as plt**

**from sklearn import datasets**

**from sklearn.cluster import KMeans, AgglomerativeClustering, DBSCAN**

**from sklearn.decomposition import PCA**

**from sklearn.metrics import silhouette\_score, calinski\_harabasz\_score, davies\_bouldin\_score, adjusted\_rand\_score**

**# Fungsi untuk mengevaluasi clustering**

**def evaluate\_clustering(labels):**

**silhouette = silhouette\_score(X, labels)**

**calinski\_harabasz = calinski\_harabasz\_score(X, labels)**

**davies\_bouldin = davies\_bouldin\_score(X, labels)**

**return silhouette, calinski\_harabasz, davies\_bouldin**

**# Clustering dengan KMeans**

**kmeans = KMeans(n\_clusters=3, random\_state=0)**

**kmeans\_labels = kmeans.fit\_predict(X)**

**# Clustering dengan Agglomerative Clustering**

**agg\_clustering = AgglomerativeClustering(n\_clusters=3)**

**agg\_labels = agg\_clustering.fit\_predict(X)**

**# Clustering dengan DBSCAN**

**dbscan = DBSCAN(eps=0.5, min\_samples=5)**

**dbscan\_labels = dbscan.fit\_predict(X)**

**# Evaluasi hasil clustering**

**kmeans\_eval = evaluate\_clustering(kmeans\_labels)**

**agg\_eval = evaluate\_clustering(agg\_labels)**

**dbscan\_eval = evaluate\_clustering(dbscan\_labels)**

1. **Bandingkan Kinerja Ketiga Algoritma:**

**# Menampilkan hasil evaluasi**

**print("KMeans Clustering Evaluation:")**

**print(f'Silhouette Score: {kmeans\_eval[0]:.2f}')**

**print(f'Calinski-Harabasz Index: {kmeans\_eval[1]:.2f}')**

**print(f'Davies-Bouldin Index: {kmeans\_eval[2]:.2f}')**

**print("\nAgglomerative Clustering Evaluation:")**

**print(f'Silhouette Score: {agg\_eval[0]:.2f}')**

**print(f'Calinski-Harabasz Index: {agg\_eval[1]:.2f}')**

**print(f'Davies-Bouldin Index: {agg\_eval[2]:.2f}')**

**print("\nDBSCAN Clustering Evaluation:")**

**print(f'Silhouette Score: {dbscan\_eval[0]:.2f}')**

**print(f'Calinski-Harabasz Index: {dbscan\_eval[1]:.2f}')**

**print(f'Davies-Bouldin Index: {dbscan\_eval[2]:.2f}')**