Analisis Komparatif *Clustering* K-Means dan DBSCAN untuk Segmentasi Properti dan Prediksi Harga Rumah dengan *Approximate Nearest Neighbor*

Rangga Dwi Saputra
2341720248

Abstrak: Penelitian ini membandingkan K-Means dan DBSCAN untuk segmentasi properti harga rumah, memanfaatkan 17 fitur yang di-*preprocess* menjadi 45 dimensi. K-Means (K=5) berhasil mengidentifikasi segmen pasar yang jelas (rata-rata harga dari 110,651 hingga 338,646), meskipun metrik (Silhouette Score 0.1419) menunjukkan tumpang tindih klaster. DBSCAN (epsilon=3.0, MinSamples=50) efektif mengisolasi 6.4% data sebagai *noise* (terutama rumah-rumah berharga/berkualitas tinggi) tetapi gagal dalam segmentasi mayoritas data. Implementasi Annoy menunjukkan kemiripan tinggi antar data, mendukung potensi penggunaan ANN untuk prediksi harga jual (*SalePrice*) yang efisien pada data *test*.

Keyword : Clustering, K-Means, DBSCAN, Approximate Nearest Neighbor (ANN), Annoy, Harga Rumah (*House Price*), Segmentasi Properti

1. Introduction

Analisis *clustering* adalah teknik penting dalam *unsupervised machine learning* yang bertujuan mengelompokkan data ke dalam grup-grup (klaster) berdasarkan kemiripan fitur. Dalam konteks data House Prices, *clustering* dapat mengidentifikasi segmen-segmen properti yang berbeda (misalnya, klaster rumah murah, klaster rumah mewah, klaster rumah tua, dll.) berdasarkan karakteristik fisiknya. Segmentasi ini sangat berguna untuk pemodelan prediktif harga jual (seperti regresi dengan model per klaster) atau untuk strategi pasar.

Penelitian ini membandingkan dua metode *clustering* utama, yaitu K-Means (berbasis *centroid* yang cocok untuk klaster berbentuk bulat) dan DBSCAN (berbasis kepadatan yang cocok untuk menemukan klaster berbentuk non-linear dan mengidentifikasi *noise*). Selain itu, diuji pula implementasi Approximate Nearest Neighbor (ANN) dengan Annoy sebagai teknik untuk melakukan pencarian tetangga terdekat yang efisien pada data berdimensi tinggi setelah proses *clustering*.

2. Metodologi (Metjode)

2.1. Persiapan Data (Data Preparation)

Data *training* (train.csv) dan *test* (test.csv) digabungkan untuk memastikan konsistensi dalam *preprocessing* dan *encoding* fitur, menghasilkan total 2919 sampel.

2.1.1 Fitur yang Digunakan:

Dipilih 16 fitur utama, ditambah 1 fitur hasil *feature engineering*, menjadi 17 fitur total untuk *clustering*:

a) Numerik: GrLivArea, TotalBsmtSF, OverallQual, YearBuilt, 1stFlrSF, 2ndFlrSF, GarageCars, GarageArea, BsmtFullBath, FullBath, Fireplaces, LotArea, LotFrontage, MasVnrArea, dan fitur baru TotalSF = TotalBsmtSF + 1stFlrSF + 2ndFlrSF.

b) Kategorikal: Neighborhood, MSZoning.

2.1.2 Preprocessing Pipeline:

a) Imputasi: *Missing values* pada kolom numerik diisi dengan median, sedangkan pada kolom kategorikal diisi dengan modus atau label 'None' (untuk kolom seperti Alley, BsmtQual, dll. yang NaN nya berarti tidak ada fitur tersebut).

b) Scaling: Fitur numerik di-*scale* menggunakan StandardScaler.

c) Encoding: Fitur kategorikal di-*encode* menggunakan OneHotEncoder (handle_unknown='ignore').

2.1.3 Hasil Akhir Preprocessing:

Data memiliki 45 fitur setelah *encoding*.

2.2. Metode Clustering

2.2.1. K-Means

a) Penentuan K Optimal: Menggunakan metode Elbow (Inertia) dan metrik kualitas klaster (Silhouette Score dan Davies-Bouldin Index/DBI) dengan rentang K=2 hingga K=10.

b) Metrik Kualitas Klaster:

- Silhouette Score: Mengukur seberapa mirip suatu objek dengan klaster sendiri dibandingkan dengan klaster lain. Nilai tinggi (mendekati +1) menunjukkan klaster yang padat dan terpisah.
- Davies-Bouldin Index (DBI): Mengukur rasio antara dispersi intra-klaster (di dalam klaster) dengan jarak antar-klaster. Nilai rendah menunjukkan *clustering* yang lebih baik.

c) K Final: Berdasarkan analisis visual, K_optimal = 5 dipilih karena dianggap memberikan keseimbangan terbaik pada grafik (meskipun skor Silhouette dan DBI tidak terlalu tinggi).

2.2.2. DBSCAN

a) Penentuan Parameter Optimal: Menggunakan K-Distance Graph untuk mencari nilai epsilon (radius tetangga) optimal, dengan asumsi MinSamples = 50.

b) epsilon Final: Berdasarkan pengamatan titik 'siku' pada grafik K-Distance (meskipun tidak ditampilkan secara langsung), dipilih epsilon = 3.0.

c) Min Samples Final: MinSamples final = 50.

d) Metrik Kualitas Klaster**:** Silhouette Score dan DBI dihitung hanya pada titik **non-noise** (klaster neq -1).

2.3. Approximate Nearest Neighbor (ANN)

- Metode: Annoy (Approximate Nearest Neighbors Oh Yeah), menggunakan metrik jarak fangular.
- Hyperparameter: Jumlah pohon (*num_trees*) ditetapkan 10.
- Tujuan: Untuk mencari rumah-rumah yang paling mirip (tetangga terdekat) dengan rumah tertentu (*query point*) secara cepat.

3. Hasil dan Pembahasan (Result + Discuss)

3.1. Hasil K-Means Clustering K=5

Metrik	**Nilai**	**Interpretasi**
Silhouette Score	0.1419	Nilai relatif rendah menunjukkan klaster tidak terpisah dengan jelas dan/atau ada tumpang tindih.
Davies-Bouldin Index	2.0010	Nilai yang cukup tinggi (ideal < 1.0) mengindikasikan klaster yang kurang padat dan jarak antar klaster yang kecil.

Kesimpulan Kualitas K-Means: Hasil metrik menunjukkan bahwa klaster K-Means K=5 kurang optimal dalam memisahkan segmen rumah secara jelas, kemungkinan karena kompleksitas data dan sensitivitas K-Means terhadap penskalaan dan bentuk klaster yang tidak bulat.

Profil Klaster K-Means (Berdasarkan Data Training)

Klaster	% Data (Train)	Rata-rata SalePrice	Rata-rata OverallQual	Rata-rata GrLivArea	Ciri Khas
Cluster 0	10.1%	338,646	8.2/10	2,397 sq ft	Klaster Mewah: Kualitas, luas, dan harga tertinggi. Dibangun baru (median YearBuilt ≈ 2003). Lokasi utama: NridgHt, NoRidge.
Cluster 1	19.5%	110,651	5.0/10	1,202 sq ft	Klaster Termurah/Tua: Kualitas, luas, dan harga terendah. Dibangun sangat tua (median YearBuilt ≈ 1931). Lokasi: OldTown.
Cluster 2	20.8%	207,652	6.8/10	1,522 sq ft	Klaster Menengah-Atas & Baru: Kualitas dan harga di atas rata-rata. Dibangun relatif baru (median YearBuilt ≈ 2000). Lokasi: CollgCr.
Cluster 3	25.6%	137,233	5.2/10	1,101 sq ft	Klaster Menengah-Bawah: Harga dan kualitas sedikit di

Klaster	% Data (Train)	Rata-rata SalePrice	Rata-rata OverallQual	Rata-rata GrLivArea	Ciri Khas
					bawah rata-rata. Luas terkecil. Dibangun cukup tua (median YearBuilt ≈ 1964). Lokasi: NAmes.
Cluster 4	24.0%	194,989	6.6/10	1,835 sq ft	Klaster Luas/Baru: Luas bangunan tinggi. Harga dan kualitas di atas rata-rata. Dibangun relatif baru median YearBuilt ≈ 1995. Lokasi: Gilbert.

3.2. Hasil DBSCAN Clustering (epsilon=3.0, MinSamples=50)

Metrik	Nilai	Interpretasi
Jumlah Klaster	1	Hanya satu klaster padat yang teridentifikasi, selain *noise*.
Jumlah Noise (-1)	178 rumah (6.4%)	Sejumlah kecil *outlier* berhasil diisolasi.
Silhouette Score	nan	Tidak dapat dihitung karena hanya ada 1 klaster (>1 klaster non-noise diperlukan).

Metrik	Nilai	Interpretasi
Davies-Bouldin Index	nan	Tidak dapat dihitung karena hanya ada 1 klaster.

Kesimpulan Kualitas DBSCAN: DBSCAN dengan epsilon=3.0 dan MinSamples=50 menghasilkan satu klaster besar (93.6% data *training*) dan mengidentifikasi sisa data sebagai *noise*. Ini menunjukkan bahwa data cenderung berkelompok di satu area kepadatan tinggi, atau epsilon yang dipilih terlalu besar/kecil. Untuk data yang sudah di-*scale* dan di-*encode* seperti ini, biasanya klaster akan berbentuk satu bola besar, dan DBSCAN sulit memisahkan sub-klaster tanpa *tuning* epsilon yang sangat presisi atau menggunakan metrik jarak yang berbeda.

Profil Klaster DBSCAN

a) Cluster 0 (Mayoritas - 93.6%): Merepresentasikan populasi umum data harga rumah: Rata-rata Harga 175,010, OverallQual≈6.1/10, GrLivArea≈ 1,463 sq ft.

b) Cluster Noise (-1 - 6.4%): Rata-rata Harga ≈ 266,828, Median OverallQual = 7.0/10, Rata-rata GrLivArea ≈ 2,276 sq ft. Ciri-ciri ini menunjukkan bahwa *noise* DBSCAN sebagian besar adalah rumah-rumah dengan kualitas dan/atau ukuran yang jauh di atas rata-rata populasi umum, menjadikannya *outlier* dalam hal kepadatan data.

3.3. Hasil Approximate Nearest Neighbor (Annoy)

Annoy berhasil menemukan 5 tetangga terdekat yang sangat cepat untuk 4 rumah *query* yang dipilih.

Query Index	Kualitas (OQ)	Luas (GrLivArea)	Harga Jual	Tetangga Terdekat	Jarak (Angular)	Kualitas (OQ)	Luas (SF)	Harga Jual
0 (Train)	7	1,710	208,500	1240 (Train)	0.1173	7	1,768	224,900
				1366 (Train)	0.1427	7	1,790	193,000
				2167 (Test)	0.3585	6	1,771	N/A

Query Index	Kualitas (OQ)	Luas (GrLivArea)	Harga Jual	Tetangga Terdekat	Jarak (Angular)	Kualitas (OQ)	Luas (SF)	Harga Jual
10 (Train)	5	1,040	129,500	2139 (Test)	0.1548	5	992	N/A
				2476 (Test)	0.1690	5	999	N/A
1459 (Train)	5	1,256	147,500	1571 (Test)	0.2265	5	1,172	N/A
				985 (Train)	0.2435	5	1,164	125,000

Kesimpulan ANN: Pencarian tetangga terdekat menunjukkan kemiripan yang tinggi (jarak angular yang kecil) antara rumah di data *training* (yang memiliki SalePrice) dengan rumah di data *test* (yang SalePrice-nya N/A), bahkan untuk rumah dengan kualitas dan harga yang berbeda (seperti Index 0). Ini menunjukkan bahwa Annoy berhasil mengidentifikasi properti serupa di seluruh *dataset* gabungan. Teknik ini dapat digunakan sebagai langkah awal dalam metode *imputasi* atau *regresi K-Nearest Neighbors* untuk memprediksi harga jual rumah *test* berdasarkan harga rumah *training* yang paling mirip.

4. Kesimpulan (Conclusion)

4.1. Hasil Riset Terbaik

Riset *clustering* pada data harga rumah ini menunjukkan bahwa:

4.1.1 K-Means (K=5) adalah pendekatan *clustering* yang lebih informatif untuk segmentasi pasar karena berhasil mengidentifikasi 5 klaster dengan profil karakteristik dan harga jual yang berbeda secara intuitif (Klaster Mewah, Menengah-Atas, Menengah-Bawah, dan Termurah/Tua). Namun, klaster-klaster ini tidak terpisah dengan baik (Silhouette Score 0.1419 dan DBI 2.0010), menunjukkan adanya tumpang tindih antar segmen.

4.1.2 DBSCAN dengan parameter yang dipilih (epsilon=3.0, MinSamples=50) kurang efektif untuk segmentasi pasar utama, karena hanya menghasilkan 1 klaster mayoritas. Namun, DBSCAN sangat berguna untuk identifikasi *outlier*, di mana klaster *noise* (-1) sebagian besar adalah rumah-rumah besar dan berkualitas tinggi yang jauh dari populasi umum, yang mungkin memerlukan penanganan khusus dalam model prediksi.

4.2. Pemanfaatan Annoy

Teknik ANN dengan Annoy berhasil diterapkan dan menunjukkan potensi besar untuk memperkirakan harga rumah di data *test*. Dengan menemukan tetangga terdekat dari data *training* (yang memiliki harga jual), kita dapat menggunakan harga rata-rata atau tertimbang dari tetangga tersebut sebagai prediksi harga jual untuk rumah-rumah di data *test*.