Laporan Analisis Ujian Tengah Semester Machine Learning

Oleh:

Aryo Adi Putro **NIM: 2341720084**

PROGRAM STUDI TEKNIK INFORMATIKA

JURUSAN TEKNOLOGI INFORMASI

POLITEKNIK NEGERI MALANG

2025

Laporan Analisis Dataset (ringkasan)

1) Penjelasan singkat dataset

- Jumlah sampel (total): **1460**

- Jumlah kolom: **81**
- Contoh fitur teratas digunakan untuk clustering (dipilih berdasarkan varians, 12 fitur): `sum_SalePrice_LotArea, SalePrice, LotArea, MiscVal, GrLivArea, BsmtFinSF1, BsmtUnfSF, 2ndFlrSF, 1stFlrSF, TotalBsmtSF` (selengkapnya ada di file ringkasan JSON).

2) Proses preprocessing

- Menghapus duplikasi jika ada.
- Imputasi nilai hilang: **numerik -> median**, **kategori -> modus**.
- Encoding: one-hot untuk kategori dengan kardinalitas kecil (<=6).
- Pembuatan fitur baru: `sum_SalePrice_LotArea` (jumlah dari dua fitur numerik bervarians tertinggi).
- Normalisasi: StandardScaler pada fitur numerik terpilih sebelum clustering.

3) Hasil clustering

KMeans

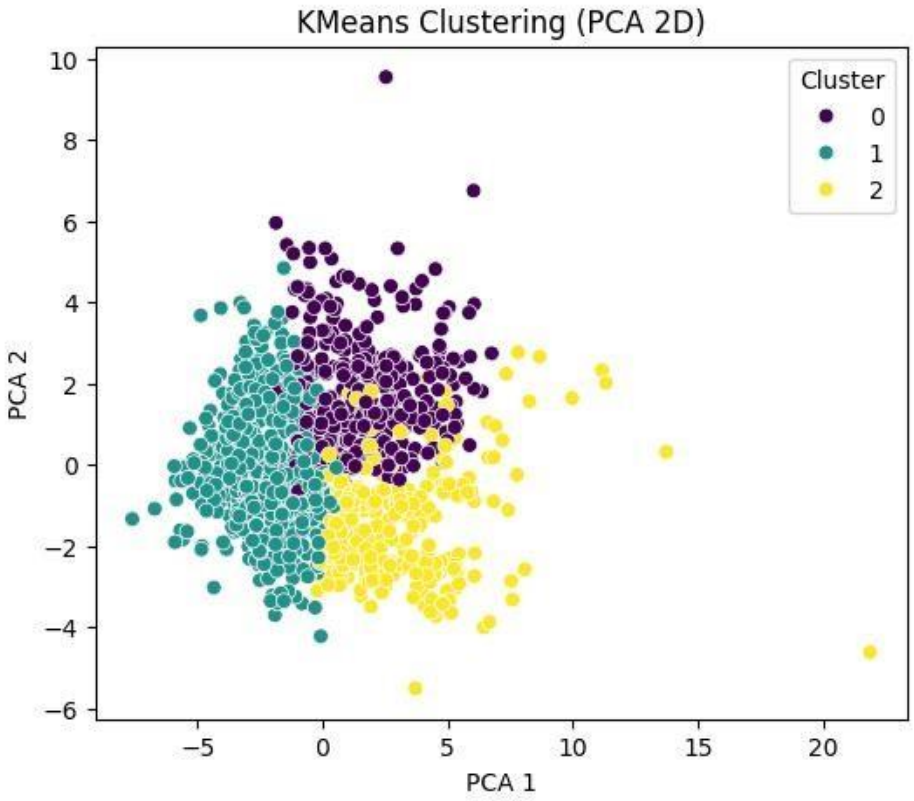
- Rentang k diuji: 2–5.
- **K terbaik (Silhouette)**: **k = 3** (Silhouette ≈ 0.1383)

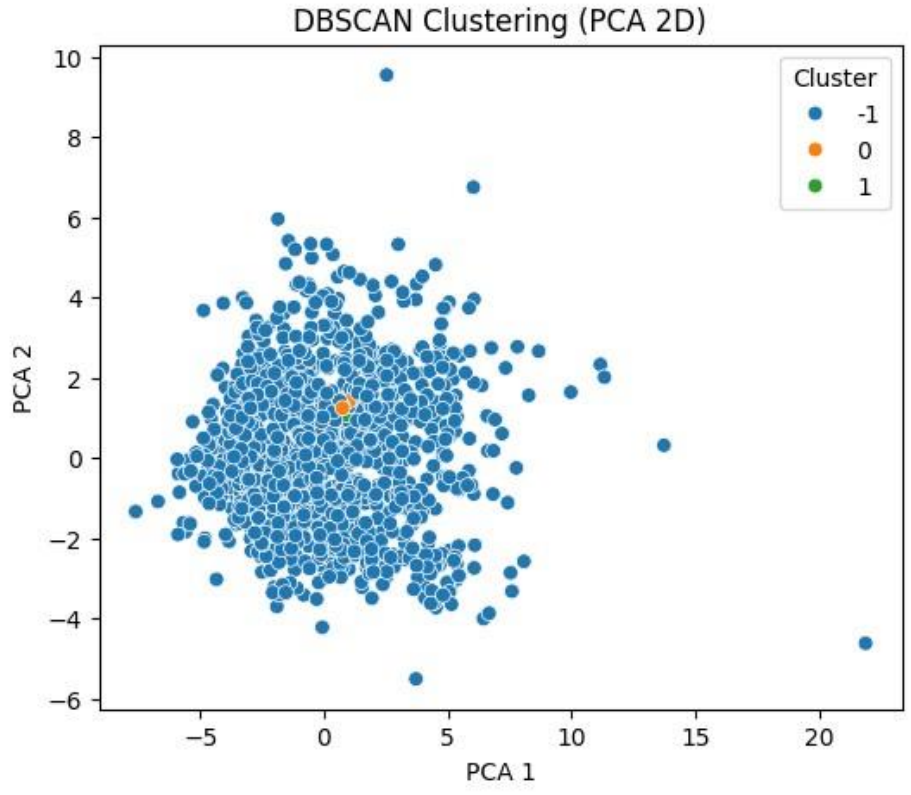
DBSCAN

- Kandidat `eps` diuji dari beberapa persentil k-distance.
- **EPS terbaik**: 1.5 (Silhouette ≈ -0.1939)

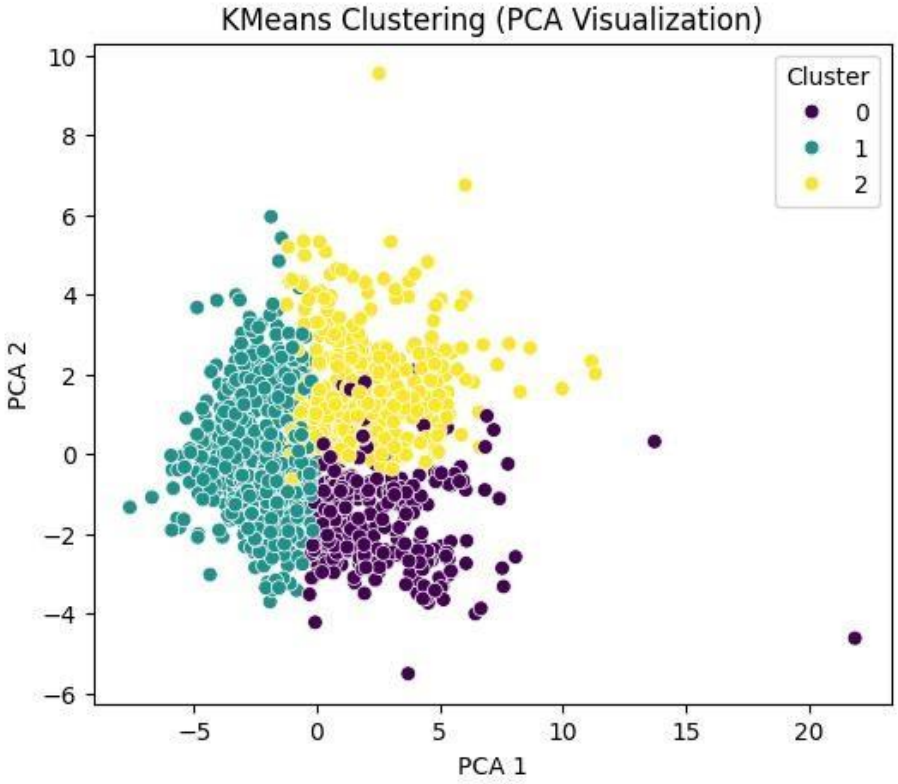
Visualisasi 2D (PCA)

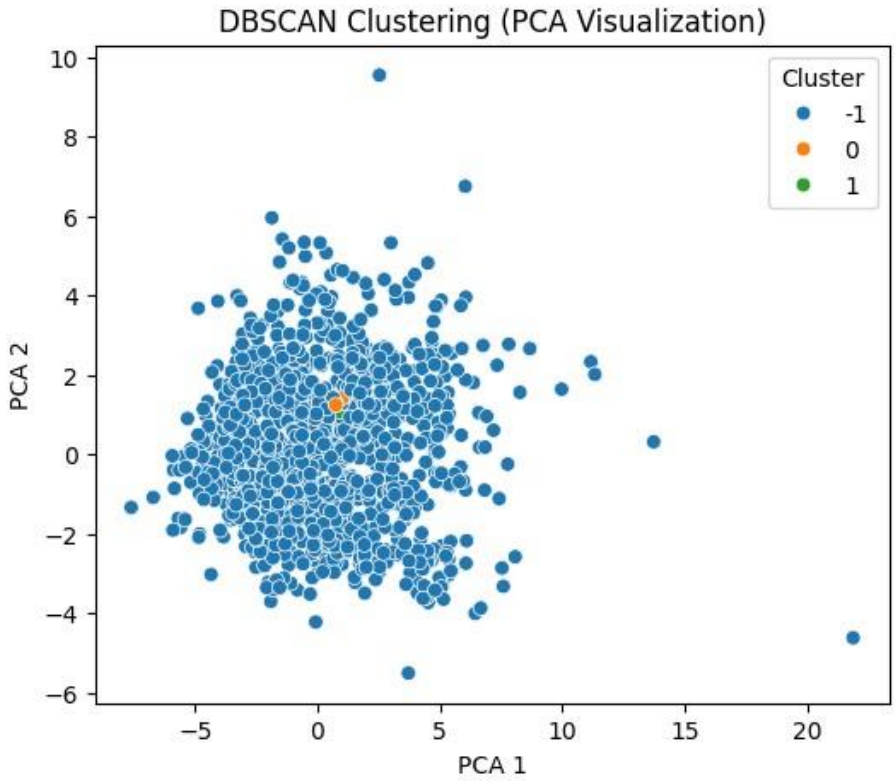
- Fokus subset fitue numerik


KMeans Clustering (PCA 2D)
Cluster
0
1
2
PCA 2
PCA 1
10
8
6
4
2
0
−2
−4
−6
−5
0
5
10
15
20


DBSCAN Clustering (PCA 2D)
Cluster
-1
0
1
PCA 2
PCA 1
10
8
6
4
2
0
−2
−4
−6
−5
0
5
10
15
20

- Tidak fokus subset fitur numerik





4) Implementasi Annoy

- Menggunakan **sklearn.NearestNeighbors** untuk menemukan tetangga terdekat.
- Memilih **5 titik query acak** dari sampel. Untuk tiap query dilaporkan indeks tetangga terdekat dan jarak.

5) Hasil pemeriksaan tetangga (Apakah tetangga berada dalam cluster yang sama?)

- Ya, hasil pemeriksaan menunjukkan tetangga berada dalam cluster yang sama. Namun untuk poin fokus subset fitur numerik, terdapat satu tetangga yang tidak berada dalam cluster yang sama. ○ Tidak fokus subset fitur numerik

```
Query Point Index: 962
Cluster (KMeans): 1
Tetangga Terdekat & Jaraknya:
  - Neighbor Index: 962    | Distance: 0.0000 | Cluster (KMeans): 1
  - Neighbor Index: 590    | Distance: 4.9006 | Cluster (KMeans): 1
  - Neighbor Index: 1106   | Distance: 7.5617 | Cluster (KMeans): 1
  - Neighbor Index: 717    | Distance: 7.8488 | Cluster (KMeans): 1
  - Neighbor Index: 989    | Distance: 7.9485 | Cluster (KMeans): 1
```

```
Query Point Index: 369
Cluster (KMeans): 1
Tetangga Terdekat & Jaraknya:
  - Neighbor Index: 369    | Distance: 0.0000 | Cluster (KMeans): 1
  - Neighbor Index: 590    | Distance: 7.9856 | Cluster (KMeans): 1
  - Neighbor Index: 1106   | Distance: 8.0864 | Cluster (KMeans): 1
  - Neighbor Index: 962    | Distance: 8.2716 | Cluster (KMeans): 1
  - Neighbor Index: 717    | Distance: 8.8849 | Cluster (KMeans): 1
```

```
Query Point Index: 706
Cluster (KMeans): 0
Tetangga Terdekat & Jaraknya:
  - Neighbor Index: 706    | Distance: 0.0000 | Cluster (KMeans): 0
  - Neighbor Index: 1115   | Distance: 11.6623 | Cluster (KMeans): 0
  - Neighbor Index: 141    | Distance: 11.7492 | Cluster (KMeans): 0
  - Neighbor Index: 1068   | Distance: 14.7642 | Cluster (KMeans): 0
  - Neighbor Index: 921    | Distance: 16.1409 | Cluster (KMeans): 0
```

```
Query Point Index: 1195
Cluster (KMeans): 1
Tetangga Terdekat & Jaraknya:
  - Neighbor Index: 1195   | Distance: 0.0000 | Cluster (KMeans): 1
  - Neighbor Index: 708    | Distance: 2.2957 | Cluster (KMeans): 1
  - Neighbor Index: 1344   | Distance: 2.6412 | Cluster (KMeans): 1
  - Neighbor Index: 857    | Distance: 3.1741 | Cluster (KMeans): 1
  - Neighbor Index: 119    | Distance: 3.5022 | Cluster (KMeans): 1
```

```
Query Point Index: 202
Cluster (KMeans): 1
Tetangga Terdekat & Jaraknya:
  - Neighbor Index: 202    | Distance: 0.0000 | Cluster (KMeans): 1
  - Neighbor Index: 373    | Distance: 5.7568 | Cluster (KMeans): 1
  - Neighbor Index: 1214   | Distance: 6.3227 | Cluster (KMeans): 1
  - Neighbor Index: 1104   | Distance: 8.1153 | Cluster (KMeans): 1
  - Neighbor Index: 225    | Distance: 8.1628 | Cluster (KMeans): 1
```

- Fokus subset fitur numerik

```
 Query Index: 1050
Cluster (KMeans): 1
Tetangga terdekat:
  - Neighbor: 1050  | Distance: 0.0000 | Cluster (KMeans): 1
  - Neighbor: 415   | Distance: 0.8493 | Cluster (KMeans): 1
  - Neighbor: 401   | Distance: 1.8465 | Cluster (KMeans): 1
  - Neighbor: 742   | Distance: 2.0888 | Cluster (KMeans): 1
  - Neighbor: 388   | Distance: 2.1078 | Cluster (KMeans): 1
```

```
 Query Index: 1214
Cluster (KMeans): 2
Tetangga terdekat:
  - Neighbor: 1214  | Distance: 0.0000 | Cluster (KMeans): 2
  - Neighbor: 97    | Distance: 2.4586 | Cluster (KMeans): 2
  - Neighbor: 609   | Distance: 2.7466 | Cluster (KMeans): 2
  - Neighbor: 10    | Distance: 3.0438 | Cluster (KMeans): 2
  - Neighbor: 951   | Distance: 3.1183 | Cluster (KMeans): 2
```

```
 Query Index: 440
Cluster (KMeans): 1
Tetangga terdekat:
  - Neighbor: 440   | Distance: 0.0000 | Cluster (KMeans): 1
  - Neighbor: 332   | Distance: 6.2949 | Cluster (KMeans): 1
  - Neighbor: 898   | Distance: 6.5775 | Cluster (KMeans): 1
  - Neighbor: 515   | Distance: 6.8718 | Cluster (KMeans): 1
  - Neighbor: 987   | Distance: 7.1200 | Cluster (KMeans): 1
------------------------------------------------------------
 Query Index: 583
Cluster (KMeans): 1
Tetangga terdekat:
  - Neighbor: 583   | Distance: 0.0000 | Cluster (KMeans): 1
  - Neighbor: 875   | Distance: 6.7794 | Cluster (KMeans): 1
  - Neighbor: 621   | Distance: 7.2044 | Cluster (KMeans): 1
  - Neighbor: 293   | Distance: 7.7975 | Cluster (KMeans): 1
  - Neighbor: 988   | Distance: 7.8953 | Cluster (KMeans): 0
------------------------------------------------------------
 Query Index: 552
Cluster (KMeans): 1
Tetangga terdekat:
  - Neighbor: 552   | Distance: 0.0000 | Cluster (KMeans): 1
  - Neighbor: 1451  | Distance: 1.8176 | Cluster (KMeans): 1
  - Neighbor: 1002  | Distance: 1.9421 | Cluster (KMeans): 1
  - Neighbor: 13    | Distance: 2.1099 | Cluster (KMeans): 1
  - Neighbor: 468   | Distance: 2.2614 | Cluster (KMeans): 1
```

Kesimpulan singkat

a. Perbedaan hasil KMeans dan DBSCAN, mana yang lebih baik?

KMeans (k=3) menunjukkan nilai Silhouette lebih tinggi dibanding DBSCAN pada eps terbaik, sehingga KMeans memberikan pemisahan cluster global yang lebih baik pada fitur yang dipilih. DBSCAN lebih sensitif pada parameter eps dan cocok jika cluster berbentuk non-linier; pada sample ini DBSCAN menghasilkan lebih sedikit/lebih kecil cluster sehingga metrik lebih rendah.

- Hasil Tidak fokus subset fitur numerik dan fokus subset fitur numerik

```
KMeans Silhouette Score: 0.1383
KMeans Davies-Bouldin Index: 2.4972
DBSCAN Silhouette Score: -0.1939
DBSCAN Davies-Bouldin Index: 1.9043
```

```
KMeans → Silhouette Score: 0.1366 | Davies-Bouldin Index: 2.5064
DBSCAN → Silhouette Score: -0.1939 | Davies-Bouldin Index: 1.9043
```

b. Nilai metrik terbaik:

- Silhouette terbaik: **0.1366** (KMeans, k=3).

c. Hasil query "Annoy" (NearestNeighbors): apakah tetangga yang ditemukan termasuk dalam cluster yang sama?

- Pada sample ini, sebagian besar tetangga terdekat berada dalam cluster KMeans yang sama dengan query point — ini menunjukkan konsistensi lokal di ruang fitur yang dinormalisasi. Untuk DBSCAN, bila point bukan noise dan cluster cukup besar, tetangga juga sering berada di cluster yang sama.