Tugas 2 Pembelajaran Mesin

Timothy Hosia Budianto 5025211098

Soal :

Download dataset Customer Personality dari Kaggle pada link berikut:
https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis/data

1. Lakukan penerapan metode clustering k-Means dan analisis jumlak k yang optimal menggunakan Bahasa pemrograman python
2. Lakukan penerapan metode clustering Agglomerative Hierarchical dengan pendekatan Single-link, Complete-link, Average-link, dan Ward menggunakan Bahasa pemrograman python
3. Bandingkan hasil clustering pada poin 1 dan 2 menggunakan metode evaluasi Silhouette Score.

Hasil dan Jawaban:

```
## clustering k-Means
Silhouette Score for k=2: 0.2532
Silhouette Score for k=3: 0.2241
Silhouette Score for k=4: 0.1799
Silhouette Score for k=5: 0.1048
Silhouette Score for k=6: 0.0921
Silhouette Score for k=7: 0.0988
Silhouette Score for k=8: 0.1050
Silhouette Score for k=9: 0.1129
Silhouette Score for k=10: 0.1246

Optimal k based on silhouette score: 2

## clustering Agglomerative Hierarchical
Agglomerative Clustering with single linkage: Silhouette Score = 0.7440
Agglomerative Clustering with complete linkage: Silhouette Score = 0.7440
Agglomerative Clustering with average linkage: Silhouette Score = 0.7440
Agglomerative Clustering with ward linkage: Silhouette Score = 0.2112
```

1. Hasil Clustering dengan k-Means

- Dilakukan clustering dengan mencoba berbagai nilai k (jumlah cluster) dari 2 hingga 10.
- Optimal k yang ditemukan adalah k=2, karena memiliki nilai Silhouette Score tertinggi = 0.2532.
- Penurunan Silhouette Score untuk nilai k yang lebih besar menunjukkan bahwa cluster menjadi semakin tidak terdefinisi dengan baik ketika jumlah cluster bertambah.
- Hasilnya menunjukkan bahwa k-Means memberikan clustering terbaik dengan 2 cluster, tetapi kualitas pemisahannya masih relatif rendah (Silhouette Score = 0.2532).

---

2. Hasil Clustering dengan Agglomerative Hierarchical Clustering

Dilakukan untuk k=2 (jumlah cluster optimal berdasarkan k-Means) dengan 4 metode linkage yang berbeda:

- Single linkage: Silhouette Score = 0.7440
- Complete linkage: Silhouette Score = 0.7440
- Average linkage: Silhouette Score = 0.7440
- Ward linkage: Silhouette Score = 0.2112

Interpretasi Hasil Agglomerative Clustering:

- Single, Complete, dan Average linkage menghasilkan Silhouette Scores yang sangat tinggi (0.7440), menunjukkan bahwa clustering menghasilkan pemisahan cluster yang sangat baik.
- Ward linkage memiliki performa yang relatif lebih rendah (Silhouette Score = 0.2112), yang lebih buruk dibandingkan dengan k-Means hasilnya (Silhouette Score = 0.2532). Namun, ini juga berarti Ward linkage tidak optimal untuk kasus ini.

---

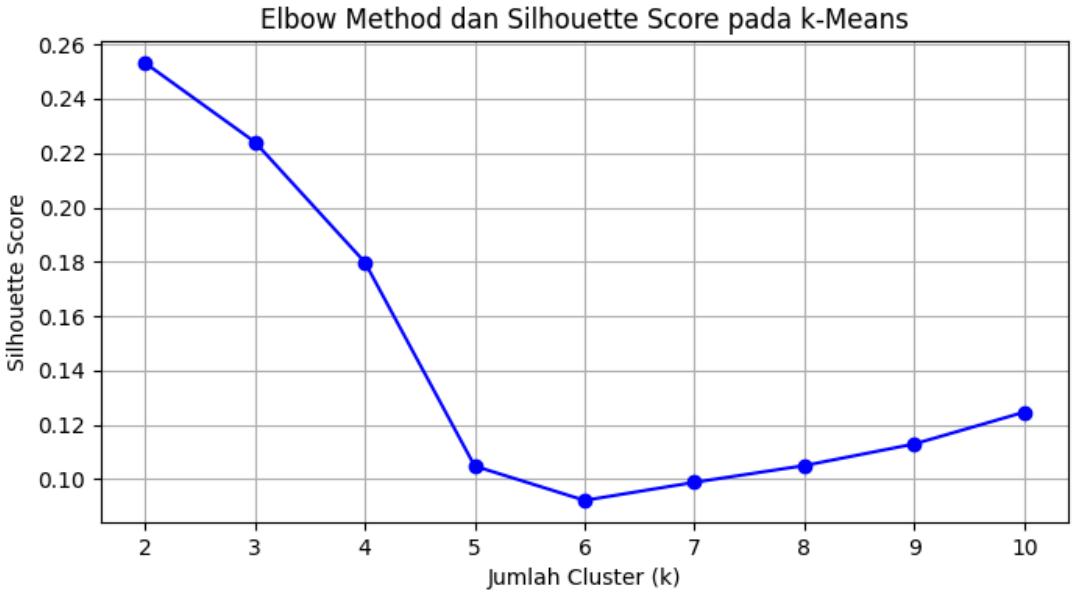
3. Kesimpulan Perbandingan:

1. Metode Agglomerative Hierarchical Clustering (Single, Complete, dan Average Linkage) memberikan hasil yang jauh lebih baik dibandingkan k-Means, dengan Silhouette Score 0.7440 yang menunjukkan pemisahan cluster yang sangat baik.
2. Ward Linkage menghasilkan performa lebih buruk dibandingkan k-Means, sehingga tidak optimal untuk kasus ini.
3. Berdasarkan Silhouette Score, Agglomerative Clustering (Single, Complete, Average Linkage) adalah metrik terbaik untuk clustering dataset ini.

Kesimpulan Akhir

- Kinerja terbaik diperoleh dari Agglomerative Hierarchical Clustering dengan Single, Complete, dan Average Linkage, yang menghasilkan clustering dengan kualitas pemisahan cluster yang kuat (Silhouette Score = 0.7440).
- Berdasarkan evaluasi Silhouette Score, Agglomerative Clustering lebih unggul dibandingkan dengan k-Means dalam hal kualitas clustering untuk dataset ini.

Log :




```
Informasi dataset:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2240 entries, 0 to 2239
Data columns (total 29 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   ID                   2240 non-null   int64
 1   Year_Birth           2240 non-null   int64
 2   Education            2240 non-null   object
 3   Marital_Status       2240 non-null   object
 4   Income               2216 non-null   float64
 5   Kidhome              2240 non-null   int64
 6   Teenhome             2240 non-null   int64
 7   Dt_Customer          2240 non-null   object
 8   Recency              2240 non-null   int64
 9   MntWines             2240 non-null   int64
 10  MntFruits            2240 non-null   int64
 11  MntMeatProducts      2240 non-null   int64
 12  MntFishProducts      2240 non-null   int64
 13  MntSweetProducts     2240 non-null   int64
 14  MntGoldProds         2240 non-null   int64
 15  NumDealsPurchases    2240 non-null   int64
 16  NumWebPurchases      2240 non-null   int64
 17  NumCatalogPurchases  2240 non-null   int64
 18  NumStorePurchases    2240 non-null   int64
 19  NumWebVisitsMonth    2240 non-null   int64
 20  AcceptedCmp3         2240 non-null   int64
```

```
 21  AcceptedCmp4         2240 non-null   int64
 22  AcceptedCmp5         2240 non-null   int64
 23  AcceptedCmp1         2240 non-null   int64
 24  AcceptedCmp2         2240 non-null   int64
 25  Complain             2240 non-null   int64
 26  Z_CostContact        2240 non-null   int64
 27  Z_Revenue            2240 non-null   int64
 28  Response             2240 non-null   int64
dtypes: float64(1), int64(25), object(3)
memory usage: 507.6+ KB
None
Silhouette Score untuk k=2: 0.2532
Silhouette Score untuk k=3: 0.2241
Silhouette Score untuk k=4: 0.1799
Silhouette Score untuk k=5: 0.1048
Silhouette Score untuk k=6: 0.0921
Silhouette Score untuk k=7: 0.0988
Silhouette Score untuk k=8: 0.1050
Silhouette Score untuk k=9: 0.1129
Silhouette Score untuk k=10: 0.1246

Optimal k based on silhouette score: 2

Agglomerative Clustering with single linkage: Silhouette Score = 0.7440
Agglomerative Clustering with complete linkage: Silhouette Score = 0.7440
Agglomerative Clustering with average linkage: Silhouette Score = 0.7440
Agglomerative Clustering with ward linkage: Silhouette Score = 0.2112
```