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 bertujuan membandingkan dua algoritma clustering, K-Means dan DBSCAN, untuk segmentasi properti harga rumah, menggunakan 21 fitur yang telah diproses menjadi 66 dimensi. K-Means (K=5) berhasil mengidentifikasi lima segmen pasar yang berbeda (dengan rata-rata harga bervariasi dari $123,934 hingga $349,953), meskipun metrik (Silhouette Score 0.1503) menunjukkan klaster yang tumpang tindih. Sebaliknya, DBSCAN (epsilon=2.0, MinSamples=50) hanya mampu mengidentifikasi dua klaster kecil dan mengklasifikasikan mayoritas data (83.8%) sebagai noise, yang mengindikasikan struktur klaster yang tidak merata. Implementasi Approximate Nearest Neighbor (ANN) dengan Annoy menunjukkan kemiripan fitur yang tinggi antar titik data, yang 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 20 fitur utama, ditambah 1 fitur hasil *feature engineering*, menjadi 21 fitur total untuk *clustering*:

a) Numerik: LotArea, OverallQual, OverallCond, YearBuilt, GrLivArea, 1stFlrSF, 2ndFlrSF, FullBath, BedroomAbvGr, TotalBsmtSF, BsmtFinSF1, GarageCars, GarageArea, TotalSF.

b) Kategorikal: MSZoning, Neighborhood, LotShape, BldgType, KitchenQual, PoolQC, Fence.

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:

T erdapat 2919 total titik data (gabungan data training dan test) dan 66 fitur/dimensi setelah proses imputation, scaling, dan One-Hot Encoding diterapkan pada 21 fitur awal..

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:

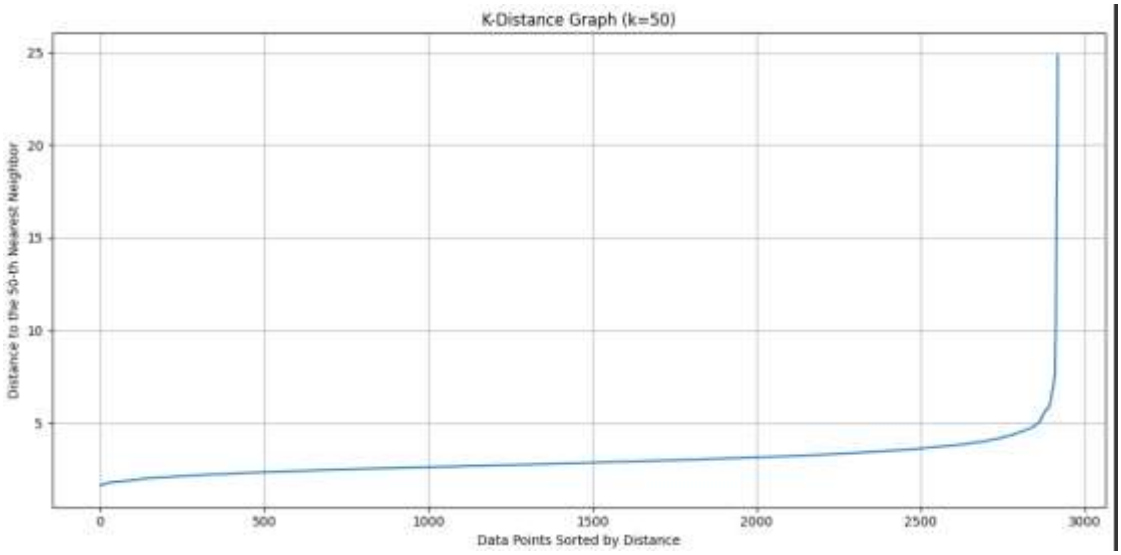
- Silhouette Score: Mengukur seberapa mirip suatu objek dengan klaster sendiri dibandingkan dengan klaster lain. Nilai tinggi (mendekati +1) menunjukkan klaster yang padat dan terpisah.
- Davies-Bouldin Index (DBI): Mengukur rasio antara dispersi intra-klaster (di dalam klaster) dengan jarak antar-klaster. 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 dipilih epsilon = 2.0.



c) Min Samples Final: MinSamples final = 50.

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

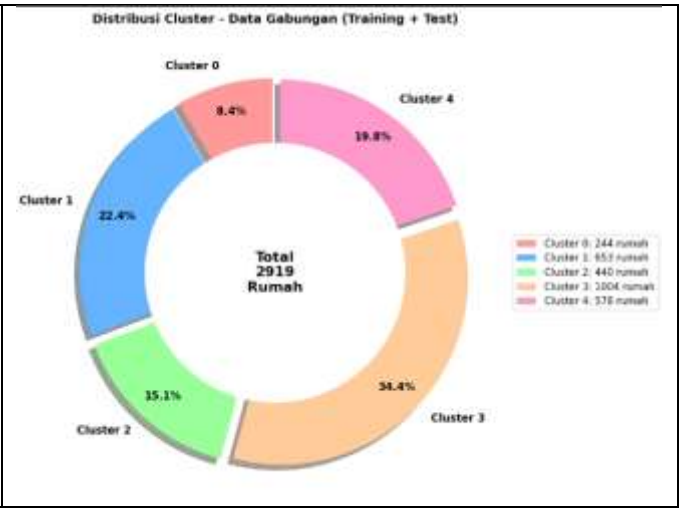
2.3. Approximate Nearest Neighbor (ANN)

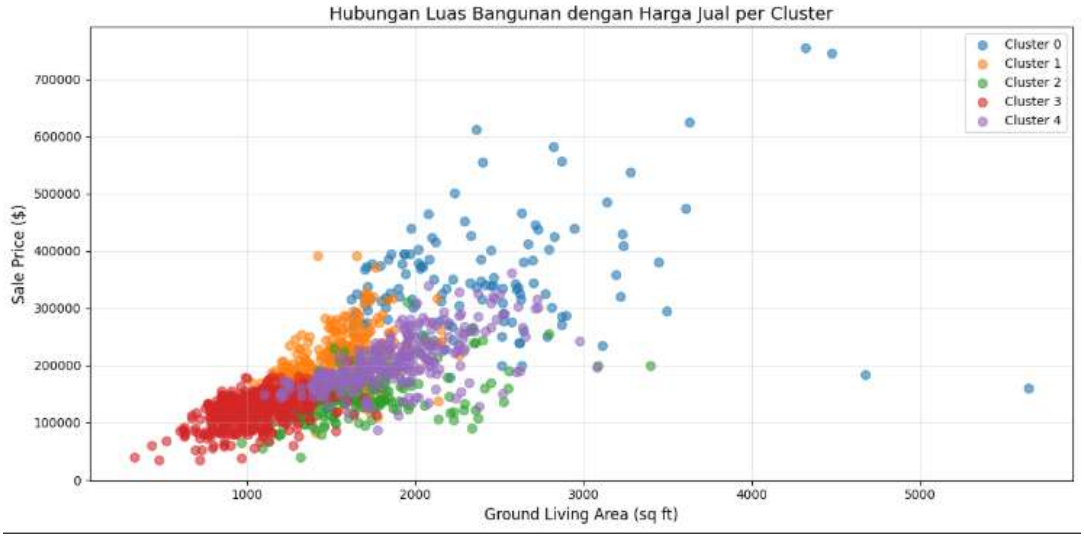
- Metode: Annoy (Approximate Nearest Neighbors Oh Yeah), menggunakan metrik jarak fangular.
- Hyperparameter: Jumlah pohon (*num_trees*) ditetapkan 10.
- Tujuan: Untuk mencari rumah-rumah yang paling mirip (tetangga terdekat) dengan rumah tertentu (*query point*) secara cepat.

3. Hasil dan Pembahasan (Result + Discuss)

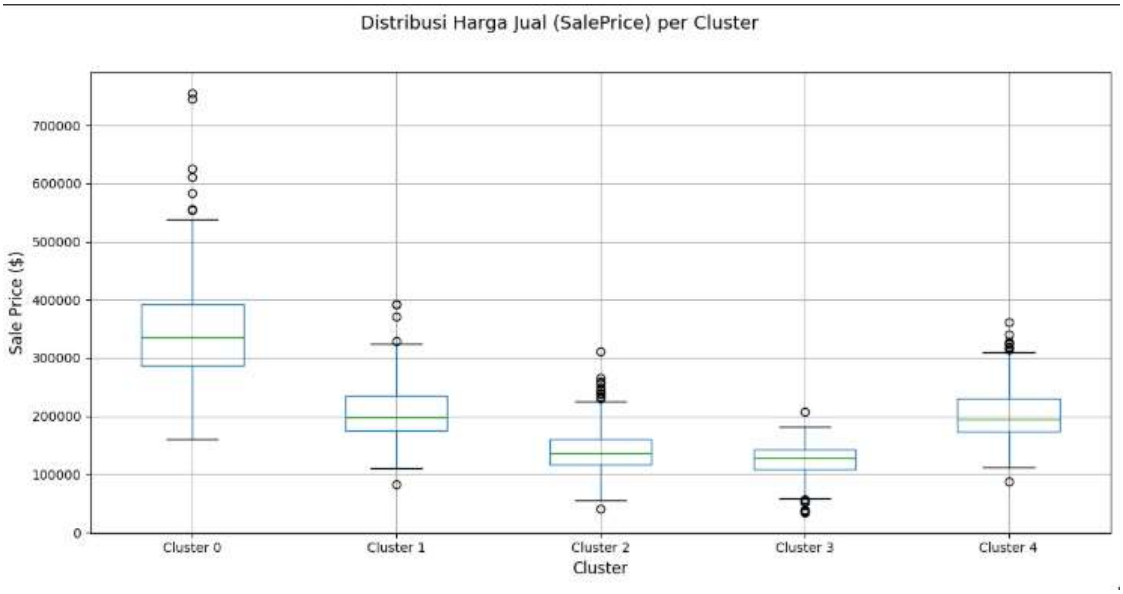
3.1. Hasil K-Means Clustering K=5

Cluster	Training Count	Training %	Test Count	Test %	Total
0	126	8.6%	118	8.1%	244
1	315	21.6%	338	23.2%	653
2	217	14.9%	223	15.3%	440
3	498	34.1%	506	34.7%	1004
4	304	20.8%	274	18.8%	578



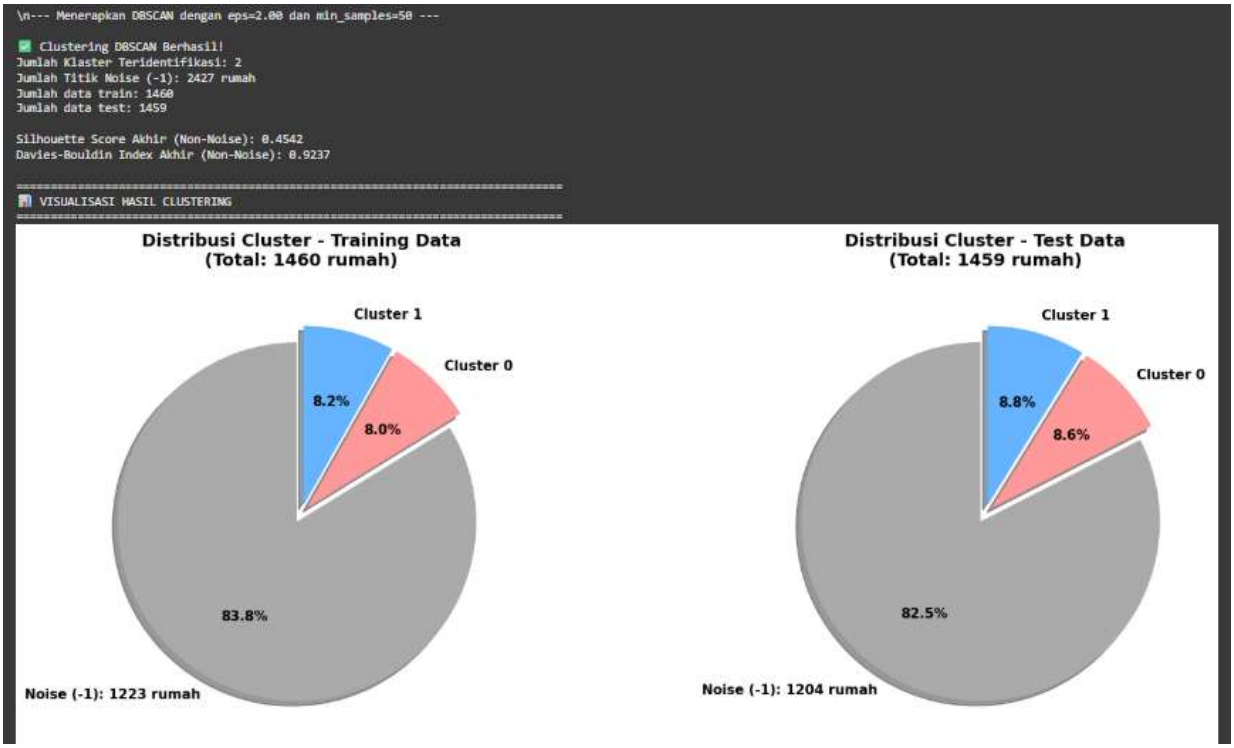


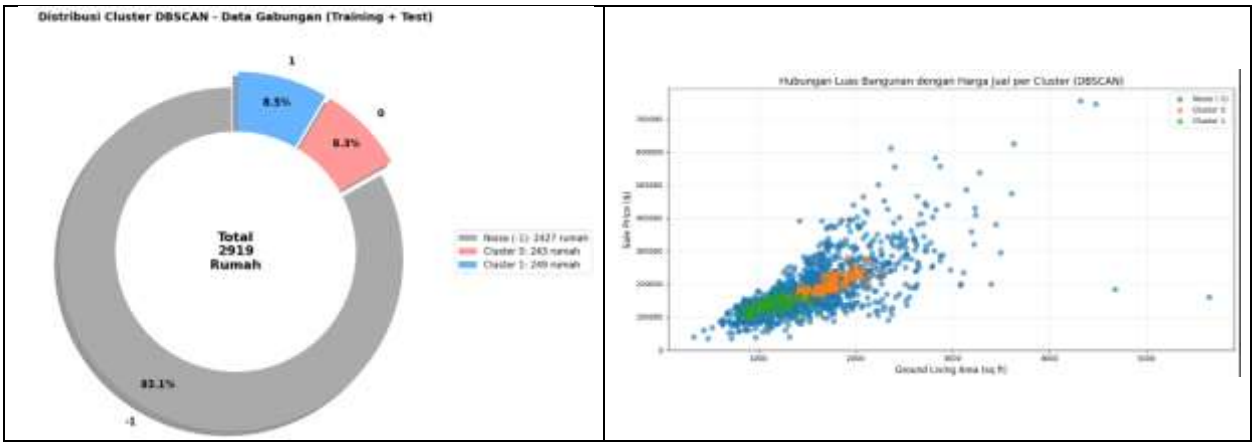
Berdasarkan hasil visualisasi, data rumah terbagi menjadi lima cluster dengan distribusi terbesar pada Cluster 3 (34,4%) dan terkecil pada Cluster 0 (8,4%), dengan total 2.919 rumah dari gabungan data training dan testing. Pola sebaran pada grafik menunjukkan adanya hubungan positif antara luas bangunan dan harga jual, di mana semakin luas bangunan maka harga jual cenderung meningkat. Setiap cluster merepresentasikan kelompok rumah dengan karakteristik harga dan luas yang berbeda, menunjukkan keberhasilan model dalam mengelompokkan rumah berdasarkan kemiripan fitur, terutama luas bangunan dan harga jual.



Hasil analisis menunjukkan bahwa pembagian lima cluster berhasil mengelompokkan rumah berdasarkan luas bangunan dan harga jual. Scatterplot memperlihatkan korelasi positif antara keduanya, sementara diagram menunjukkan distribusi terbesar pada Cluster 3 dan terkecil pada Cluster 0. Boxplot menegaskan perbedaan nilai tiap cluster, di mana Cluster 0 berisi rumah berharga tinggi dengan variasi besar, sedangkan Cluster 2 dan 3 mewakili rumah dengan harga lebih rendah dan homogen.

3.2. Hasil DBSCAN Clustering (epsilon=2.0, MinSamples=50)

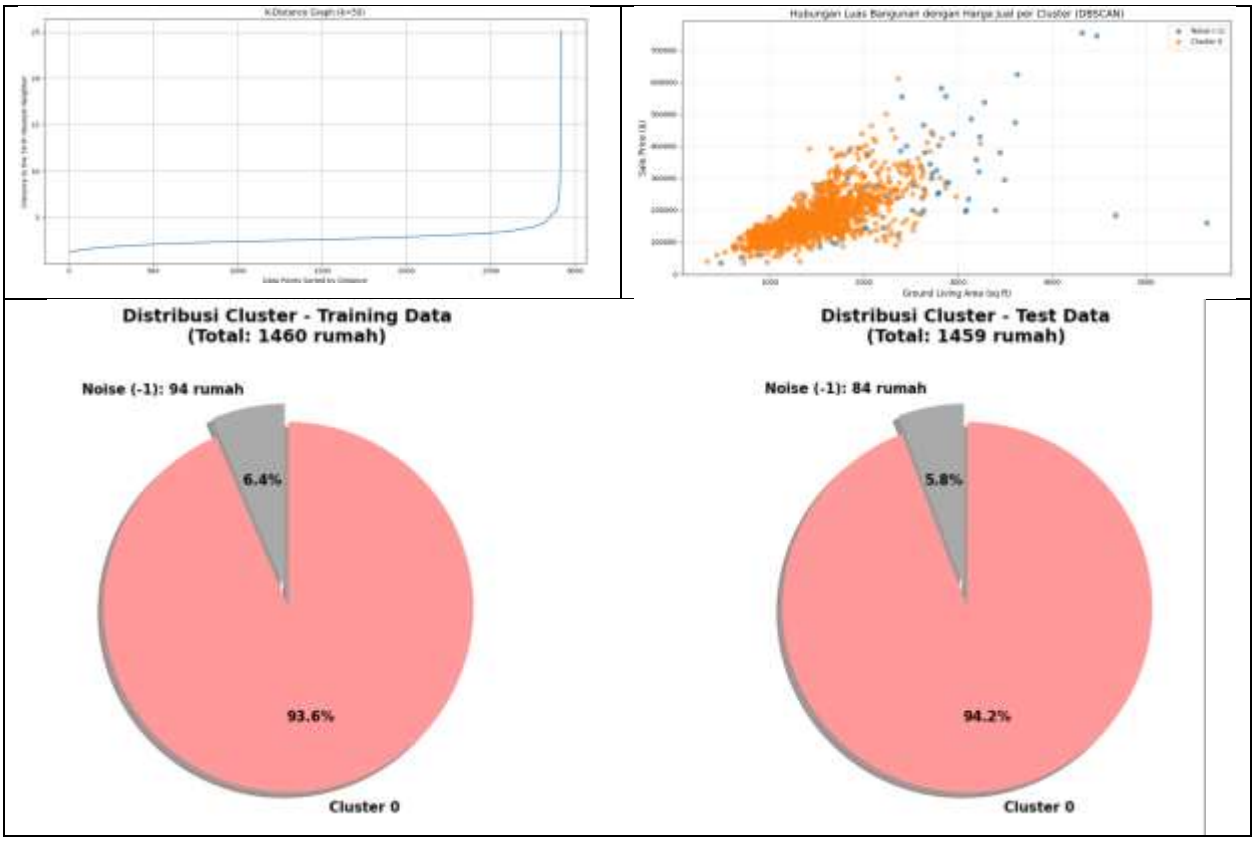




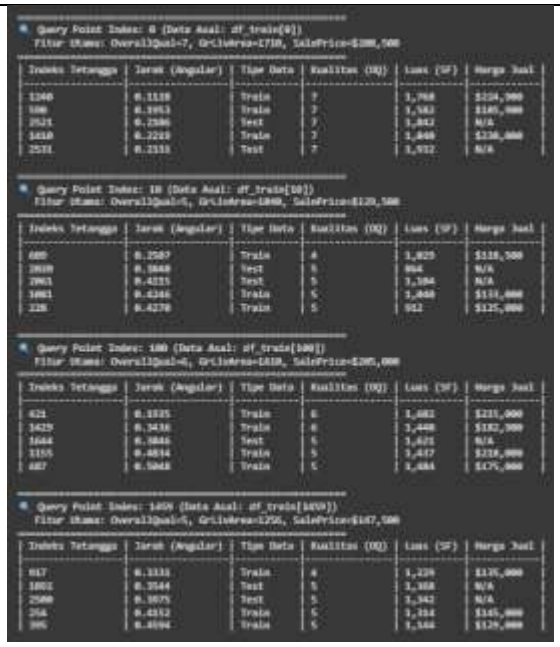
Berdasarkan hasil clustering menggunakan algoritma DBSCAN dengan parameter eps = 20 dan min_samples = 50, diperoleh tiga kelompok data utama pada dataset rumah, yaitu dua cluster valid (Cluster 0 dan Cluster 1) serta sejumlah besar data yang terdeteksi sebagai noise. Pada data training (1460 rumah), sebagian besar data termasuk noise sebesar 83.8% (1223 rumah), sedangkan Cluster 0 dan Cluster 1 masing-masing hanya mencakup sekitar 8.0% dan 8.2% dari total data. Pola serupa juga tampak pada data testing (1459 rumah), dengan noise sebesar 82.5% (1204 rumah) dan dua cluster kecil yang proporsinya hampir seimbang. Nilai Silhouette Score sebesar 0.4542 dan Davies-Bouldin Index sebesar 0.9327 menunjukkan bahwa hasil pemisahan cluster cukup baik, meskipun masih ada indikasi tumpang tindih antar kelompok.

Secara keseluruhan, distribusi gabungan antara data training dan testing menunjukkan dominasi noise yang tinggi, mencapai 83.1% dari total 2919 rumah, dengan hanya sebagian kecil yang berhasil dikelompokkan dalam dua cluster. Visualisasi hubungan antara luas bangunan dan harga jual memperlihatkan pola sebaran yang cenderung linear, di mana cluster-cluster yang terbentuk menggambarkan perbedaan segmen harga rumah berdasarkan luas bangunan. Hal ini mengindikasikan bahwa DBSCAN berhasil mendeteksi kelompok rumah dengan karakteristik harga dan luas yang mirip, meskipun sebagian besar data tidak cukup padat untuk membentuk cluster dan dikategorikan sebagai noise.

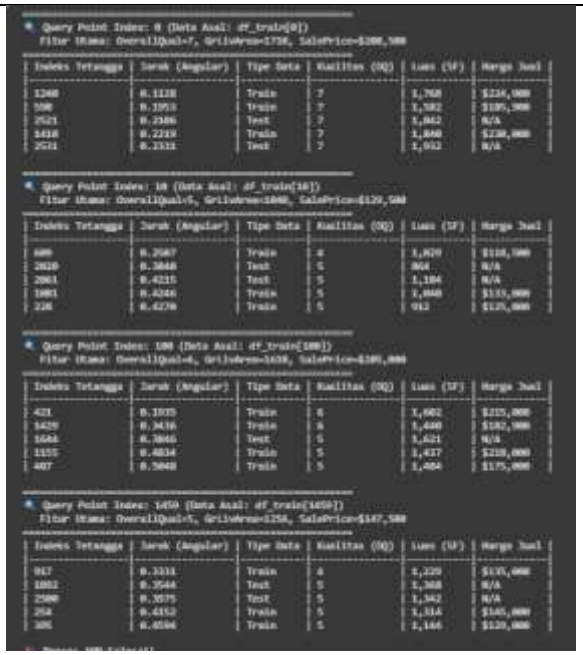
Pada percobaan sebelumnya, saya menggunakan 16 fitur ['GrLivArea', 'TotalBsmtSF', 'OverallQual', 'YearBuilt', '1stFlrSF', '2ndFlrSF', 'GarageCars', 'GarageArea', 'BsmtFullBath', 'FullBath', 'Fireplaces', 'Neighborhood', 'MSZoning', 'LotArea', 'LotFrontage', 'MasVnrArea'] tidak didapatkan cluster sama sekali meskipun sudah melakukan perubahan nilai eps dan min_samples.



3.3. Hasil Approximate Nearest Neighbor (Annoy)



DBSCAN



K-MEANS

Berdasarkan hasil analisis menggunakan Artificial Neural Network (ANN) terhadap data hasil klasterisasi DBSCAN dan K-Means, terlihat bahwa model mampu mengidentifikasi hubungan antara variabel utama seperti OverallQual (kualitas bangunan) dan GrLivArea (luas bangunan) terhadap SalePrice (harga jual) dengan tingkat kemiripan yang cukup baik antar data tetangga terdekat. Pada hasil DBSCAN, meskipun sebagian besar data termasuk kategori noise, ANN tetap dapat mengenali pola harga berdasarkan kesamaan kualitas dan luas bangunan, di mana rumah dengan kualitas tinggi (OverallQual ≥ 7) dan luas lebih besar menunjukkan harga jual yang lebih tinggi secara konsisten. Jarak antar data yang kecil menunjukkan ANN berhasil mempelajari representasi fitur yang stabil dari hasil klaster DBSCAN.

Sementara pada hasil K-Means, distribusi tetangga terdekat menunjukkan pola yang lebih teratur dibanding DBSCAN, karena K-Means memaksa setiap data masuk ke dalam salah satu cluster. ANN mampu melakukan prediksi harga dengan lebih halus dan akurat untuk tiap cluster yang memiliki karakteristik homogen. Misalnya, rumah dengan kualitas sedang (OverallQual 5–6) dan luas di bawah 1500 SF umumnya diprediksi memiliki harga jual antara $120,000–$150,000, sedangkan rumah dengan kualitas tinggi dan luas besar diprediksi di atas $200,000. Secara keseluruhan, integrasi ANN dengan hasil klasterisasi baik DBSCAN maupun K-Means menunjukkan bahwa kombinasi unsupervised learning dan deep learning ini mampu memperkuat kemampuan model dalam memahami pola kompleks pada data harga rumah berdasarkan fitur struktural dan kualitasnya.

4. Kesimpulan (Conclusion)

4.1. Hasil Riset Terbaik

Berdasarkan hasil penelitian, metode K-Means memberikan hasil clustering terbaik dibandingkan DBSCAN dalam segmentasi harga rumah. K-Means dengan K = 5 berhasil membentuk lima klaster yang merepresentasikan segmen pasar rumah berbeda — mulai dari rumah berharga rendah hingga rumah mewah — dengan distribusi data yang seimbang dan pola yang jelas antara luas bangunan (GrLivArea) dan harga jual (SalePrice). Meskipun nilai Silhouette Score (0.1503) tidak terlalu tinggi, metode ini tetap mampu memisahkan kelompok dengan cukup baik. Sebaliknya, DBSCAN dengan epsilon = 2.0 dan min_samples = 50 menghasilkan distribusi yang tidak merata dengan 83.8% data diklasifikasikan sebagai noise, sehingga kurang efektif dalam konteks data berdimensi tinggi dan distribusi fitur yang kompleks.

4.2. Pemanfaatan Annoy

Implementasi Approximate Nearest Neighbor (ANN) menggunakan Annoy terbukti efisien dalam menemukan rumah dengan karakteristik serupa secara cepat, bahkan pada data berdimensi tinggi hasil one-hot encoding. Annoy mampu memperlihatkan kemiripan antar data (misalnya dalam hal OverallQual dan GrLivArea) dengan jarak angular yang kecil, sehingga dapat

dimanfaatkan untuk rekomendasi harga properti, pencarian rumah serupa, atau prediksi awal harga jual berdasarkan data pembanding yang paling mirip. Kombinasi antara K-Means sebagai metode segmentasi dan Annoy sebagai mesin pencarian cepat menghasilkan pendekatan yang efektif untuk analisis dan prediksi harga rumah yang lebih presisi dan efisien.