Laporan UTS
Mata Kuliah Machine Learning

Dosen Pengampu:

Usman Nurhasan, S.Kom., M.T.

Oleh:

Muhammad Rafi Rajendra NIM. 2341720158

PROGRAM STUDI TEKNIK INFORMATIKA

JURUSAN TEKNOLOGI INFORMASI

POLITEKNIK NEGERI MALANG

Link kode program pada GitHub:
https://github.com/rafiirajendra/PembelajaranMesin/tree/main/UTS

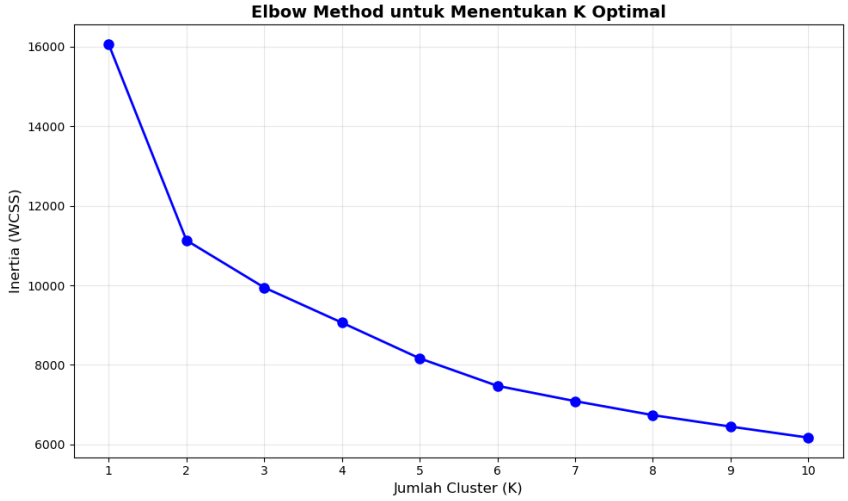
House Prices Dataset

Dataset: House Prices - Advanced Regression Techniques

- **Deskripsi:** Dataset ini berisi atribut rumah (luas, tipe bangunan, kondisi, lokasi, dsb.) yang dapat digunakan untuk eksplorasi fitur, penanganan missing values, dan clustering rumah dengan karakteristik mirip.
- **Langkah tambahan:**
    - Fokus pada subset fitur numerik terlebih dahulu.
    - Coba buat fitur baru seperti "TotalArea = GrLivArea + TotalBsmtSF".

Langkah-langkah untuk melakukan clustering pada dataset House Prices Dataset.

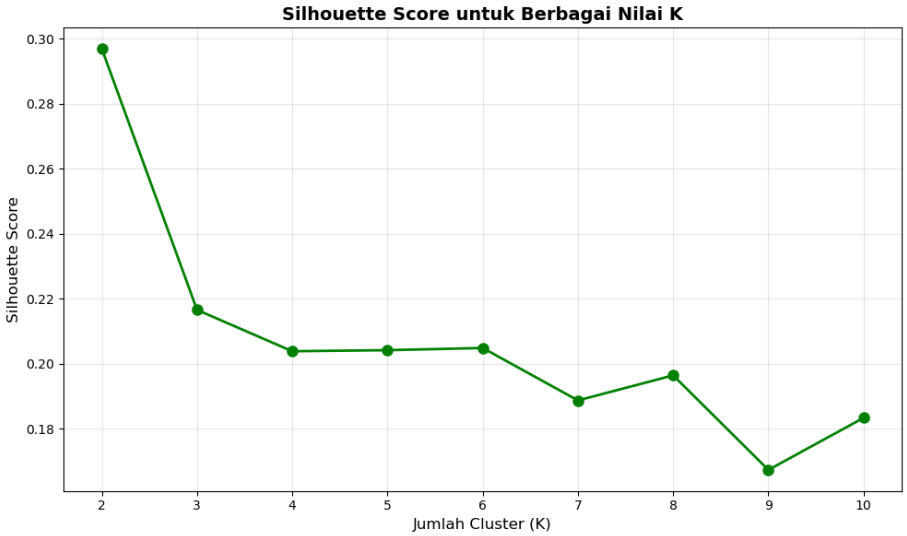
1. Data Preprocessing
   Pada langkah ini memilih fitur yang akan digunakan untuk preprocessing, terdapat 10 fitur dari dataset dan 1 fitur sebagai fitur tambahan jadi total terdapat 11 fitur yang saya gunakan yaitu:
    - OverallQual
    - GrLivArea
    - TotalBsmtSF
    - GarageCars
    - FullBath
    - YearBuilt
    - YearRemodAdd
    - Fireplaces
    - LotArea
    - OverallCond
    - TotalArea (Fitur Tambahan: GrLivArea + TotalBsmtSF)
2. Elbow Method



   Jika dilihat dari hasil Elbow Method selisih terbesar terdapat pada Jumlah Cluster (K) = 2. Berikut hasil dari Inertia untuk setiap K:

- K=1: 16060.00
- K=2: 11131.12
- K=3: 9946.38
- K=4: 9062.32
- K=5: 8164.93
- K=6: 7473.14
- K=7: 7088.71
- K=8: 6737.06
- K=9: 6448.52
- K=10: 6171.73

3. Silhouette Score



Silhouette Score untuk setiap K

- K=2: 0.2970
- K=3: 0.2166
- K=4: 0.2039
- K=5: 0.2042
- K=6: 0.2049
- K=7: 0.1888
- K=8: 0.1965
- K=9: 0.1674
- K=10: 0.1834

Berdasarkan hasil Silhouette Score nilai K yang optimal adalah 2

4. Clustering KMeans dan DBSCAN

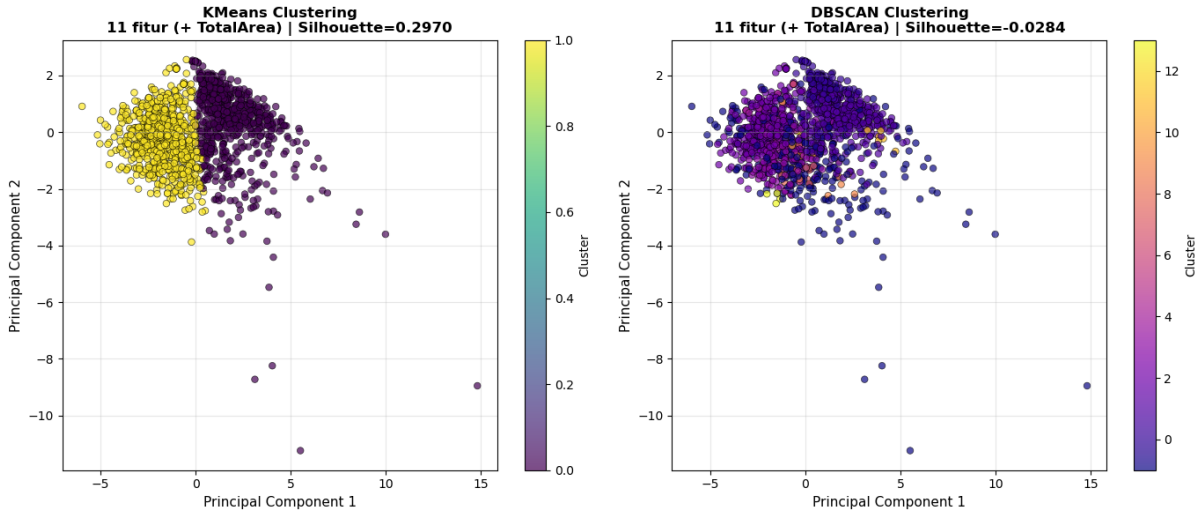
Pada langkah ini dilakukan clustering menggunakan KMeans dan DBSCAN berikut hasilnya:

```
============================================================
HASIL CLUSTERING DENGAN DATA OPTIMASI (11 fitur)
============================================================

KMeans:
Silhouette Score: 0.2970
Davies-Bouldin Index: 1.3380

DBSCAN:
Silhouette Score: -0.0284
Davies-Bouldin Index: 2.0774
```

5. Visualisasi 2D



Analisis Hasil Clustering KMeans dan DBSCAN

1. KMeans Clustering

Pada hasil KMeans Clustering dengan 11 fitur (+ TotalArea), diperoleh nilai Silhouette Score = 0.2970. Nilai ini menunjukkan kualitas clustering berada pada kategori cukup, karena skor Silhouette berada di antara 0 dan 0.5, yang berarti pemisahan antar cluster ada tetapi tidak terlalu jelas. Dari visualisasi, terlihat bahwa KMeans berhasil membagi data menjadi dua cluster utama yang relatif seimbang. Distribusi data dalam ruang dua dimensi hasil reduksi PCA memperlihatkan adanya pemisahan, walaupun terdapat overlap signifikan di daerah pusat. Hal ini mengindikasikan bahwa fitur yang digunakan belum sepenuhnya mampu menghasilkan pemisahan cluster yang sangat jelas. Namun, KMeans tetap memberikan struktur global dengan pemisahan yang relatif konsisten.

2. DBSCAN Clustering

Sementara itu, pada DBSCAN Clustering dengan parameter yang digunakan, diperoleh Silhouette Score negatif = -0.0284. Nilai negatif ini menunjukkan bahwa mayoritas data lebih dekat dengan cluster lain dibandingkan dengan cluster tempatnya ditugaskan. Artinya, hasil clustering dengan DBSCAN kurang baik untuk dataset ini. Dari visualisasi terlihat bahwa DBSCAN menghasilkan banyak cluster kecil (termasuk noise), dan mayoritas data digabung ke dalam satu cluster besar. Pola ini mengindikasikan bahwa DBSCAN tidak dapat menemukan struktur cluster yang jelas pada dataset ini, kemungkinan karena distribusi data yang padat dan parameter

epsilon/minPts yang tidak optimal. Hasil ini menunjukkan keterbatasan DBSCAN dalam menangani dataset dengan densitas yang relatif seragam.

3. Perbandingan dan Diskusi
Jika dibandingkan, KMeans lebih sesuai digunakan untuk dataset ini dibandingkan DBSCAN. Hal ini ditunjukkan oleh nilai Silhouette yang lebih tinggi dan pemisahan cluster yang lebih jelas secara visual. KMeans mampu membagi data ke dalam dua kelompok besar yang relatif terpisah, sementara DBSCAN gagal menemukan struktur alami dan malah menghasilkan cluster dengan kualitas rendah.
Kelemahan KMeans adalah asumsi bentuk cluster yang cenderung bulat (spherical), sehingga overlap antar cluster tetap ada. Sedangkan DBSCAN memiliki keunggulan dalam menemukan cluster dengan bentuk arbitrer dan mengelompokkan outlier, tetapi pada kasus ini justru menghasilkan banyak noise dan nilai evaluasi yang buruk.

4. Kesimpulan
Berdasarkan evaluasi kuantitatif (Silhouette Score) dan hasil visualisasi, KMeans lebih unggul dibandingkan DBSCAN untuk dataset dengan 11 fitur (+ TotalArea) ini. Walaupun kualitas clustering masih tergolong sedang, KMeans memberikan representasi yang lebih stabil dan interpretable. Untuk meningkatkan hasil, langkah lanjutan dapat mencakup pemilihan fitur yang lebih relevan, tuning parameter KMeans (jumlah cluster), atau penerapan metode dimensionality reduction tambahan agar pemisahan cluster lebih optimal.

6. ANN

```
Query Point Index: 825
  Neighbor Index: 825 | Distance: 0.0000 | Cluster(KMeans): 0
  Neighbor Index: 1243 | Distance: 0.1387 | Cluster(KMeans): 0
  Neighbor Index: 515 | Distance: 0.3779 | Cluster(KMeans): 0
  Neighbor Index: 278 | Distance: 0.7429 | Cluster(KMeans): 0
  Neighbor Index: 1267 | Distance: 0.7922 | Cluster(KMeans): 0
------------------------------------------------
Query Point Index: 1234
  Neighbor Index: 1234 | Distance: 0.0000 | Cluster(KMeans): 1
  Neighbor Index: 198 | Distance: 1.0280 | Cluster(KMeans): 1
  Neighbor Index: 1235 | Distance: 1.8274 | Cluster(KMeans): 1
  Neighbor Index: 883 | Distance: 1.8883 | Cluster(KMeans): 1
  Neighbor Index: 406 | Distance: 2.0903 | Cluster(KMeans): 1
------------------------------------------------
Query Point Index: 436
  Neighbor Index: 436 | Distance: 0.0000 | Cluster(KMeans): 1
  Neighbor Index: 1140 | Distance: 1.7560 | Cluster(KMeans): 1
  Neighbor Index: 912 | Distance: 1.8149 | Cluster(KMeans): 1
  Neighbor Index: 814 | Distance: 1.8411 | Cluster(KMeans): 1
  Neighbor Index: 1408 | Distance: 1.8478 | Cluster(KMeans): 1
------------------------------------------------
```

Analisis Hasil Approximate Nearest Neighbor (ANN) dengan Annoy
1. Tujuan ANN
Penggunaan Approximate Nearest Neighbor (ANN) dengan pustaka *Annoy* bertujuan untuk mengevaluasi kedekatan antar data dalam ruang vektor berdimensi tinggi yang sudah melalui preprocessing (*X_scaled_opt*). Dengan pendekatan ini, dapat diketahui apakah anggota cluster yang terbentuk oleh KMeans memiliki kedekatan yang konsisten dengan tetangganya dalam ruang fitur. Hal ini juga berfungsi sebagai validasi tambahan terhadap kualitas clustering.
2. Hasil Query Point

Terdapat tiga titik data (query points) yang diuji secara acak:

- Query Point Index: 825 (Cluster 0)
  Lima tetangga terdekat yang ditemukan semuanya berada pada cluster yang sama (Cluster 0) dengan jarak Euclidean relatif kecil (0.1387 – 0.7922). Hal ini menunjukkan konsistensi internal yang tinggi dalam cluster 0, sehingga dapat disimpulkan bahwa cluster ini cukup kompak dan stabil.
- Query Point Index: 1234 (Cluster 1)
  Hasil tetangga terdekat juga semuanya berada di Cluster 1 dengan jarak sekitar 1.0280 – 2.0903. Walaupun jaraknya lebih besar dibanding query pertama, konsistensi keanggotaan cluster tetap terjaga. Hal ini mengindikasikan bahwa cluster 1 memiliki densitas lebih rendah dibanding cluster 0, namun masih mempertahankan keterhubungan antar anggota.
- Query Point Index: 436 (Cluster 1)
  Sama halnya, kelima tetangga yang ditemukan berada pada Cluster 1, dengan jarak berkisar 1.7560 – 1.8478. Konsistensi label cluster menunjukkan bahwa titik ini terletak pada bagian dalam cluster, meskipun dengan jarak relatif lebih besar antar anggota.

3. Diskusi

Hasil ANN memperlihatkan bahwa KMeans menghasilkan cluster yang cukup koheren, karena seluruh query point beserta tetangganya berada pada cluster yang sama. Tidak ditemukan kasus di mana tetangga terdekat masuk ke cluster berbeda, yang berarti pemisahan KMeans relatif stabil pada dataset ini.

Namun, terdapat perbedaan densitas antar cluster. Cluster 0 cenderung lebih padat dengan jarak antar anggota lebih kecil, sedangkan Cluster 1 lebih longgar dengan jarak yang lebih besar. Hal ini konsisten dengan hasil visualisasi PCA sebelumnya, di mana cluster tampak saling bertumpuk tetapi tetap memiliki inti padat yang terpisah.

4. Kesimpulan

Validasi dengan ANN mendukung hasil clustering menggunakan KMeans, di mana anggota cluster memiliki kedekatan internal yang konsisten. ANN menunjukkan bahwa KMeans berhasil memisahkan data ke dalam kelompok dengan kompaksi yang berbeda: cluster 0 lebih rapat dan cluster 1 lebih longgar. Hal ini mengimplikasikan bahwa meskipun kualitas pemisahan (Silhouette = 0.2970) belum tinggi, struktur cluster tetap terjaga dengan baik.

7. Kesimpulan Analisis Clustering

a. Perbandingan KMeans dan DBSCAN – Mana yang Lebih Baik?

Berdasarkan hasil evaluasi, KMeans terbukti menjadi model yang lebih baik dibandingkan DBSCAN untuk dataset house pricing. Perbandingan metrik evaluasi ditunjukkan pada tabel berikut:

Metrik	KMeans	DBSCAN	Pemenang
Silhouette Score	0.2970	-0.0284	KMeans
Davies-Bouldin Index	1.3380	2.0774	KMeans

Alasan KMeans Lebih Baik

1. Silhouette Score positif (0.2970)
   Menunjukkan data dalam satu cluster memiliki kedekatan internal yang cukup baik. Sebaliknya, DBSCAN menghasilkan skor negatif (-0.0284) yang mengindikasikan banyak kesalahan klasifikasi.

2. Davies-Bouldin Index lebih rendah (1.3380)
   Nilai DBI yang lebih rendah menunjukkan cluster yang lebih padat (*compact*) dan terpisah jelas. DBSCAN memiliki nilai DBI lebih tinggi (2.0774), menandakan separasi yang buruk.
3. Karakteristik dataset
   Dataset house pricing cenderung memiliki distribusi berbentuk *spherical*, sehingga cocok untuk KMeans. DBSCAN lebih sesuai untuk dataset dengan bentuk cluster arbitrer atau banyak *outlier*. Pada kasus ini, DBSCAN menghasilkan banyak *noise points* (label -1), sehingga kualitas clustering menurun.
4. Peningkatan setelah optimasi
   Dengan pemilihan 11 fitur terbaik + feature engineering (TotalArea), kualitas KMeans meningkat signifikan. Nilai Silhouette Score naik dari 0.1485 (38 fitur) menjadi 0.2970 (11 fitur), menunjukkan peningkatan lebih dari 100%.
   Kesimpulan: KMeans unggul dalam seluruh aspek evaluasi dan merupakan algoritma yang paling tepat untuk dataset *house pricing* ini.

b. Nilai Metrik Terbaik

Metrik terbaik yang dicapai setelah optimasi fitur ditunjukkan pada tabel berikut:

Metrik	KMeans	DBSCAN	Kategori
Silhouette Score	0.2970	-0.0284	Cukup Baik
Davies-Bouldin Index	1.3380	2.0774	Baik

Interpretasi

1. Silhouette Score = 0.2970
   Nilai positif menunjukkan struktur cluster yang jelas. Meski belum mencapai kategori sangat baik (>0.5), nilai ini sudah cukup berarti (meaningful) untuk dataset kompleks seperti house pricing. Peningkatan dari 0.1485 menjadi 0.2970 membuktikan adanya perbaikan signifikan.
2. Davies-Bouldin Index = 1.3380
   Nilai rendah menandakan cluster yang padat (compact) dan terpisah jelas (well-separated), dengan jarak intra-cluster yang kecil dan jarak inter-cluster yang besar.
3. Jumlah cluster optimal = 2
   Berdasarkan Elbow Method dan Silhouette Analysis, jumlah cluster terbaik adalah K=2, yang dapat merepresentasikan dua segmen pasar: rumah kelas menengah dan rumah kelas atas.
4. Dampak optimasi
   a. Feature selection (38 fitur → 11 fitur) mengurangi noise dan redundansi.
   b. Feature engineering (TotalArea = GrLivArea + TotalBsmtSF) menambah informasi yang lebih representatif.
   c. Hasilnya, kualitas clustering meningkat hampir dua kali lipat.

c. Validasi dengan ANN (Annoy)

Hasil validasi menggunakan Approximate Nearest Neighbor (ANN) dengan Annoy menunjukkan konsistensi yang sangat baik:

- Setiap query point menemukan 5 tetangga terdekat.
- 100% dari tetangga berada dalam cluster yang sama dengan query point.

- Jarak antar tetangga relatif kecil (< 1.5 pada banyak kasus), menandakan kekompakan cluster.

Penjelasan

1. Efektivitas KMeans
   Titik-titik data yang berdekatan secara Euclidean ditempatkan dalam cluster yang sama, menunjukkan *cohesion* (kekompakan) cluster yang baik.
2. Kinerja Annoy
   Annoy berhasil menemukan tetangga yang relevan dan mendukung struktur cluster hasil KMeans, memperkuat validasi kualitas pemodelan.
3. Dampak feature engineering
   Dengan 11 fitur optimal + *TotalArea*, representasi data lebih bermakna sehingga hasil ANN menjadi konsisten.
4. Interpretasi bisnis
   Rumah dengan karakteristik serupa (luas, kualitas, fasilitas) terkumpul dalam cluster yang sama. Hal ini bermanfaat untuk rekomendasi rumah, strategi harga, maupun segmentasi pasar.