Journal of Applied Informatics and Computing (JAIC)
Vol.9, No.3, June 2025, pp. 741~748
e-ISSN: 2548-6861

Improving House Price Clustering Results with K-means through the Implementation of One-hot Encoding Pre-processing Technique

Vicka Rizqi Maulani[1*], Mula Agung Barata[2*], Pelangi Eka Yuwita[3]**
*Teknik Informatika, Universitas Nahdlatul Ulama Sunan Giri
**Teknik Mesin, Universitas Nahdlatul Ulama Sunan Giri
vickarizqimaulani@gmail.com[1], mula.ab26@gmail.com[2], pelangi.ardata@gmail.com[3]

Article Info

Article history:

Received 2025-04-23
Revised 2025-04-27
Accepted 2025-05-06

Keyword:

Clustering,
K-means,
One-hot Encoding,
House Price.

ABSTRACT

Basic human needs include a house that serves as a place to live and a shelter from everything. In Indonesia, owning a house is still a challenging aspect due to its high price. Information on house prices is needed for prospective buyers or consumers, so that buyers can adjust their needs and finances, and for producers or sellers it is used as a way to determine the segmentation of targeted market groups. House prices are influenced by several factors including, building area, number of bedrooms, number of bathrooms, location, condition and the presence of a garage. This research aims to improve the quality of house price clustering with K-means and the application of one-hot encoding in the data pre-processing process in representing categorical data. The dataset used has two types of data, namely numeric and categorical. The cluster evaluation is based on the silhouette score matrix and the determination of k is based on the elbow graph. The results showed an increase in the silhouette score value after applying one-hot encoding 0.15 which was previously 0.09, with the number of k = 3. The 0.15 matrix result is relatively low, which is caused by the overlap of house price values in the dataset, but it has been shown that one-hot encoding can represent categorical data well in the data pre-processing process so that the data can be processed with the k-means algorithm.

I. PENDAHULUAN

Rumah merupakan bagunan yang dibuat oleh manusia untuk melindungi diri dari berbagai hal, serta tempat untuk beraktivitas [1]. Dalam undang – undang No. 4 Tahun 1992 membahas mengenai apa arti dari rumah, rumah merupakan sebuah bangunan yang memiliki fungsi dasar sebagai tempat tinggal. Dalam buku Freedom To Build menurut John F.C Turner, 1972 rumah merupakan rangkaian dari permukiman secara utuh namun membutuhkan waktu untuk memperoleh rumah, karena rumah tidak bisa jadi dalam sekejap.[2].

Pemilihan rumah sangat penting bagi kehidupan karena sejatinya manusia akan bertahun – tahun menempati rumah yang ia miliki, di Indonesia terdapat 12,7 juta keluarga yang belum memiliki rumah[3]. Di Indonesia bahkan di dunia aspek yang dilihat oleh orang yang akan membeli rumah adalah biaya, mereka membutuhkan informasi harga rumah agar nantinya mereka bisa menentukan pilihan sesuai kondisi dan kebutuhan sehingga tidak akan terjadinya upprice.

Oleh karena itu, informasi terkait pengelompokan harga rumah sangat dibutuhkan agar calon pembeli rumah dapat mengetahui murah atau tidaknya harga rumah berdasarkan semua fitur yang didapat. Karena, Harga merupakan faktor penting dalam pemilihan rumah, karena harga rumah relative berbeda – deda berdasarkan kualitas dan fasilitas yang akan didapat seperti jumlah kamar tidur, jumlah kamar mandi, lokasi yang strategis, garasi dan luas tanah.

Berdasarkan permasalahan diatas, akan dilakukan olah data harga rumah. Olah data yang dimaksud adalah mengelompokkan data harga rumah menjadi beberapa kelompok. Dalam mengelompokkan data harga rumah ini menggunakan algoritma k-means. Teneshya Lidia Putri, dkk menerapkan metode K-Means dalam penelitian dengan topik Clusterisasi Harga Rumah. Dataset Harga Rumah di Jakarta yang bersumber dari website rumah123.com adalah data yang

digunakan dalam penelitiannya, dengan record data 1942 data dan atribut yang digunakan seperti kode pos, daerah, luas tanah, dan harga. Hasil dari penerapan metode K-Means adalah cluster optimal k=3 dengan nilai dbi sebesar 0.480[4].

Algoritma K-means merupakan sebuah metode yang berfungsi mengclusterkan data objek yang berdasarkan atribut menjadi sebuah k partisi, dengan data yang bersifat numerik [5]. K-means memiliki keunggulan dalam hal komputasi dan mudah diimplementasikan, dapat menangani jumlah data yang besar serta fleksibel terhadap semua jenis data[6][7], seperti pengelompokkan pelanggan, analisis segmentasi pasar, pemrosesan gambar

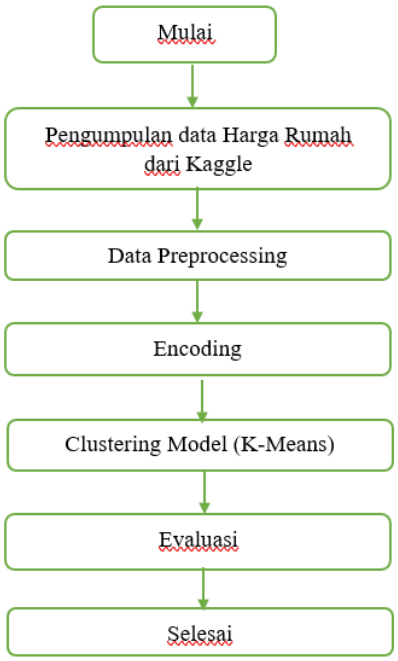
Penelitian terkait algoritma k-means telah banyak dilakukan untuk mengelompokkan data berdasarkan karakteristik kemiripan anatar fitur. Seperti, penelitian yang dilakukan oleh Briyan gifari aji, dan kawan - kawan dengan menggunakan metode k-means. Penelitian dengan studi kasus data harga rumah di bandung bersumber dari website Kaggle.com. Hasil k-means dalam clusterisasi harga rumah di bandung adalahi nilai cluster optimal k=2 dengan berdasarkan nilai silhouette score sebesar 0.887 [1] dan k-means tidak bisa memproses data bersifat kategorik, dengan menerapkan tahap preprocessing data dengan encoded dan one – hot encoding data yang bertipe non numeric dapat dikonversi menjadi data bertipe numerik agar data dapat diolah menggunakan algoritma klusterisasi seperti k-means [8]. Hasil pengelompokkan terbaik, diperoleh berdasarkan metrics evaluasi menggunakan silhouette score, digunakan sebagai dasar informasi terkait pola kelompok harga rumah. Informasi ini sangat berharga bagi pembisnis bidang properti dalam memahami segemtasi pasar untuk menentukan strategi bisnis yang lebih efektif.

Dalam penerapan preprocessing data bertipe kategorikal dengan one-hot encoding untuk mengkonversi data yang bertipe kategorik menjadi numerik agar dapat diolah menggunakan algoritma data mining. Penelitian lain terkait penerapan one-hot encoding pernah dilakukan silviana dan kawan - kawan, dengan dataset kasus covid-19 di Riau 2019 menggunakan one-hot encoding dalam mengoptimalisasi hasil silhouette score k-means untuk mengclusterisasikan daerah yang memiliki potensi rawan covid-19. Hasil dari penelitian ini didapatkan nilai silhouette score cukup tinggi yakni 0,7 dapat disimpulkan bahwa Teknik one-hot encoding dapat menghasilkan nilai yang tinggi sehingga hasil dapat dijadikan sebagai dasar informasi dalam penelitian tersebut terkait pengelompokkan data covid-19 di Riau [9]

Penelitian mengenai konversi data yang dilakukan pada tahap *preprocessing* pernah dilakukan oleh Cevi Herdiana dkk, untuk meningkatkan hasil akurasi dalam sebuah algoritma *Linear Regresi* dengan *Encod* dan *One-hot Encoding*, dengan topik Perbandingan nilai akurasi menggunakan dataset berupa Data Teks. Tekning *Encoding* dilakukan untuk menghasilkan sebuah data yang bertipe numerik. Hasil yang diperoleh dari penelitian ini adalah *One-hot Encoding* terbaik dalam meningkatkan hasil sebesar 0.85%[10]. Tujuan dari penelitian ini untuk mengetahui seberapa baik penerapan *one-hot encoding* dalam proses *preprocessing* harga rumah dengan algoritma *k-means.*

II. METODE

Metode penelitian ini berisi terkait konsep alur yang dilakukan oleh peneliti yang dilakukan dari awal hingga diperolehnya hasil agar memperoleh hasil terbaik dan terstruktur. Metode penelitian yang peneliti terapkan berdasarkan refrensi penelitian yang mengimplementasikan perhitungan K-Means oleh Silviana dkk, dengan dataset Kasus Covid-19 di Riau 2019. Peneliti menggunakan One-hot Encoding dalam preprocessing data, K-Means untuk mengklusterisasikan daerah yang memiliki potensi rawan virus covid-19 yang sama berdasarkan karakteristik gejala. Hasil dari penelitian ini didapatkan nilai Silhouette Score sebesar 0,7 dengan hasil tersebut dapat dikatakan teknik One-hot Encoding dalam algoritma K-Means dapat digunakan untuk mengelompokkan data yang mirip berdasarkan wilayah pandemic covid-19 [9].



Gambar 1. Metode Penelitian

A. *Tahap pengumpulan*

Tahap Pengumpulan data merupakan salah satu tahap untuk mencari informasi dan bahan yang dibutuhkan dalam penelitian ini. Data penelitian didapat dari situs open source terpercaya Kaggle.com yakni House Price Prediction Dataset yang bisa diakses melalui link berikut https://www.kaggle.com/datasets/zafarali27/house-price-prediction-dataset. Dataset yang digunakan merupakan data prediksi harga rumah , dengan jumlah record data sebanyak 2000 record dan terdiri 10 atribut. Isi dari dataset dapat dilihat pada Tabel dataset.

TABEL I.
DATASET

Id	Area	Bedrooms	Bathrooms	…	Location	Condition	Garage	Price
1	1360	5	4	…	Downtown	Excellent	No	149919
2	4272	5	4	…	Downtown	Excellent	No	424998
3	3592	2	2	…	Downtown	Good	No	266746
4	966	4	2	…	Suburban	Fair	Yes	244020
5	4926	1	4	…	Downtown	Fair	Yes	636056
6	3944	1	2	…	Urban	Poor	No	93262
7	3671	1	1	…	Rural	Poor	Yes	448722
…	…	…	…	…	…	…	…	…
2000	2989	5	1	…	Suburban	Fair	No	482525

Setelah mendapatkan data dilakukan analisis data yang bertujuan menentukan atribut apa saja yang digunakan dalam penelitian, berikut atribut data yang digunakan peneliti dapat dilihat pada tabel atribut.

TABEL 2.
TIPE DATA

No	Nama Atribut	Tipe Data
1.	Area	Numerik
2.	Bedrooms	Numerik
3.	Bathrooms	Numerik
4.	Floors	Numerik
5.	YearBuilt	Numerik
6.	Location	Kategorik
7.	Condition	Kategorik
8.	Garage	Kategorik
9.	Price	Numerik

B. Data Preprocessing

Preprocessing data membersihkan data dari noise, missing value serta mengubah data yang tidak relevan [11], proses preprocessing data sangat perlu dilakukan agar mendapatkan hasil yang clustering yang optimal [12]. Dalam data penelitian harga rumah ini tidak ditemukan missing value ataupun noise, selanjutnya dilakukan proses encoding untuk mengubah data kategorik kedalam numerik. Dengan jumlah data 2000 records.

Encoding dalam penelitian ini ada 2 yakni encoded dan one-hot encoding.

1. Label encod

Label encod Teknik untuk mengubah data tipe kateorik kedalam numerik berdasarkan urutan nilai setiap variabel [10]. Dalam penelitian ini *encoded* diterapkan pada atribut *condition* karena dalam dataset tersebut kondisi merupakan berdasarkan suatu hal yang dapat diukur dengan nilai seperti: *excellent, good, fair dan poor.*

2. One-hot encoding

One-hot encoding mengubah atau mengkonveris data kategorik kedalam numerik dengan cara merepresentasikan nilai dengan 0 dan 1 [9]. 0 apabila nilai data tidak sesuai dengan nilai variabel dalam atribut, dan nilai 1 untuk nilai variabel yang sesuai dengan atribut, dengan cara membuat atau memuarakan kolom baru, dengan membuat kolom baru berdasarkan nilai atribut kolomnya.

Atribut yang menerapkan *one-hot encoding* adalah: location, garage. Berikut data yang sudah dilakukan pengkonversian nilai pada tabel preprocessing dibawah ini pada atribut lokasi.

TABEL 3
PRPROCESSING ONE-HOT ENCODING

Location_Rural	Location_Suburban	Location_Urban	Location_Rural
0	0	0	0
0	0	0	0
0	0	0	0
0	1	0	0
0	0	0	0
0	0	1	0

Berikut adalah penjelasan table preprocessing terkait hasil proses one-hot encoding dimana semua nilai akan 0 jika nilai atribut baru tidak sesuai dengan nilai variabel. Seperti kolom location suburban akan berisi 0 karena kemungkinan nilai pada record data tersebut adalah lokasi lain yakni diantaranya lokasi lain dari suburban adalah location downtown, location ural, location rural. Berlaku untuk kolom atribut garage_yes.

TABEL 4
PREPROCESSING ENCOD

Area	Location	…	Garage
1360	0	…	0
4272	0	…	0
3592	2	…	0
…	…	…	…
2989	2	…	0

Berikut adalah penjalasan tabel preprocessing encod terkait hasil proses encoded data bertipe kategorik telah diubah dengan numerik dengan memberikan nilai atas dasar urutan atau bisa disebut dengan data yang memiliki tingkat level kelas.

C. Clustering model K-means

Clustering pada data harga rumah ini menggunakan algoritma k-means. Atribut dipilih digunakan untuk mendapatkan kelas dari harga rumah. Data harga rumah tersebut akan dikelompokkan ke dalam cluster berdasarkan jarak ke centroid, dan proses ini diulangi sampai centroid stabil atau tidak ada perubahan signifikan dalam pengelompokan data sehingga menghasilkan pengelompokkan data yang sesuai dan akurat. Hal tersebut dapat diselesaikan dengan melibatkan Teknik perhitungan diantaranya adalah menentukan nilai k secara random atau dengan elbow untuk pemilihan k dengan menggunakan elbow dapat dilakukan dengan rumus

$$SSE = \sum_{k=1}^{k} |x_i - c_k|^2 \quad (2.1)$$

memilih centroid secara acak dan yang terakhir menghitung dengan rumus Euclidean distance.

$$d(x,y) = \sqrt{\sum_{i=1}^{n}} \ (x_i - y_i) \quad ^2 = \sqrt{[(x_i - x_j)^2 + (y_i - y_i)^2]} \quad (2.2)$$

D. Evaluasi

Pada tahap evaluasi kluster optimal yang diperoleh akan dilakukan pengukuran performa dengan berdasarkan matrik silhouette score. Rumus dari matriks silhouette score adalah.

$$s(K) = \frac{1}{|k|} \sum_{i=1}^{k} s(i) \quad (2.3)$$

Proses dari tahap preprocessing sampai dengan evaluasi dilakukan dengan menggunkan tools google collab.

III. HASIL DAN PEMBAHASAN

Berikut penelitian terkait penelitian pentingnya clustering dengan k-means. Pada Penelitian terkait metode K-Means dalam studi kasus clusterisasi harga rumah juga pernah dilakukan oleh Nuraeni Septiani, dkk dengan topik Implementasi K-means Harga Rumah. Pada penelitian ini menggunakan dataset Harga Rumah di Jakarta selatan 2020, dengan jumlah data record 1010 data. Hasil clusterisasi dengan K-Means harga rumah dikelompokkan menjadi 10 kluster dengan nilai dbi 0.129[13]. Selanjutnya Penelitian lain terkait clusterisasi dengan metode K-Means harga rumah pernah dilakukan oleh Zyhan Faradilla Daldiri, dkk dengan topik pengelompkan harga rumah di Jakarta. Pada penelitian ini data yang digunakan bersumber dari website Kaggle.com menggenai daftar harga rumah di Jakarta, dengan atribut data sebanyak 8 atribut diantaranya, nomor data, nama rumah, harga, luas bangunan, luas tanah, jumlak kamar, jumlah kamar mandi dan luas garasi. Hasil dari penerapan k-means untuk data harga rumah di Jakarta berupa jumlah kluster k = 5 berdasarkan Silhouette score 0.626[14].

Kemudian penelitian terkait pentingnya Teknik one-hot encoding pada tahap preprocessing sebelum dilakukan pemodelan menggunakan algoritma data mining. Penelitian terkait penggunaan Teknik One-hot Encoding pernah dilakukan oleh Zahra risky fadilah dkk. Topik dari penelitian yang dilakukan adalah perbandingan teknik olah data bertipe campuran. Dataset yang digunakan dalam penelitian ini Chronic Kidney Disease dengan hasil penelitian penggunaan Teknik One-hot Encoding dapat meningkatkan hasil dari perhitungan menggunkan algoritma K-Means, dan hasil Silhouette Score terbaik diperoleh dari K-Prototype 0,3796 [8]. Dan Penelitian terkait penerapan One-hot encoding pada tahap preprocessing dilakukan oleh Bin-Bin Jia dan Min-Ling Zhang dengan dataset Benchmark seperti data klasifikasi teks. Hasil dari penelitian ini mengatakan bahwa penerapan One-hot Encoding memberikan pengaruh yang signifikan terhadap keberhasilan kinerja sebuah metode Machine learning, karena dapat merepresentasikan setiap kategori data dengan jelas menjadi biner 1 dan 0[15].

Pada sub bab hasil dan pembahasan ini akan dijelaskan terkait perihitungan atau analisis data harga rumah dengan *k-means one-hot encoding* dengan *tools google collab.*

A. Preprocessing Data

Preprocessing merupakan tahap awal dalam analisis data yang bertujuan untuk meningkatkan kualitas data dan sudah siap uji dengan algoritma yang dipilih dengan melakukan pengamatan apakah data yang akan di olah mengandung *missing value, noise, atau outlier* serta implementasi metode *encoding* pada proses Preprocessing data.

1. Normalisasi data One-hot encoding

Berikut merupakan table data normalisasi one-hot encoding dengan data penelitian.

TABLE 5
NORMALISASI ONE-HOT ENCODING

Location_Rural	Location_Suburban	Location_Urban
-0,55733	-0,56426	-0,5658
-0,55733	-0,56426	-0,5658
-0,55733	-0,56426	-0,5658
-0,55733	1,772226	-0,5658
-0,55733	-0,56426	-0,5658

Normalisasi ini dilakukan untuk memastikan bahwa semua atribut dalam penelitian ini memiliki skala rentang nilai yang sama agar terhindar dari dominasi fitur tertentu dengan menghitung mean dan standard deviation.

2. Normalisasi data Encod pra-pemrosesan tanpa one-hot encoding

Berikut merupakan table data encod yang telah dinormalisasikan.

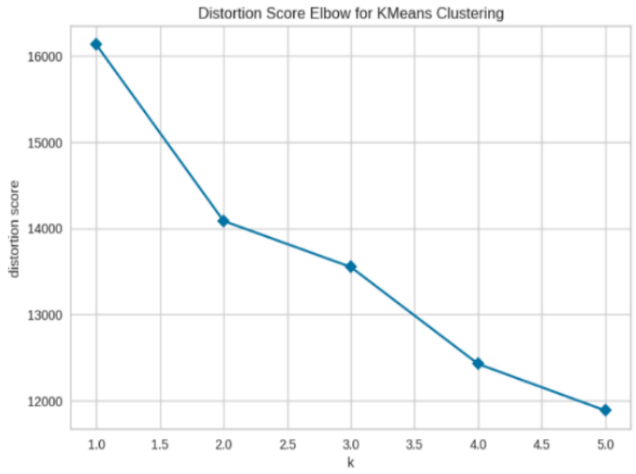
TABLE 6
NORMALISASI ENCOD

No	Location	Price
1.	-1,27433	-1,40309
2.	-1,27433	-0,40772
…	…	…
2000.	0,486403	-0,19957

Berdasarkan table 5 dan 6 yang telah disajikan diatas, sebenarnya nilai normalisasi untuk semua atribut sama, yang berbeda hanya pada atribut location dan garage. Karena one-hot encoding menghasilkan nilai biner yakni 0 dan 1 sehingga normalisasi tidak akan memberikan perubahan yang besar terhadap setiap nilai di atribut tersebut. Sedangkan encoding menghasilkan nilai berdasarkan rentang dan lebih besar, sehingga apabila dilakukan normalisasi maka perubahan nilai akan terjadi kedalam skala rentang yang lebih kecil.

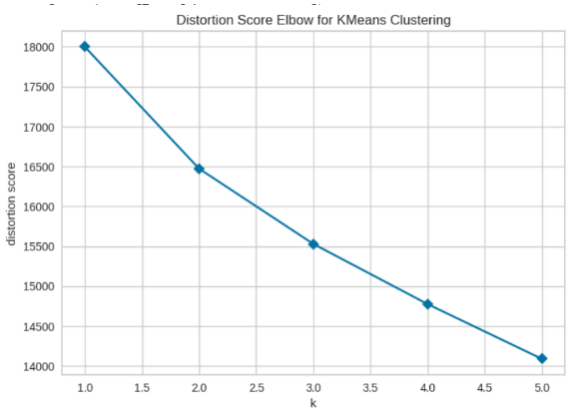
B. *Penetuan Nilai K*

Penentuan k optimal dapat dilakukan berdasarkan hasil titik siku elbow menggunakan nilai k =1 sampai k=n. untuk melihat k optimal yang dihasilkan dari metode elbow dapat dilihat pada gambar 2.



Gambar 2 *Elbow One-hot Encoding*

Berdasarkan visualisasi gambar 2 dengan data yang telah dilakukan preprocessing dengan one-hot encoding, diperoleh bahwa k optimal = 3 dapat dilihat berdasarkan letak titik siku pada grafik.



Gambar 3 Elbow Encod

Pada gambar 3 grafik elbow tidak memberikan hasil klustering yang optimal. Sebab grafik terus menurun dengan jumlah banyaknya kluster dan tidak menunjukkkan adanya titik siku, sehingga dibutuhkan evaluasi tekait nilai kluster dengan silhoute score agar dapat mengetahui seberapa akurat pembagian kluster yang terbentuk.

C. *Pemodelan K-Means*

Pemodelan dengan algoritma k-means untuk mengelompokkan data harga rumah dengan jumlah cluster 3 berdasarkan nilai k yang telah di evaluasi menggunakan matriks silhouette score (pada table nilai sil nanti). Berikut merupakan implementasi kode pemodelan algoritma k-means.

```
kmeans = KMeans(n_clusters=3,
random_state=42)
df3['Cluster'] =
kmeans.fit_predict(df3_numeric)
```

Gambar 4 Pemodelan K-means

Penjelasan terkait implementasi code pemodelan k-means.

KMeans(n_clusters=3, random_state=42): Inisialisasi model K-Means dengan 3 cluster dan random state untuk reproducibility.

kmeans.fit(df1[numerik_kolom]): Melakukan clustering pada data yang telah dipilih.

df1['Cluster'] = kmeans.labels_: Menambahkan kolom Cluster ke dataset untuk menyimpan label hasil clustering.

Berdasarkan hasil pemodelan k-means pada data yang telah dilakukan one-hot encoding bahwa data berhasil

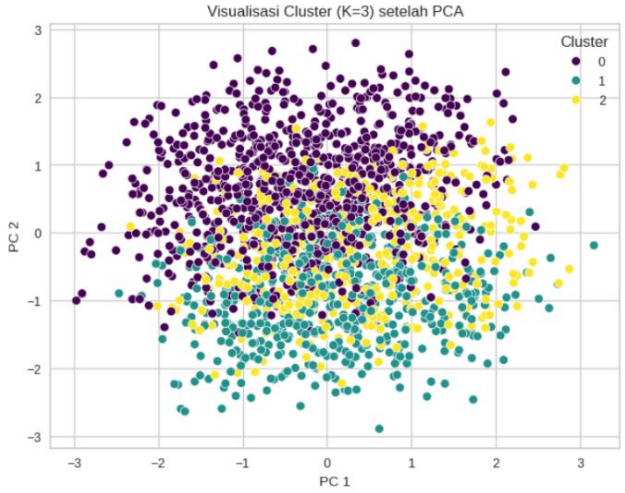
dikelompokkan kedalam 3 kluster. Yakni mahal, murah, dan sangat murah,

D. Hasil Pengujian Dataset

Setelah dilakukan preprocessing data baik dengan ataupun tanpa one-hot encoding, selanjutnya dilakukan algoritma k-means yang digunakan peneliti dalam mengklasterisasikan harga rumah.

1. Hasil pemodelan Algoritma K-means pada data preprocessing tanpa one-hot encoding.

Hasil penelitian ini divisualisasikan dengan menggunakan *principal component analysis* (PCA). Umunya PCA digunakan seblum pemodelan untuk melakukan reduksi dimensi, namun dalam penelitian ini pca digunakan setelah pemodelan yang bertujuan untuk memvisualisasikan hasil klasterisasi dengan dua dimensi yang tidak berpengaruh terhadap performance k-means[16]. Sebab, model telah dilatih menggunakan data asli yang telah dilakukan pra-pemrosesan dalam bentuk nyata tanpa reduksi. Berikut gambar tampilan hasil klasterisasi dengan PCA pada dataset pra-pemrosesan tanpa one-hot encod.



Gambar 5 PCA Hasil klasterisasi tanpa one-hot encoding

Pada gambar hasil kluaterisasi pra-pemrosesan tanpa one-hot encoding, sebaran data sangatlah bertumpang tindih. Meskipun telah dilakukan inisialisasi nilai k 3 berdasarkan titik siku elbow, yang artinya sebaran tidak terbagi secara jelas memungkinkan data yang berbeda masuk kedalam kluster lain. Hal ini menunjukkan bahwa representasi nilai data kategorik tanpa one-hot encoding belum menghasilkan pemisahan kluster dengan baik.

Analisis terkait hasil klasterisasi tanpa one-hot encoding, dilakukan dengan matriks nilai silhouette score serta nilai inertia. Hal ini bertujuan untuk melihat dan mengukur seberapa baik pemisahan kluster dan kepadatan anggota yang diperoleh dari pengujian dengan algoritma k-means. Berikut hasil evaluasi dapat dilihat pada tabel evaluasi matrik tanpa one-hot encoding.

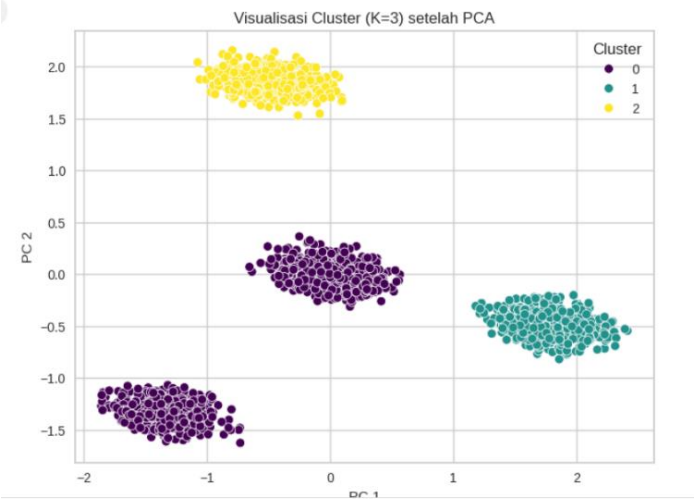
TABEL 8
MATRIKS EVALUASI ENCOD

No.	Cluster	Silhouette score	Inertia
1.	2	0.11	15992
2.	3	0.09	15113
3.	4	0.09	14360
4.	5	0.09	13807

Dari tabel evaluasi matrik tanpa one-hot encoding diperoleh hasil score silhouette tertinggi pada cluster 2 sebesar 0.11. namun nilai tersebut merupakan nilai yang tergolong rendah dalam merepresentasikan nilai pemisah paling baik dalam k-means ini karena kluster yang baik atau optimal nilai silhouette score seharusnya diatas 0.50. Dengan nilai inertia relative menurun saat jumlah kluster bertambah menunjukkan bahawa data akan semakin terbagi.

2. Hasil pemodelan Algoritma K-means pada data preprocessing tanpa one-hot encoding.

Seperti hasil pemodelan k-means tanpa one-hot encoding yang divisualisasikan menggunakan pca, pemodelan dengan k-means pada data yang dilakukan pra-pemrosesan dengan one-hot encoding juga divisualisasikan menggunakan pca guna melihat persebaran cluster secar dua dimensi. Berikut adalah gambar tampilan hasil klasterisasi dengan PCA pada dataset pra-pemrosesan dengan one-hot encoding.



Gambar *6* PCA Hasil klasterisasi dengan one-hot encoding

Berdasarkan visualisasi hasil klasterisasi k-means pra-pemrosesan dengan teknik one-hot encoding, pembagian kluster sudah semakin terlihat jelas dan jauh lebih baik dari hasil yang pertama. Hal ini merupakan tanda bahwa representasi data kategorik dengan one-hot encoding

merupakan salah satu cara yang cukup tepat dalam studi kasus ini, terlihat dari perbedaan visualisasi persebaran kluster sehingga dapat dikatakan bahwa penerapan one-hot encoding dapat membantu meningkatkan kualitas kluster dan dapat membagi kluster dengan pemisah yang cukup jelas.

Analisis terkait hasil klasterisasi pra-pemrosesan dengan one-hot encoding, dilakukan dengan matriks silhouette score serta nilai inertia. Hal ini beretujuan untuk melihat dan mengukur seberapa baik pemisahan kluster dan kepadatan anggota yang diperoleh dari pengujian dengan algoritma k-means. Berikut hasil evaluasi dapat dilihat pada tabel evalusi matriks one-hot encodin.

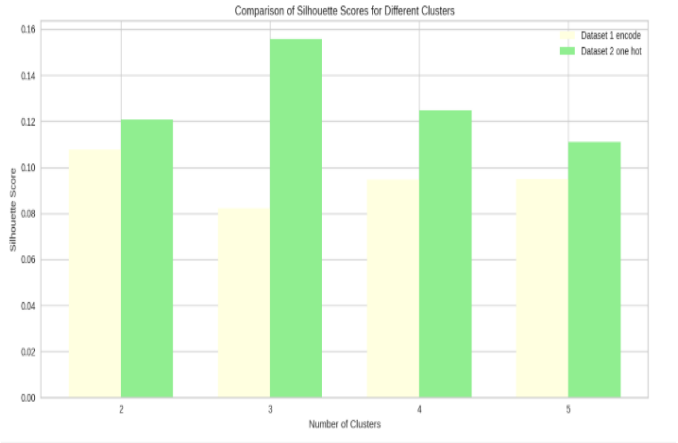
TABEL 9
MATRIKS EVALUASI ONE-HOT ENCODING

No.	Cluster	Silhouette score	Inertia
1.	2	0.12	19589,23
2.	3	0.15	17401,11
3.	4	0.14	15979,99
4.	5	0.11	15422,33

Dari hasil tabel matrik evaluasi one-hot encoding diperoleh hasil silhouette score tertinggi pada k=3 sebesar 0.15. nilai silhouette score 0.15 memiliki peningkatan dibandingakan nilai silhouette score pra-pemrosesan tanpa one-hot encoding, hal ini memberikan hasil bahwa one-hot encoding lebih baik dalam merepresentasikan data sehingga menghasilkan pembagian kluster yang lebih baik dari pada hasil pertama tanpa one-hot encoding. Selain itu nilai inerti juga terlihat jelas penurunnanya pada kluster k=3

E. Visualisasi dan analisis hasil clustering

Hasil penelitian ini, adalah clustering harga rumah menggunakan metode K-Means dengan dua pendekatan preprocessing yang dalam menangani variabel kategori, yaitu tanpa one-hot encoding (Label Encod) dan preprocessing dengan One-Hot Encoding.



Gambar 7 Diagram batang perbandingan hasil

Hasil clustering dari kedua metode encoding ditunjukkan pada diagram batang warna kuning untuk pra-pemrosesan tanpa one-hot encoding dan warna hijau untuk pra-pemrosesan dengan one-hot encoding.

IV. KESIMPULAN

Berdasarkan pembahasan yang telah diuraikan peneliti diperoleh hasil perbandingan nilai Silhouette score antara model clustering tanpa One-Hot Encoding dan dengan One-Hot Encoding menunjukkan adanya sedikit peningkatan pada nilai silhouette score yang mana hasil pra-pemrosesan tanpa one-hot encoding nilai paling tinggi terletak pada kluster 2 dengan nilai sebesar 0.11, sedangakn pra-pemrosesan dengan one-hot encoding memiliki peningkatan secara signifikan baik kluster 2, 3, 4, 5 dan yang terting dimiliki oleh k=3 sebesar 0.15. Meskipun nilai Silhouette score yang diperoleh tidak terlalu tinggi dan masih menunjukkan bahwa clustering belum sepenuhnya baik dan optimal, hal ini tetap menunjukkan bahwa One-Hot Encoding mampu memberikan kontribusi dalam memperbaiki pemisahan antar klaster, meskipun perbaikan tersebut terbatas. Hal ini menunjukkan bahwa untuk dataset ini, representasi kategorikal yang lebih mendalam dan berpengaruh melalui One-Hot Encoding ini dapat sedikit memberikan manfaat dalam membantu model memahami pola yang ada, meskipun masih diperlukan perbaikan lebih lanjut dalam penggunaan metode clustering yang lebih kompleks atau pengoptimalan parameter lainnya untuk mencapai hasil yang lebih baik.

DAFTAR PUSTAKA

[1] B. G. Aji, D. C. A. Sondawa, M. R. Gifari, and S. Wijayanto, “Penerapan Algoritma K-Means Untuk Clustering Harga Rumah Di Bandung,” *J. Ilm. Inform. Glob.*, vol. 14, no. 2, pp. 17–23, 2023, doi: 10.36982/jiig.v14i2.3189.

[2] dpu, “No Title,” 12 Maret 2019. [Online]. Available: https://dpu.kulonprogokab.go.id/detil/52/rumah-perumahan-dan-permukiman

[3] S. Y. Safitri Kiki, “No Title,” Kompas.com. Accessed: Feb. 18, 2025. [Online]. Available: https://money.kompas.com/read/2023/03/01/123000726/12-7-juta-rumah-tangga-belum-punya-rumah-jumlahnya-berpotensi-naik

[4] T. Lidia Putri and R. Danar Dana, “Penerapan Data Mining Pada Clustering Data Harga Rumah Dki Jakarta Menggunakan Algoritmak-Means,” *JATI (Jurnal Mhs. Tek. Inform.*, vol. 8, no. 1, pp. 1174–1179, 2024, doi: 10.36040/jati.v8i1.8957.

[5] N. Wahidah, O. Juwita, and F. N. Arifin, “Pengelompokkan Daerah Rawan Bencana di Kabupaten Jember Menggunakan Metode K-Means Clustering,” *INFORMAL Informatics J.*, vol. 8, no. 1, p. 22, 2023, doi: 10.19184/isj.v8i1.29542.

[6] P. M. Putri, L. Pujiastuti, I. Parlina, and Solikhun, “Pengelompokan Data Rasio Penggunaan Gas Rumah Tangga Berdasarkan Provinsi di Indonesia Menggunakan Metode K-Means Clustering,” *Semin. Nas. Teknol. Komput. Sains*, pp. 236–240, 2020.

[7] M. Barata, I. S. Ayuni, A. Y. Kartini, and Z. Alawi, “Algorima K-Means dalam Clustering Produk Skincare untuk Menentukan Strategi Pemasaran,” *J. Inform. Polinema*, vol. 10, no. 3, pp. 421–428, 2024, doi: 10.33795/jip.v10i3.5167.

[8] Z. R. Fadilah and A. W. Wijayanto, “Perbandingan Metode

Klasterisasi Data Bertipe Campuran: One-Hot-Encoding, Gower Distance, dan K-Prototype Berdasarkan Akurasi (Studi Kasus: Chronic Kidney Disease Dataset)," *J. Appl. Informatics Comput.*, vol. 7, no. 1, pp. 57–67, 2023, doi: 10.30871/jaic.v7i1.5857.

[9] Silviana et al., "STMIK Dian Cipta Cendikia Kotabumi," no. 1, 2022.

[10] C. Herdian, A. Kamila, and I. G. Agung Musa Budidarma, "Studi Kasus Feature Engineering Untuk Data Teks: Perbandingan Label Encoding dan One-Hot Encoding Pada Metode Linear Regresi," *Technol. J. Ilm.*, vol. 15, no. 1, p. 93, 2024, doi: 10.31602/tji.v15i1.13457.

[11] H. Syahputra, "Clustering Tingkat Penjualan Menu (Food and Beverage) Menggunakan Algoritma K-Means," *J. KomtekInfo*, vol. 9, pp. 29–33, 2022, doi: 10.35134/komtekinfo.v9i1.274.

[12] D. E. Kurniawan, and A. Fatulloh, 'Clustering of Social Conditions in Batam, Indonesia Using K-Means Algorithm and Geographic Information System', Int. J. Earth Sci. Eng., vol. 10, no. 05, pp. 1076–1080, 2017.

[13] N. Septiani and R. Herdiana, "Penerapan Algoritma K-Means Clustering Untuk Harga Rumah di Jakarta Selatan Nuraeni Septiani Sekolah Tinggi Manajemen Informatika dan Komputer (STMIK) IKMI Cirebon Saeful Anwar Sekolah Tinggi Manajemen Informatika dan Komputer (STMIK) IKMI Cirebon," *Trending J. Ekon. Akunt. dan Manaj.*, vol. 1, no. 2, 2023.

[14] Z. F. Daldiri, M. Rafly, and I. Veritawati, "Clustering Daftar Harga Rumah di Jakarta Dengan Algoritma K-Means," *J. Informatics Adv. Comput.*, vol. 3, no. 2, pp. 155–160, 2022, [Online]. Available: https://www.kaggle.com/datasets/wisnuanggara/daf

[15] B. Bin Jia and M. L. Zhang, "Multi-Dimensional Classification via Sparse Label Encoding," *Proc. Mach. Learn. Res.*, vol. 139, no. Mdc, pp. 4917–4926, 2021.

[16] M. Islam and M. Nasser, "PCA versus ICA in Visualization of Clusters," *Statru.Org*, no. October, pp. 978–984, 2012, [Online]. Available: http://www.statru.org/conference/wp-content/uploads/2012/01/000_Contrbuted_Part-2.pdf