Jurnal Komputer Multidisipliner

Vol.7 No.10, Oktober 2024
ISSN: 24559633

SEGMENTASI HARGA RUMAH DI KOTA BANDUNG MENGGUNAKAN ALGORITMA GAUSSIAN MIXTURE MODEL (GMM)

Syafira Istiara[1], Radiatun Maya Sari[2], Intan Aisyah Pratiwi[3], Muhammad Rafli Dewantara Siregar[4], Arnita[5]
Universitas Negeri Medan
E-mail: syafiraistiara@gmail.com[1], radiatunmayasari2@gmail.com[2], intanaisyahpratiwi11@gmail.com[3], mraflidewantarasiregar@gmail.com[4], arnita@unimed.ac.id[5]

Abstrak

Penelitian ini bertujuan untuk melakukan segmentasi harga rumah di Kota Bandung menggunakan algoritma Gaussian Mixture Model (GMM). Harga rumah merupakan salah satu indikator penting dalam ekonomi, dipengaruhi oleh faktor seperti lokasi, ukuran properti, dan kondisi ekonomi. Dengan menggunakan GMM, penelitian ini berhasil mengidentifikasi tiga klaster utama harga rumah yang berbeda. Klaster ini dapat membantu dalam memahami dinamika harga rumah di Bandung dan memberikan wawasan bagi pengambilan keputusan terkait investasi properti. Hasil menunjukkan bahwa GMM lebih efektif dibandingkan metode clustering lainnya, seperti K-Means, dalam menangani distribusi data yang kompleks. Purpose. Penelitian ini bertujuan untuk mengelompokkan harga rumah di Kota Bandung berdasarkan fitur-fitur yang mempengaruhi harga, seperti luas tanah, luas bangunan, dan jumlah kamar tidur. Pengelompokan ini diharapkan dapat memberikan wawasan yang lebih mendalam terkait variasi harga rumah di berbagai wilayah dan membantu pemangku kepentingan dalam membuat keputusan yang lebih baik terkait kebijakan perumahan dan investasi properti. Methods/Study design/approach. Pendekatan penelitian menggunakan algoritma Gaussian Mixture Model (GMM) untuk segmentasi data harga rumah. Data yang digunakan diambil dari dataset harga rumah di seluruh kecamatan Kota Bandung, diakses melalui Kaggle. Tahapan penelitian meliputi pengumpulan data, pra-pemrosesan data untuk menangani nilai hilang dan outlier, serta normalisasi data. Setelah itu, dilakukan proses clustering dengan GMM, dan hasilnya divisualisasikan menggunakan Principal Component Analysis (PCA) untuk mempermudah interpretasi klaster yang terbentuk. Result/Finding. Penelitian ini menghasilkan tiga klaster harga rumah yang berbeda di Kota Bandung: klaster rendah, menengah, dan tinggi. Klaster-klaster ini berbeda secara signifikan berdasarkan fitur seperti luas tanah, jumlah kamar tidur, dan jumlah kamar mandi. Evaluasi menggunakan metrik AIC dan BIC menunjukkan bahwa model GMM dengan tiga klaster adalah yang paling optimal untuk dataset ini. Visualisasi hasil klastering memperlihatkan pola distribusi harga rumah yang jelas di setiap klaster, yang memberikan informasi penting bagi analisis pasar properti. Novelty/Originality/Value. Penelitian ini memberikan nilai kebaruan dalam penerapan algoritma Gaussian Mixture Model untuk segmentasi harga rumah di Kota Bandung, yang belum banyak diterapkan sebelumnya. Metode GMM terbukti lebih fleksibel dan akurat dibandingkan metode lain seperti K-Means, terutama dalam menangani data dengan distribusi yang kompleks. Hasil penelitian ini memberikan kontribusi penting dalam bidang perencanaan kota dan pengambilan keputusan terkait properti, serta dapat digunakan sebagai referensi dalam penelitian selanjutnya di bidang real estate dan analisis data besar (big data).

Kata Kunci — Gaussian Mixture Model, Clustering, Home Received Month 2024 / Revised Month 2024 / Accepted Month 2024.

1. PENDAHULUAN

Rumah merupakan salah satu bentuk investasi yang menarik. Perumahan dan tempat tinggal juga merupakan kebutuhan dasar manusia untuk perbaikan martabat, kualitas hidup, penghidupan, dan sebagai refleksi dalam upaya meningkatkan taraf hidup dan pembentukan kepribadian kebangsaan dan karakter. Pertumbuhan penduduk dan kepadatan penduduk adalah dua factor yang mempengaruhi pembangunan pembangunan di Kota Bandung mengakibatkan harga rumah menjadi fluktuatif [1].

Harga rumah di Kota Bandung merupakan salah satu indikator penting dalam perekonomian yang mempengaruhi berbagai sektor, mulai dari perencanaan kota hingga kebijakan ekonomi makro. Fluktuasi harga rumah dipengaruhi oleh berbagai faktor seperti lokasi, ukuran properti, kondisi ekonomi regional, hingga perubahan dalam kebijakan moneter. Oleh karena itu, penting bagi para peneliti dan pengambil kebijakan untuk memahami pola dan karakteristik harga rumah guna mendukung pengambilan keputusan yang lebih baik.

Salah satu pendekatan yang dapat digunakan untuk memahami pola harga rumah adalah dengan melakukan pengelompokan (clustering) berdasarkan kemiripan fitur-fitur tertentu. Metode clustering memungkinkan untuk mengidentifikasi kelompok-kelompok yang memiliki karakteristik serupa sehingga pola pergerakan harga di setiap kelompok dapat dianalisis lebih lanjut.

Clustering adalah metode analisis data yang bertujuan untuk mengelompokkan objek-objek ke dalam kelompok-kelompok (cluster) berdasarkan karakteristik atau kemiripan di antara mereka. Objek-objek dalam satu cluster memiliki kemiripan yang tinggi, sementara objek-objek di antara cluster berbeda memiliki perbedaan yang signifikan. Dalam penelitian ini, kami menggunakan algoritma Gaussian Mixture Model (GMM) untuk melakukan clustering terhadap data harga rumah di Amerika.

Algoritma Gaussian Mixture Model merupakan metode analisis cluster non-hierarchical yang bekerja untuk memodelkan beberapa data dalam distribusi Gaussian dengan parameter mean dan varians tertentu [2]. GMM merupakan metode clustering yang lebih fleksibel dibandingkan metode lain seperti K-Means karena mampu menangkap variasi dalam bentuk distribusi data yang lebih kompleks. Berbeda dengan K-Means, yang biasanya membagi data menjadi kelompok-kelompok berbentuk bulat (spherical clusters), GMM dapat memodelkan distribusi yang lebih variatif, sehingga lebih sesuai untuk data yang memiliki pola harga yang tidak seragam. GMM dapat menangkap variasi dalam bentuk dan ukuran kelompok, sehingga menghasilkan pengelompokan yang lebih akurat dan representatif.

Algoritma Gaussian Mixture Model sendiri merupakan model statistik yang sangat populer dan umum digunakan[3]. Pada penelitian sebelumnya, ada penelitian yang meggunakan Gaussian Mixture Model untuk mengidentifikasi kepadatan kendaraan di jalan tol, dan menghasilkan tingkat akurasi sebesar 91% (Putra, 2017). Lalu terdapat penelitian yang melakukan perbandingan clustering cloud workloads dengan menggunakan dua metode yaitu K-Means dan Gaussian Mixture Model [4].

Dari penelitian tersebut menghasilkan Gaussian Mixture Model memberikan cluster yang lebih baik, dan lebih rinci daripada model yang lain [5]. Dengan menggunakan GMM, diharapkan dapat dihasilkan kelompok harga rumah yang lebih representatif, yang nantinya dapat digunakan untuk menganalisis perbedaan harga antar wilayah dan faktor-faktor yang memengaruhinya.

2. METODE PENELITIAN

Pengumpulan Data

Data yang digunakan dalam penelitian ini diambil dari dataset "Harga Rumah di Seluruh Kecamatan di Kota Bandung", yang diakses melalui platform Kaggle. Dataset ini dipublikasikan pada tahun 2023 dan dapat diakses melalui tautan berikut

https://www.kaggle.com/datasets/khaleeel347/harga-rumah-seluruh-kecamatan-di-kota bandung. Dataset ini berisi informasi mengenai harga rumah dan karakteristik properti di berbagai lokasi di Kota Bandung dan sekitarnya. Data terdiri dari 13.404 entri dengan 11 kolom yang mencakup berbagai informasi berikut:

- Price: Harga rumah (dalam Rupiah)
- Location: Nama lokasi atau kecamatan tempat rumah berada
- Bedroom: Jumlah kamar tidur
- Bathroom: Jumlah kamar mandi
- Carport: Jumlah tempat parkir mobil
- Land: Luas tanah (dalam meter persegi)
- Building: Luas bangunan (dalam meter persegi)
- City/Regency: Nama kota atau kabupaten tempat rumah berada
- Latitude: Koordinat lintang lokasi rumah
- Longitude: Koordinat bujur lokasi rumah

Data ini dikumpulkan dengan tujuan untuk menganalisis faktor-faktor yang mempengaruhi harga rumah di wilayah Kota Bandung. Informasi lokasi, serta ukuran tanah dan bangunan, diharapkan memberikan wawasan yang komprehensif mengenai variasi harga rumah di wilayah tersebut.

Pra-Pemrosesan Data

Pra-pemrosesan data merupakan langkah awal yang sangat penting dalam analisis data, terutama ketika berhadapan dengan dataset mentah yang seringkali mengandung inkonsistensi[6]. Proses ini bertujuan untuk meningkatkan kualitas data sebelum masuk ke tahap analisis atau pemodelan. Pra-pemrosesan mencakup beberapa langkah seperti pemuatan data, penanganan nilai hilang (missing values), deteksi outlier, dan normalisasi data. Setiap tahap dilakukan secara sistematis agar model yang dibangun mampu memberikan hasil yang lebih akurat. Tahapan-tahapan yang dilakukan dalam pra-pemrosesan data ini meliputi beberapa langkah penting sebagai berikut:

1. Pemuatan Data

Tahap pertama dalam pra-pemrosesan adalah pemuatan data ke dalam program analisis. Dalam penelitian ini, data dimuat menggunakan pustaka Pandas [7], yang merupakan pustaka standar untuk manipulasi data di Python. Pemuatan data ini penting karena memungkinkan data ditampilkan dalam format yang mudah diolah, seperti data frame yang terdiri dari baris dan kolom. Tahap ini memastikan bahwa data yang akan diproses memiliki struktur yang konsisten sehingga memudahkan proses analisis selanjutnya.

2. Penanganan Missing Value

Setelah data dimuat, langkah selanjutnya adalah menangani nilai yang hilang atau missing values. Kehadiran missing values dapat memengaruhi hasil analisis, sehingga penting untuk menangani masalah ini dengan baik. Ada beberapa metode yang dapat digunakan, termasuk menghapus baris yang memiliki nilai hilang atau melakukan imputasi dengan menggantinya menggunakan nilai rata-rata, median, atau modus [8]. Imputasi nilai hilang memungkinkan kita untuk meminimalkan distorsi dalam analisis tanpa kehilangan informasi yang signifikan dari dataset.

3. Mendeteksi Outlier

Outliers adalah data yang secara signifikan berbeda dari data lain di dalam dataset. Outlier dapat memengaruhi hasil analisis secara signifikan, terutama pada metode statistik yang sensitif terhadap nilai ekstrim. Oleh karena itu, penting untuk mendeteksi dan menangani outlier secara hati-hati. Metode yang umum digunakan untuk mendeteksi outlier adalah IQR. IQR (Interquartile Range) adalah metode statistik yang digunakan untuk mengukur variabilitas data dengan melihat rentang antara kuartil pertama (Q1) dan

kuartil ketiga (Q3) [9]. IQR sangat efektif dalam mendeteksi outliers karena nilai yang jatuh jauh di luar batas IQR dianggap sebagai pencilan atau data yang menyimpang dari pola umum. Rumus untuk menghitung IQR:

$$IQR = Q_3 - Q_1 \quad (1)$$

Setelah menghitung IQR, batas bawah dan batas atas untuk mendeteksi outliers dihitung dengan rumus sebagai berikut:

$$Batas\ Bawah = Q_1 - (1.5 \times IQR)$$
$$Batas\ Atas = Q_3 + (1.5 \times IQR) \quad (2)$$

Data yang berada di bawah Batas Bawah atau di atas Batas Atas dianggap sebagai outliers.

Selanjutnya untuk mengatasi outlier bisa menggunakan z-score. Z-Score adalah metode statistik yang digunakan untuk menstandarkan nilai berdasarkan jarak dari rata-rata populasi dalam satuan standar deviasi [10]. Z-Score digunakan untuk mengidentifikasi outliers dengan melihat data yang memiliki Z-Score sangat tinggi atau sangat rendah (biasanya Z-Score di atas 3 atau di bawah -3 dianggap sebagai outliers). Rumus untuk menghitung Z-Score adalah sebagai berikut:

$$Z = \frac{X-\mu}{\sigma} \quad (3)$$

Di Mana:

- Z adalah Z-Score,
- X adalah nilai data,
- μ\muμ adalah nilai rata-rata data,
- σ\sigmaσ adalah standar deviasi data.

Jika nilai Z-Score melebihi ambang batas tertentu (misalnya, Z > 3 atau Z < −3), maka data tersebut dianggap sebagai outliers. Setelah terdeteksi, outliers dapat diatasi dengan menghapus data tersebut atau menggantinya dengan nilai lain, seperti median atau rata-rata dari data yang tidak dianggap sebagai pencilan.

4. Normalisasi Data

Setelah outliers diatasi, tahap normalisasi dilakukan untuk memastikan data berada dalam skala yang seragam. Salah satu metode yang umum digunakan adalah MinMax Scaler. Metode ini merubah data ke dalam rentang yang ditentukan, biasanya antara 0 dan 1[11].

Rumus untuk MinMax Scaler adalah sebagai berikut:

$$Xscaled = \frac{X-Xmin}{Xmax-Xmin} \quad (4)$$

Di mana:

- Xscaled adalah nilai data setelah normalisasi,
- X adalah nilai data asli,
- Xmin dan Xmax adalah nilai minimum dan maksimum dari dataset

Eksplorasi Data

Setelah tahap pengolahan, dilakukan eksplorasi data untuk mendapatkan pemahaman awal mengenai distribusi variabel-variabel utama. Analisis deskriptif dilakukan untuk mengetahui distribusi harga rumah di seluruh kota, ukuran properti, dan hubungan antar variabel. Visualisasi seperti boxplot, dan heatmap digunakan untuk menggambarkan bagaimana fitur-fitur ini berhubungan satu sama lain, serta bagaimana harga rumah bervariasi tergantung pada variabel seperti luas tanah, jumlah kamar tidur, kamar mandi, dan luas bangunan. Eksplorasi ini membantu dalam memahami karakteristik dasar dari data dan memberikan panduan untuk langkah-langkah clustering selanjutnya.

Penerapan Gaussian Mixture Model (Gmm)

Clustering pada penelitian ini dilakukan menggunakan algoritma Gaussian Mixture Model (GMM), yang merupakan metode probabilistik untuk mengidentifikasi distribusi dalam data. Gaussian Mixture Model (GMM) adalah algoritma yang sangat berguna dalam

pemodelan data yang kompleks dan tidak dapat diatasi dengan pendekatan yang lebih sederhana, seperti k-Means. Gaussian Mixture Model (GMM) digunakan untuk melakukan klasterisasi data berdasarkan distribusi Gaussian [12]. Tahapan pertama dimulai dengan mengimpor library yang diperlukan, termasuk `GaussianMixture` dari `sklearn.mixture`, yang menyediakan implementasi GMM, dan `PCA` dari `sklearn.decomposition` untuk mereduksi dimensi data agar dapat divisualisasikan secara dua dimensi. Hal ini penting untuk mempermudah pemahaman hasil klastering secara visual, terutama ketika bekerja dengan data berdimensi tinggi.

Selanjutnya, dilakukan Principal Component Analysis (PCA) untuk mereduksi dimensi data ke dalam dua komponen utama (2D). PCA memungkinkan kita untuk merangkum informasi yang ada di dalam data dengan mempertahankan sebanyak mungkin variasi dari fitur aslinya [13]. Pada proses ini, data numerik diolah dan direduksi, menghasilkan dua komponen yang digunakan untuk memvisualisasikan hasil klastering.

Setelah mereduksi dimensi, model Gaussian Mixture Model dibangun dengan menggunakan `GaussianMixture` dari Scikit-learn. Parameter `n_components` menentukan jumlah klaster yang diinginkan, dalam hal ini dipilih tiga komponen (klaster). GMM bekerja dengan mengasumsikan bahwa data dihasilkan dari kombinasi beberapa distribusi Gaussian yang berbeda, dan parameter distribusi tersebut diestimasi menggunakan teknik Expectation-Maximization (EM). Proses ini mengoptimalkan probabilitas bahwa setiap titik data berasal dari distribusi Gaussian tertentu.

Setelah model dilatih dengan data numerik, hasil klastering dari GMM kemudian diterapkan kembali pada dataset. Untuk setiap titik data, algoritma GMM memprediksi klaster mana yang paling mungkin untuk dianggotai oleh titik tersebut. Informasi klaster ini ditambahkan ke dalam dataset sebagai fitur baru yang disebut `Cluster`.

Akhirnya, visualisasi hasil klastering dilakukan dalam ruang dua dimensi yang diperoleh dari PCA. Visualisasi ini memanfaatkan scatter plot dengan warna yang berbeda untuk tiap klaster. Dengan cara ini, distribusi titik data dalam klaster-klaster yang dihasilkan oleh GMM dapat divisualisasikan, memberikan gambaran yang lebih jelas tentang bagaimana data terkelompok dalam ruang berdimensi rendah.

Secara matematis, distribusi campuran Gaussian untuk data x dapat dinyatakan sebagai berikut:

$$log[p(X|\theta)] = \sum_{i=1}^{N} log \left(\sum_{j=1}^{K} w_j \, P(X|\mu_j, \sigma_j)\right) \quad (5)$$

Di mana:
N = banyaknya data
K = jumlah cluster
θ = parameter
w = bobot atau peluang mixing
X = data harga rumah
μ = mean
σ = variansi

Evaluasi Model

Setelah model GMM diterapkan, perlu dilakukan evaluasi untuk menilai kualitas clustering yang dihasilkan. Metrik seperti AIC digunakan untuk mengukur keseimbangan antara kualitas kecocokan model dan kompleksitas model (dalam hal ini jumlah parameter). Rumus AIC adalah sebagai berikut:

$$\mathrm{AIC} = 2k - 2\ln(\hat{L}) \quad (6)$$

AIC menyeimbangkan kecocokan model (diukur oleh likelihood L^) dan jumlah parameter k. Model yang lebih baik memiliki nilai AIC yang lebih rendah. Model dengan lebih banyak parameter cenderung memiliki likelihood lebih tinggi, tetapi AIC menghukum penggunaan parameter yang terlalu banyak untuk menghindari overfitting [14].

BIC mirip dengan AIC, namun memberikan penalti yang lebih besar terhadap jumlah parameter. BIC juga digunakan untuk memilih model terbaik dengan mempertimbangkan kecocokan model dan jumlah parameter [15]. Rumusnya adalah:

$$\mathrm{BIC} = k\ln(n) - 2\ln(\hat{L}) \tag{7}$$

Sama seperti AIC, model dengan BIC yang lebih rendah dianggap lebih baik. Karena ada penalti yang lebih besar untuk jumlah parameter dalam BIC, model yang dipilih cenderung lebih sederhana dibandingkan dengan yang dipilih oleh AIC.

Selain itu, Silhouette Score digunakan untuk menilai seberapa baik objek di dalam kluster berada. Nilainya berkisar antara -1 hingga 1, dengan nilai mendekati 1 menunjukkan bahwa data dikelompokkan dengan baik [16]. Rumusnya adalah sebagai berikut:

$$S(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \tag{8}$$

Di Mana:

- S(i) adalah Silhouette Score untuk titik data ke-i,
- a(i) adalah rata-rata jarak dari titik data i ke semua titik dalam kluster yang sama,
- b(i) adalah rata-rata jarak dari titik data i ke semua titik dalam kluster terdekat (kluster terdekat yang bukan milik i).

Jika nilai S(i) mendekati 1, maka titik data i berada dalam kluster yang tepat. Jika mendekati -1, maka titik data lebih dekat dengan kluster lain daripada kluster yang seharusnya.

3. HASIL DAN PEMBAHASAN

Tabel 1 menampilkan sebagian data harga rumah di Kota Bandung dan sekitarnya, yang mencakup informasi tentang harga, lokasi, jumlah kamar tidur, kamar mandi, carport, luas tanah, serta luas bangunan. Data ini mencakup beberapa properti di wilayah seperti Parongpong, Bojongsoang, Ngamprah, dan Lengkong. Sebagai contoh, properti di Parongpong memiliki harga sebesar Rp775 juta dengan 2 kamar tidur, 1 kamar mandi, 1 carport, luas tanah 100 m², dan luas bangunan 60 m².

Tabel 1. Data Harga Rumah di Kota Bandung

	Unnamed: 0	Price	Location	Bedroom	Bathroom	Carport	Land	Building	City/Regency	Latitude	Longitude
0	0	7.750000e+08	Parongpong	2	1	1	100	60	West Bandung Regency	-6.803228	107.581804
1	1	8.300000e+08	Bojongsoang	3	2	0	50	54	Bandung Regency	-6.993945	107.643700
2	3	5.500000e+08	Ngamprah	4	2	1	105	90	West Bandung Regency	-6.843730	107.508553
3	4	2.500000e+09	Lengkong	3	3	0	130	108	Bandung City	-6.934348	107.626219
4	5	2.450000e+09	Lengkong	4	3	2	100	170	Bandung City	-6.934348	107.626219

Tabel 2 menunjukkan deskripsi statistik dari data tersebut, yang merangkum informasi penting seperti jumlah data (count), nilai rata-rata (mean), standar deviasi (std), nilai minimum (min), persentil ke-25, median (persentil ke-50), persentil ke-75, dan nilai maksimum (max). Dari tabel ini, rata-rata harga rumah di Kota Bandung berada pada kisaran Rp2,511 miliar, dengan standar deviasi sebesar Rp1,864 miliar, yang menunjukkan adanya variasi harga yang signifikan antar properti. Nilai harga minimum sebesar Rp18,5 juta, sedangkan harga maksimum mencapai Rp9,25 miliar.

Tabel 2. Deskripsi Data

	Price	Bedroom	Bathroom	Carport	Land	Building
count	1.340400e+04	13404.000000	13404.000000	13404.000000	13404.000000	13404.000000
mean	2.511813e+09	3.421591	2.406819	1.161892	175.789317	171.633020
std	1.864898e+09	1.143596	0.919526	1.001778	103.701420	106.290707
min	1.850000e+07	1.000000	1.000000	0.000000	38.000000	37.000000
25%	1.120000e+09	3.000000	2.000000	1.000000	100.000000	94.000000
50%	1.950000e+09	3.000000	2.000000	1.000000	144.000000	150.000000
75%	3.300000e+09	4.000000	3.000000	2.000000	220.000000	220.000000
max	9.250000e+09	8.000000	4.000000	32.000000	588.000000	600.000000

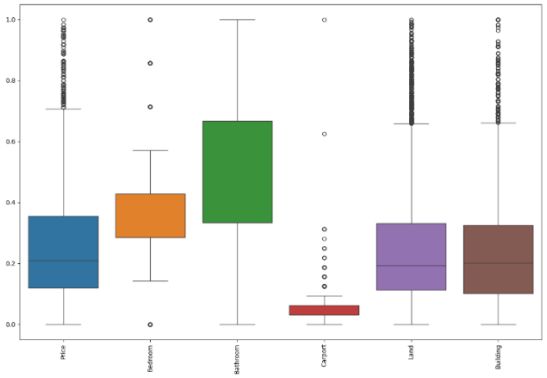
Dari sisi karakteristik rumah, jumlah rata-rata kamar tidur adalah 3,4 dengan standar deviasi 1,14, yang berarti sebagian besar rumah memiliki sekitar 3 hingga 4 kamar tidur. Jumlah kamar mandi rata-rata adalah 2,4, dan rata-rata properti memiliki 1,16 carport. Luas tanah rata-rata adalah 175,79 m^2, sementara luas bangunan rata-rata mencapai 171,63 m^2. Namun, terlihat ada variasi yang cukup besar untuk luas tanah dan bangunan, dengan nilai maksimum mencapai 588 m^2 untuk luas tanah dan 600 m^2 untuk luas bangunan.

Pada pra-pemrosesan data, dilakukan pemeriksaan terhadap keberadaan missing value pada seluruh variabel. Hasil pemeriksaan menunjukkan bahwa tidak ada missing value di dalam dataset ini. Hal ini berarti bahwa setiap entri pada setiap variabel telah terisi dengan data yang valid dan tidak ada informasi yang hilang.

Gambar 1 menunjukkan visualisasi boxplot dari beberapa variabel yang telah distandarisasi, meliputi Price, Bedroom, Bathroom, Carport, Land, dan Building. Dengan standarisasi, variabel-variabel ini berada dalam rentang nilai yang sama (0 hingga 1), sehingga distribusi antar variabel dapat lebih mudah dibandingkan.

- Variabel Price menunjukkan persebaran yang cukup besar, dengan rentang data yang luas, namun sedikit outlier pada batas bawah distribusi.
- Carport memiliki distribusi yang sangat terpusat di bagian bawah dengan sedikit outlier, menunjukkan bahwa sebagian besar properti memiliki carport dalam jumlah kecil atau bahkan tidak memiliki carport sama sekali.
- Variabel Land dan Building menunjukkan adanya banyak outlier, yang menunjukkan bahwa terdapat beberapa properti dengan luas tanah dan bangunan yang signifikan di luar rentang normal.

Gambar 1. Visualisasi Boxplot

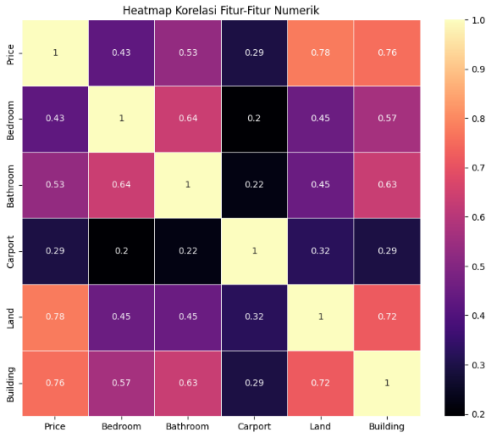


Gambar 2 menunjukkan heatmap korelasi antara variabel-variabel numerik yang ada dalam dataset. Heatmap ini memvisualisasikan kekuatan hubungan antara variabel-variabel tersebut menggunakan koefisien korelasi Pearson.

- Korelasi antara Price dan variabel Land serta Building sangat tinggi, dengan nilai korelasi masing-masing sebesar 0.78 dan 0.76. Hal ini mengindikasikan bahwa luas tanah dan bangunan adalah faktor penting dalam menentukan harga properti.
- Bedroom dan Bathroom memiliki korelasi sebesar 0.64, menunjukkan adanya keterkaitan antara jumlah kamar tidur dan kamar mandi di dalam rumah.
- Carport memiliki korelasi yang lemah terhadap variabel lainnya, terutama pada

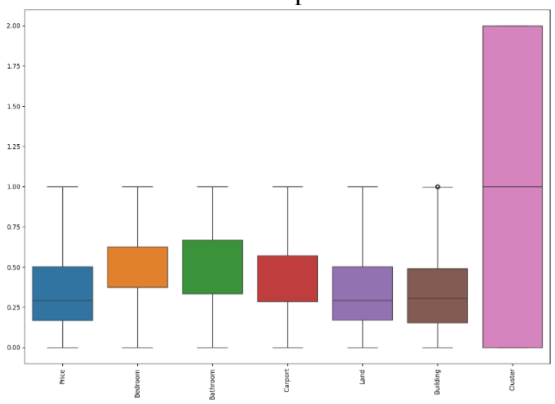
Bedroom dan Bathroom, dengan korelasi sekitar 0.20 hingga 0.22. Ini menunjukkan bahwa jumlah carport cenderung tidak berkaitan erat dengan karakteristik lain dari properti.

Gambar 2. Heatmap Korelasi antar Variabel



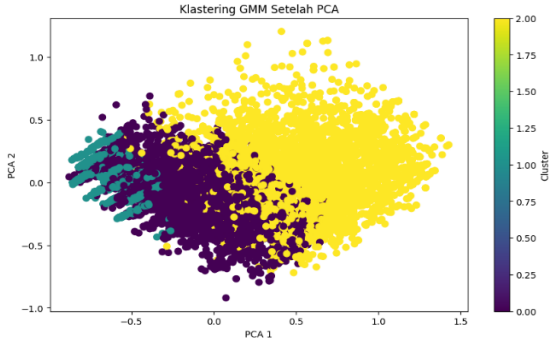
Gambar 3 menampilkan boxplot dari beberapa fitur dataset setelah dilakukan proses standarisasi. Boxplot ini menggambarkan persebaran data dari berbagai fitur seperti Price, Bedroom, Bathroom, Carport, Land, Building, dan Cluster. Hasil menunjukkan bahwa mayoritas fitur memiliki persebaran data yang relatif serupa, kecuali fitur Cluster yang memiliki skala berbeda dengan fitur lainnya.

Gambar 3. Visualisasi boxplot setelah standarisasi



Gambar 4 menampilkan hasil klastering menggunakan Gaussian Mixture Model (GMM) setelah dilakukan Principal Component Analysis (PCA). Pada visualisasi ini, data direduksi menjadi dua komponen utama, yaitu PCA 1 dan PCA 2, yang membantu memperjelas distribusi data dalam ruang dua dimensi. Terdapat tiga klaster yang teridentifikasi, ditandai dengan warna yang berbeda: ungu, hijau-biru, dan kuning. Klaster-klaster ini mengindikasikan bahwa data memiliki beberapa kelompok yang berbeda berdasarkan fitur-fitur utama yang ditemukan melalui PCA.

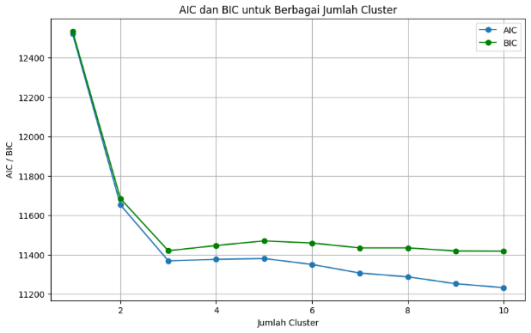
Gambar 4. Klastering GMM Setelah PCA



Dari hasil evaluasi, dapat disimpulkan bahwa model dengan 3 klaster adalah pilihan yang optimal berdasarkan nilai AIC dan BIC, karena terjadi penurunan yang tajam dari 1 klaster ke 3 klaster. Setelah itu, tidak ada perbaikan signifikan pada AIC dan BIC dengan menambah jumlah klaster, yang berarti model dengan lebih dari 3 klaster hanya menambah kompleksitas tanpa memberikan keuntungan yang signifikan dalam hal fit model.

Dengan demikian, model GMM dengan 3 klaster adalah yang paling sesuai untuk data ini, karena memberikan keseimbangan yang baik antara kompleksitas model dan kemampuan untuk memisahkan klaster.

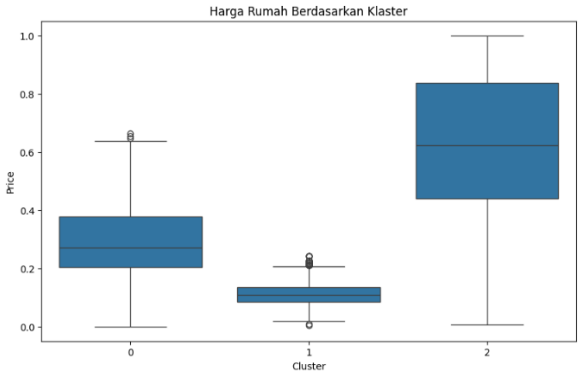
Gambar 5. Evaluasi dengan AIC dan BIC



Gambar menunjukkan distribusi harga rumah berdasarkan tiga klaster yang dihasilkan oleh Gaussian Mixture Model (GMM). Setiap klaster menunjukkan pola distribusi harga yang berbeda-beda, dengan karakteristik berikut:

- Klaster 0: Memiliki median harga yang lebih rendah dibandingkan dengan klaster lainnya, dengan rentang interkuartil yang lebih luas. Distribusi harga menunjukkan adanya beberapa pencilan (outliers).
- Klaster 1: Memiliki rentang harga yang sangat sempit dan berada di sekitar harga rendah, serta terlihat beberapa pencilan di luar batas distribusi utama.
- Klaster 2: Memiliki rentang harga yang paling lebar dengan median harga yang lebih tinggi. Klaster ini menunjukkan bahwa sebagian besar rumah dalam klaster ini memiliki harga yang lebih tinggi daripada klaster lain.

Gambar 6. Visualisasi harga rumah berdasarkan klaster



4. KESIMPULAN

Kesimpulan dari penelitian ini adalah bahwa penggunaan algoritma Gaussian Mixture Model (GMM) dalam segmentasi harga rumah di Kota Bandung berhasil mengidentifikasi tiga klaster harga yang berbeda: rendah, menengah, dan tinggi. Klaster ini memberikan gambaran lebih jelas mengenai distribusi harga rumah berdasarkan luas tanah, luas bangunan, dan fitur lainnya. Metode ini terbukti lebih fleksibel dibandingkan metode lain seperti K-Means, terutama karena GMM mampu menangani distribusi data yang lebih kompleks. Evaluasi menggunakan metrik AIC dan BIC menunjukkan bahwa

model dengan tiga klaster adalah yang paling optimal, dengan Silhouette Score menunjukkan kualitas pengelompokan yang baik. Hasil dari penelitian ini dapat memberikan wawasan penting bagi pembeli, penjual, dan pengembang properti dalam memahami dinamika harga rumah serta mendukung pengambilan keputusan strategis terkait pemasaran dan investasi properti.

DAFTAR PUSTAKA

B. C. Putra and Y. N. Afifah, "Gaussian Mixture Model Untuk Penghitungan Tingkat Kebersihan Sungai Berbasis Pengolahan Citra," Tek. Eng. Sains J., vol. 2, no. 1, p. 53, 2018, doi: 10.51804/tesj.v2i1.231.53-58.

C. M. Weber, D. Ray, A. A. Valverde, J. A. Clark, and K. S. Sharma, "Gaussian mixture model clustering algorithms for the analysis of high-precision mass measurements," Nucl. Instruments Methods Phys. Res. Sect. A Accel. Spectrometers, Detect. Assoc. Equip., vol. 1027, 2022, doi: 10.1016/j.nima.2021.166299.

D. H. Prakoso, B. Purnama, and F. Sthevanie, "Penghitungan Orang dengan Metode Gaussian Mixture Model dan Human Presence Map Studi Kasus: Penghitungan Orang di dalam Kelas," e-Proceeding Eng., vol. 2, no. 1, pp. 1562–1573, 2015.

D. K. Putra, I. I. Triasmoro, R. D. Atmaja, I. Iwut, and R. D. Atmaja, "Simulasi Dan Analisis Speaker Recognition Menggunakan Metode Mel Frequency Cepstrum Coefficient (MFCC) Dan Gaussian Mixture Model (GMM)," eProceedings Eng., vol. 4, no. 2, pp. 1766–1772, 2017, [Online]. Available: http://libraryeproceeding.telkomuniversity.ac.id/index.php/engineering/article/view/487/460

D. Y. Faidah, A. M. Hudzaifa, N. Theresia, and C. E. Widiantoro, "Optimalisasi Strategi Pengelompokkan Potensi Padi Sebagai Solusi Efektif Kelangkaan Beras Di Jawa Barat," J. Lebesgue J. Ilm. Pendidik. Mat. Mat. dan Stat., vol. 5, no. 1, pp. 529–537, 2024, doi: 10.46306/lb.v5i1.592.

E. Patel and D. S. Kushwaha, "Clustering Cloud Workloads: K-Means vs Gaussian Mixture Model," Procedia Comput. Sci., vol. 171, no. 2019, pp. 158–167, 2020, doi: 10.1016/j.procs.2020.04.017.

F. Amaluddin, M. Muslim, and A. Naba, "Klasifikasi Kendaraan Menggunakan Gaussian Mixture Model (GMM) Dan Fuzzy Cluster K Means (FCM)," J. EECCIS, vol. 9, no. 1, p. pp.19-24, 2015.

H. Vazirani, A. Kautsar, K. Adi, and J. Fisika, "Implementasi Object Tracking Untuk Mendeteksi Dan Menghitung Jumlah Kendaraan Secara Otomatis Menggunakan Metode Kalman Filter Dan Gaussian Mixture Model," Youngster Phys. J., vol. 5, no. 1, pp. 13–20, 2016.

I. A. Nafiudin, R. T. Hidayat, A. M. Putri, and A. R. Maulana, "Deteksi Jumlah Kendaraan Dengan Algoritma Gaussian Mixture Model Di Area Jalan Raya," Method. J. Tek. Inform. dan Sist. Inf., vol. 7, no. 1, pp. 37–44, 2021, doi: 10.46880/mtk.v7i1.258.

J. Riyono, S. D. Puspa, and C. E. Pujiastuti, "Simulasi Clustering Provinsi di Indonesia dalam Penyebaran Covid-19 Berdasarkan Indikator Kesehatan Masyarakat Menggunakan Algoritma Gaussian Mixture Model," MAJAMATH J. Mat. dan Pendidik. Mat., vol. 5, no. 1, pp. 43–60, 2022.

K. E. Setiawan and A. Kurniawan, "Pengelompokan Rumah Sakit di Jakarta Menggunakan Model DBSCAN, Gaussian Mixture, dan Hierarchical Clustering," J. Inform. Terpadu, vol. 9, no. 2, pp. 149–156, 2023, doi: 10.54914/jit.v9i2.995.

L. Rahmawati and K. Adi, "Rancang bangun penghitung dan pengidentifikasi kendaraan menggunakan Multiple Object Tracking," Youngster Phys. J., vol. 6, no. 1, pp. 70–75, 2017.

N. F. Deofanny, A. A. Rohmawati, and I. Indwiarti, "Model Gaussian Mixture Pada Distribusi Kecepatan Angin Dengan Algoritma Em," e-Proceedings Eng., vol. 9, no. 3, pp. 1978–1984, 2022.

R. Rahmattullah, Indwiarti Indwiarti, and A. A. Rohmawati, "Clustering Harga Rumah: Perbandingan Model K-Means dan Gaussian Mixture Model," e-Proceeding Eng., vol. 10, no. 3, pp. 3441–3449, 2023.

R. Ummami and B. Winarno, “Gaussian Mixture Model dengan Algoritme Expectation Maximization untuk Pengelompokan Data Distribusi Air Bersih di Jawa Barat,” Prism. Pros. Semin. Nas. Mat., vol. 6, pp. 745–750, 2023, [Online]. Available: https://journal.unnes.ac.id/sju/index.php/prisma/

Y. F. Munawaroh and I. Salamah, “Analisa Perbandingan Algoritma Histogram of Oriented Gradient (HOG) dan Gaussian Mixture Model (GMM) Dalam Mendeteksi Manusia,” Semin. Nas. Inov. dan Apl. Teknol. Di Ind. 2018, vol. 4, no. 2, pp. 251–255, 2018.