**Segmentasi Harga Jual Rumah Menggunakan Gaussian**

**Mixture Model-Based *Clustering***

**Tugas Akhir**

**diajukan untuk memenuhi salah satu syarat**

**memperoleh gelar sarjana**

**dari Program Studi Informatika**

**Fakultas Informatika**

**Universitas Telkom**

**NIM: 1301184079**

**Muhammad Hafidh Raditya**

****

**Program Studi Sarjana Informatika**

**Fakultas Informatika**

**Universitas Telkom**

**Bandung**

**2021/2022**

LEMBAR PENGESAHAN

**Segmentasi Harga Jual Rumah Menggunakan Gaussian**

**Mixture Model-Based *Clustering***

**House Selling Price Segmentation Using Gaussian**

**Mixture Model-Based *Clustering***

**NIM : 1301184079**

**Muhammad Hafidh Raditya**

Tugas akhir ini telah diterima dan disahkan untuk memenuhi sebagian syarat memperoleh gelar pada Program Studi Sarjana Informatika  
Fakultas Informatika

Universitas Telkom

Bandung, 23 Juli 2022

Menyetujui

|  |  |  |
| --- | --- | --- |
| Pembimbing I, |  | Pembimbing II, |
|  |  |  |
| Dra. Indwiarti, M.Si  NIP : 98690022 |  | Aniq Atiqi Rohmawati, S.Si., M.Si  NIP : 98690022 |

|  |
| --- |
| Ketua Program Studi Sarjana Informatika,  Erwin Budi Setiawan, S.Si., M.T.  NIP: 00760045-1 |

LEMBAR PERNYATAAN

Bandung, 23 Juli 2022

Yang Menyatakan

Muhammad Hafidh Raditya

Segmentasi Harga Jual Rumah Menggunakan Gaussian

Mixture Model-Based *Clustering*

Muhammad Hafidh Raditya1, Indwiarti2, Aniq Atiqi Rohmawati3

1,2,3Fakultas Informatika, Universitas Telkom, Bandung

1mhafidhraditya@students.telkomuniversity.ac.id, 2indwiarti@telkomuniversity.ac.id, 3aniqatiqi@telkomuniversity.ac.id

#### Abstrak

**Rumah merupakan tempat tinggal sekaligus salah satu kebutuhan pokok bagi manusia. Seiring berkembangnya zaman, kebutuhan akan rumah semakin tinggi dan bervariasi sehingga berpengaruh terhadap harga jual rumah. Oleh karena itu perlu adanya penelitian lebih lanjut mengenai harga jual rumah. Penelitian ini berfokus pada proses segmentasi harga jual rumah di Provinsi DKI Jakarta menggunakan metode Gaussian Mixture Model-Based Clustering dengan algoritma Expectation-Maximization. Tujuan dari penelitian ini adalah untuk membuat model segmentasi harga jual rumah sehingga dapat disimpulkan suatu informasi yang bermanfaat bagi para calon pembeli. Hasil yang didapatkan adalah rumah di provinsi DKI Jakarta dapat dibagi menjadi kedalam 3 cluster. Cluster pertama sebagai cluster dengan rumah berprofil kelas rendah. Cluster kedua sebagai cluster dengan rumah berprofil menengah. Cluster ketiga sebagai cluster dengan rumah berprofil kelas atas.**

**Kata kunci : Segmentasi, clustering, Gaussian Mixture Model**

**Abstract**

**Home is a place to live for humans. The house is also one of the basic human needs. Along with the development of the times, the need for houses is getting higher and varied so that it affects the selling price of the house. Therefore, there is a need for further research on the selling price of the house. This study focuses on the process of segmenting the selling price of houses so that this research can produce useful information for prospective home buyers. One way to do the segmentation process is to use the *clustering* technique. The *clustering* technique in this study will use the Gaussian Mixture Model (GMM) method because the *clustering* technique using the GMM method can produce a more flexible *cluster* form.**

**House is a place for humans to live and also a main necessity for humans. For years, the need for houses is increasing and varied so that it affects the selling price of the house. Therefore more research is needed to learn about the selling price of houses. This research is only focusing on house price segmentation in DKI Jakarta using the Gaussian Mixture Model-Based Clustering Method with the Expectation-Maximization algorithm. The goal of this research is to make a house price segmentation model so that we can obtain useful information for the potential buyer. The result of this research is houses in DKI Jakarta can be segmented into 3 different clusters. The first cluster is for the low-profile houses. The second cluster is for the mid-profile houses. The third cluster is for the high-profile houses.**

**Keywords: segmentation, *clustering*, Gaussian Mixture Model**

1. **Pendahuluan**

**Latar Belakang**

Rumah merupakan salah satu kebutuhan yang sangat penting bagi manusia dan termasuk ke dalam tiga kebutuhan primer manusia yaitu sandang, pangan, dan papan. Rumah merupakan tempat tinggal sekaligus tempat berlindung bagi manusia. Seiring berkembangnya zaman, kebutuhan manusia untuk memiliki rumah juga semakin tinggi, begitu juga dengan harga rumah yang semakin bervariasi bergantung pada parameter yang dimiliki tiap rumah seperti luas rumah, daerah tempat rumah, jumlah ruangan, dll.

Karena harga rumah yang semakin bervariasi, perlu dilakukan penelitian lebih lanjut untuk menganalisis keterkaitan antara harga rumah terhadap parameter-parameter lain yang dimiliki rumah tersebut. Analisis yang dilakukan diharapkan dapat mengelompokkan rumah ke dalam beberapa kategori sehingga orang-orang yang ingin membeli rumah dapat memiliki wawasan yang cukup dalam membeli rumah yang sesuai dengan kebutuhan dan anggaran yang mereka miliki. Sebelumnya sudah ada yang melakukan penelitian tentang segmentasi dan penentu harga rumah menggunakan metode *hedonic regression* [1].

*Clustering* merupakan salah satu proses unsupervised learning yang sangat penting dalam machine learning. *Clustering* merupakan suatu proses pengelompokan sekumpulan data yang “mirip” dalam satu *cluster* yang sama dan berbeda jika dibandingkan pada data yang berada dalam *cluster* yang lain [2]. *Clustering* biasanya digunakan dalam proses segmentasi suatu data, sehingga *clustering* juga sangat cocok untuk digunakan dalam penelitian ini.

*Hierarchical* *clustering* merupakan metode *clustering* pertama yang pernah digunakan yang digunakan oleh ahli biologi dan ilmuwan sosial, dimana pada saat itu analisis *cluster* telah menjadi cabang dari statistik analisis multivariat [3]. Sedangkan pada penelitian ini akan digunakan metode *Gaussian mixture model-based clustering*. *Gaussian mixture model* (GMM) merupakan sebuah model yang cukup sederhana untuk permasalahan classification maupun *clustering* jika dibandingkan dengan metode lain [4]. Pada akhir-akhir ini sudah banyak penelitian yang menggunakan GMM sebagai metode *clustering*. Sebagai contoh ada penelitian yang menggunakan *Gaussian mixture model-based clustering* untuk mengidentifikasi profil penggunaan listrik harian [5]. Hasil dari penelitian tersebut adalah GMM-based *clustering* dapat menghasilkan informasi berguna yang berkaitan dengan pola profil listrik harian dari sebuah gedung. Selain itu, penelitian ini juga menyatakan kalau GMM-based *clustering* menghasilkan biaya komputasi yang lebih rendah jika dibandingkan dengan hierarchical *clustering*. Tabel 1 menjelaskan perbandingan antara GMM dan *hierarchical clustering* menggunakan metode *silhouette.*

Tabel 1. Perbandingan GMM-*based clustering* dengan *hierarchical clustering*

|  |  |  |
| --- | --- | --- |
| **Algoritma *clustering*** | **Jumlah *cluster* optimal** | **Nilai skor *silhouette*** |
| GMM | 2 | 0,719 |
| *Hierarchical clustering* | 2 | 0,696 |

Kedua algoritma, baik GMM maupun hierarchical *clustering* memiliki jumlah *cluster* optimal yang sama yaitu sebanyak dua *cluster*. Namun jika dilihat skor *silhouette*-nya, GMM menghasilkan skor yang lebih besar jika dibandingkan dengan *hierarchical* *clustering*.

**Topik dan Batasannya**

* Bagaimana cara melakukan optimasi parameter dari GMM?
* Bagaimana hasil segmentasi harga rumah menggunakan GMMl-*Based Clustering* dan informasi apa saja yang bisa didapat?

**Tujuan**

* Mengidentifikasi cara melakukan optimasi parameter dari GMM
* Mengidentifikasi hasil segmentasi harga jual rumah menggunakan GMM-*Based* *Clustering* dan menganalisis hasil yang didapatkan.

1. **Landasan Teori**
   1. *Clustering*

*Clustering* merupakan teknik unsupervised learning dimana setiap objek dikelompokkan kedalam beberapa kelompok yang tiap objek di dalam suatu kelompok memiliki kemiripan sifat antara satu sama lain [6].

* 1. *Univariate Tukey Boxplot*

*Univariate Tukey Boxplot* merupakan teknik visualisasi data yang memungkinkan untuk merangkum karakteristik distribusi data univariat dengan visualisasi yang mudah dan cepat dipahami [7]. **Gambar** **1** adalah ilustrasi dari penggunaan *univariate Tukey boxplot.*

Chart

Description automatically generated

Gambar 1. Ilustrasi dari *univariate tukey boxplot*

* 1. *Deteksi Data Pencilan Menggunakan Interquartile Range*

*Interquartile* range (IQR) adalah sebuah teknik yang dapat membantu dalam proses pendeteksian data pencilan yang bersifat kontinu [8]. IQR pada dasarnya adalah kuartil pertama dan kuartil ketiga; IQR=Q3-Q1. Dalam mendeteksi data pencilan, perlu diidentifikasi terlebih dahulu *lower bound* dan *upper bound* dari data tersebut. Data yang berada diluar dari *lower bound* dan *upper bound* adalah data pencilan. Persamaan (1) dan (2) merupakan persamaan yang dipakai untuk mencari nilai dari *lower bound* dan *upper bound*.

* 1. *Gaussian Mixture Model*

*Gaussian Mixture Model* (GMM) merupakan sebuah fungsi probabilitas densitas parametrik yang direpresentasikan dengan sekumpulan komponen fungsi-fungsi Gaussian [9]. Pada *Gaussian Mixture Model-Based Clustering*, setiap *cluster* direpresentasikan melalui distribusi *gaussian* atau distribusi normal yang dideskripsikan dengan tiga parameter seperti *mean* (), standar deviasi (), dan bobot () [10].

Untuk setiap data dikelompokkan kedalam *cluster* berdasarkan fungsi *probability distribution* yang memenuhi persamaan (3)

Dimana *K* merupakan banyaknya *cluster*.

Namun yang menjadi permasalahan utama pada GMM adalah bagaimana menentukan nilai optimal dari ketiga parameter tersebut. Untuk itu diperlukan algoritma *Expectation-Maximization* dalam upaya optimasi ketiga parameter tersebut.

* 1. *Expectation-Maximization*

*Expectation-maximization (EM)* merupakan suatu algoritma untuk melakukan proses estimasi *log-likelihood* pada suatu parameter untuk memastikan konvergensi pada suatu fungsi [11]. Pada GMM, yang menjadi fokus utama adalah berapa peluang suatu data akan masuk kedalam satu *cluster* tertentu. EM memiliki proses sebagai berikut [12]:

1. Inisialisasi parameter *mean* (), standar deviasi (), dan bobot () secara acak lalu evaluasi fungsi *log-likelihood* menggunakan ketiga parameter tersebut yang memenuhi persamaan (4)[12].

Dimana *N* adalah banyaknya data dan *K* adalah banyaknya *cluster*.

1. E-step (Expectation)

Tahap ini menghitung nilai probabilitas () dari suatu data () untuk masuk kedalam suatu *cluster* yang memenuhi persamaan (5).

1. M-*step* (*Maximization*)

Tahap ini melakukan pembaruan pada ketiga parameter tersebut dengan menggunakan total nilai probabilitas () pada tiap iterasinya. Persamaan (6), (7), (8), dan (9) dipakai untuk mengoptimasi ketiga parameter GMM.

1. Evaluasi fungsi *log-likelihood* menggunakan ketiga parameter baru yang telah diperbarui. Ulangi Langkah 2-3 sampai nilai fungsi *log-likelihood* memenuhi kriteria konvergensi.
   1. *Koefisien Korelasi Pearson*

Koefisien korelasi Pearson merupakan sebuah koefisien korelasi linier dalam mengukur hubungan dan keterkaitan antara dua variabel, yang dikembangkan oleh Karl Pearson [13]. Misal dan merupakan dua variabel acak berbilangan *real* dan N merupakan banyaknya data. Koefisien korelasi Pearson dapat didefinisikan melalui persamaan (10) [13].

Ada beberapa poin penting yang dimiliki oleh koefisien korelasi Pearson yaitu [13]:

1. Range nilai dari adalah (-1,1).
2. Jika , berkorelasi positif.
3. Jika , tidak terdapat korelasi linier antara dan .
4. Jika , dan berkorelasinegatif.
5. Semakin besar nilai , maka tingkat korelasi linier antara dan juga semakin besar.
   1. *Bayesian Information Criterion*

*Bayesian Information Criterion* (BIC) merupakan salah satu metode yang digunakan untuk proses seleksi model [14]. Dalam penggunaan GMM-*based clustering*, BIC sangat berguna untuk mengetahui jumlah *cluster* optimum dari data yang dimiliki yang memenuhi persamaan (11)

Dimana adalah nilai log­-*likelihood* dari model yang dimiliki, adalah banyaknya data, dan adalah banyaknya *cluster*. Semakin rendah skor BIC yang dihasilkan maka *cluster* yang dibuat semakin baik.

* 1. *Akaike Information Criterion*

Sama seperti BIC, *Akaike Information Criterion* (AIC) merupakan salah satu metode yang digunakan untuk proses seleksi model dimana pada GMM-*based clustering*, AIC juga dapat digunakan untuk mencari jumlah *cluster* optimum dari data yang dimiliki. AIC memenuhi persamaan (12)[15]

dimana adalah banyaknya data, adalah banyaknya *cluster* dan adalah nilai dari *sum square error*. Sama seperti BIC, semakin rendah skor AIC yang dihasilkan maka *cluster* yang dibuat semakin baik.

* 1. *Silhouette*

*Silhouette* merupakan salah satu metode untuk menguji kualitas *cluster* yang telah dibuat pada proses *clustering*. *Silhouette* menghitung rata-rata nilai setiap data pada setiap *cluster*, dimana perhitungan nilai setiap data merupakan selisih antara nilai *separation* dan *compactness* yang dibagi dengan maksimum antara keduanya [16]. Persamaan (12) dipakai untuk mencari nilai dari skor *silhouette*[17].

Keterangan:

nilai *silhouette*

rata-rata jarak dari objek i dengan objek berada di *cluster* yang berbeda

rata-rata jarak dari objek i dengan seluruh objek dalam *cluster* yang sama

1. **Metodologi Penelitian**

Diagram

Description automatically generated

Gambar 2. Diagram alur metodologi penelitian

1. Pengumpulan Data

Dataset yang digunakan merupakan data dari sekumpulan harga jual rumah yang ada di provinsi DKI Jakarta. Dataset ini diambil melalui website [www.olx.co.id](http://www.olx.co.id) dengan menggunakan *tools* web scraping yang tersedia secara gratis melalui ekstensi Google Chrome. *Tools* ini memanfaatkan elemen HTML yang setiap elemennya dibuat spesifik untuk masing-masing atribut harga rumah. Atribut yang terdapat pada dataset dapat dilihat pada **Tabel 2**.

Tabel 2. Atribut dataset

|  |  |
| --- | --- |
| **Nama atribut** | **Deskripsi** |
| ALT | Luas tanah |
| ALB | Luas bangunan |
| JKT | Jumlah kamar tidur |
| JKM | Jumlah kamar mandi |
| Lantai | Jumlah lantai |
| Lokasi | Lokasi rumah, terdiri dari kecamatan, kotamadya, dan provinsi |
| Harga | Harga jual rumah |

Dataset ini dapat diunduh pada tautan berikut <https://bit.ly/3ztvo23>

B. Ekplorasi dan Praproses Data

Pada tahap ini akan dilakukan proses eksplorasi data, yaitu proses yang dilakukan untuk memahami data sebelum dilakukan praproses data. Praproses data merupakan sebuah proses persiapan dan transformasi data mentah menjadi kedalam bentuk atau *format* yang siap dipakai untuk dilakukan proses *modelling*[18]*.* Tahapan yang dilakukan pada praproses data antara lain:

1. **Tabel 3** adalah bentuk dari dataset yang masih mentah.

Tabel 3. Dataset mentah

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **ALT** | **ALB** | **JKT** | **JKM** | **lantai** | **lokasi** | **harga** |
| 60 | 60 | 3 | 2 | 2.0 | Cempaka Putih, Jakarta Pusat, Jakarta D.K.I. | Rp 550.000.000 |
| 87 | 85 | 3 | 2 | 2.0 | Sawah Besar, Jakarta Pusat, Jakarta D.K.I. | Rp 795.000.007 |
| 79 | 60 | 2 | 2 | 2.0 | Cempaka Putih, Jakarta Pusat, Jakarta D.K.I. | Rp 589.300.023 |
| 144 | 41 | 2 | 1 | 1.0 | Cempaka Putih, Jakarta Pusat, Jakarta D.K.I. | Rp 495.000.000 |
| 90 | 60 | 4 | 2 | 1.0 | Kelapa Gading, Jakarta Utara, Jakarta D.K.I. | Rp 2.100.000.000 |

Dataset yang masih mentah memiliki total 2098 baris data. Langkah pertama yang peneliti lakukan adalah melakukan penamaan ulang pada tiap atribut. Setelah itu peneliti juga melakukan perubahan pada tipe data masing masing atribut. Untuk atribut numerikal diseragamkan menjadi tipe *float* agar mempermudah pemrosesan data. Sebagai contoh atribut harga pada awalnya bertipe *string* sehingga harus diubah menjadi tipe *float*. Selain itu peneliti juga melakukan pemisahan pada atribut lokasi menjadi tiga atribut baru yaitu Kecamatan, Kotamadya, dan Provinsi. **Tabel 4** adalah hasil dari dataset yang telah dilakukan proses-proses yang telah disebutkan:

Tabel 4. Dataset hasil praproses

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Luas tanah** | **Luas bangunan** | **Kamar tidur** | **Kamar mandi** | **Lantai** | **Harga** | **Kecamatan** | **Kota** | **Provinsi** |
| 60.0 | 60.0 | 3.0 | 2.0 | 2.0 | 550000000 | Cempaka Putih | Jakarta Pusat | Jakarta D.K.I. |
| 87.0 | 85.0 | 3.0 | 2.0 | 2.0 | 795000007 | Sawah Besar | Jakarta Pusat | Jakarta D.K.I. |
| 79.0 | 60.0 | 2.0 | 2.0 | 2.0 | 589300.023 | Cempaka Putih | Jakarta Pusat | Jakarta D.K.I. |
| 144.0 | 41.0 | 2.0 | 1.0 | 1.0 | 495000000 | Cempaka Putih | Jakarta Pusat | Jakarta D.K.I. |
| 90.0 | 60.0 | 4.0 | 2.0 | 1.0 | 2100000000 | Kelapa Gading | Jakarta Utara | Jakarta D.K.I. |

1. Penanganan *missing values*. Dengan mempertimbangkan jumlah baris data yang sudah cukup banyak, maka semua *missing values* akan langsung dihapus.
2. Penanganan *outlier* pada atribut target yaitu atribut harga menggunakan *univariate tukey boxplot* dan konsep *interquartile range*. Sama seperti penanganan *missing values*, semua data yang tergolong *outlier* akan langsung dihapus.

Setelah dilakukan praproses, jumlah baris data yang pada awalnya berjumlah 2098 baris data, sekarang hanya tinggal bersisa 1548 baris data.

C. Proses *Clustering*

Data yang sudah dilakukan praproses akan dilakukan proses *clustering*. Pada proses ini terbagi menjadi tiga tahap yaitu:

1. Mengidentifikasi jumlah *cluster* optimal menggunakan metode *Bayesian Information Criterion* (BIC) dan *Akaike Information Criterion* (AIC). **Gambar 3** adalah hasil identifikasinya dengan *range* jumlah *cluster* dari 1-10.

Chart, line chart

Description automatically generated

Gambar 3. Grafik identifikasi jumlah *cluster* optimal

Dapat dilihat pada gambar kalau grafik BIC dan AIC yang telah dibuat sama-sama menunjukkan kalau jumlah *cluster* optimal adalah 3 *cluster*. Namun terdapat sedikit perbedaan antara BIC dan AIC. Perbedaannya terletak pada tingkat ketelitian masing-masing metode. Pada grafik AIC yang berwarna oranye, nilai AIC yang dihasilkan untuk jumlah *cluster* diatas 3 tampak konstan hampir tidak ada perbedaan. Sedangkan pada grafik BIC yang berwarna biru, mulai dari jumlah *cluster* sebanyak 3 cluster, grafik menunjukkan tren peningkatan seiring bertambahnya jumlah *cluster*. Hal ini menunjukkan kalau metode BIC lebih teliti daripada metode AIC dalam menentukan jumlah *cluster* optimal.

1. Proses *modelling* GMM menggunakan algoritma EM dengan jumlah *cluster* optimal yang telah didapatkan pada langkah sebelumnya, yaitu sebanyak 3 *cluster*. **Gambar 4** dan **Gambar 5** adalah visualisasi sebaran data dari hasil *clustering* yang telah dibuat.

Chart, pie chart

Description automatically generated

Gambar 4. *Pie chart* sebaran data hasil *clustering*

Chart, scatter chart

Description automatically generated

Gambar 5. *Scatter plot* sebaran data hasil *clustering*

*Cluster* 1 memiliki 785 anggota, *cluster* 2 memiliki 529 anggota, dan *cluster* 3 memiliki 234 anggota. Jika dilihat dari bentuk sebaran datanya pada **Gambar 5**, *cluster* 1 dan *cluster* 2 memiliki bentuk yang cenderung seperti elips sedangkan *cluster* 3 cenderung berbentuk seperti lingkaran. Hal ini membuktikan kalau GMM-*based clustering* dapat menghasilkan bentuk cluster yang lebih fleksibel. Rincian tiga parameter *Gaussian Mixture Model*nyaadalah sebagai berikut. **Tabel 5** mendeskripsikan parameter GMM optimal dari masing-masing *cluster*.

Tabel 5. Parameter GMM untuk tiap *cluster*

|  |  |  |  |
| --- | --- | --- | --- |
| **Parameter** | ***Cluster* 1** | ***Cluster* 2** | ***Cluster* 3** |
| **Bobot** | 0.47367921 | 0.33842949 | 0.1878913 |
| **Standar deviasi** |  |  |  |
| **Mean** |  |  |  |

*Clustering* dengan 3 *cluster* ini menghasilkan skor *silhouette* yang cukup besar dan bagus yaitu sebesar 0,60866. Untuk evaluasi lebih lanjut akan dibahas pada bab berikutnya.

1. **Hasil dan Evaluasi**

Setiap atribut akan dianalisis untuk mengetahui informasi apa saja yang bisa didapatkan dari hasil *clustering*. Untuk atribut numerikal hasilnya akan ditampilkan menggunakan histogram, sedangkan untuk atribut kategorikal atributnya akan ditampilkan menggunakan diagram batang. **Gambar 6** sampai **Gambar 13** merupakan hasil visualisasi untuk setiap atribut.

Graphical user interface, chart, histogram

Description automatically generated

Gambar 6. Histogram atribut harga

Chart, histogram

Description automatically generated

Gambar 7. Histogram atribut luas tanah

Chart, histogram

Description automatically generated

Gambar 8. Histogram atribut luas bangunan

Chart

Description automatically generated

Gambar 9. Histogram atribut kamar tidur

Chart

Description automatically generated

Gambar 10. Histogram atribut kamar mandi

Chart, waterfall chart

Description automatically generated

Gambar 11. Histogram atribut lantai

Chart, bar chart

Description automatically generated

Gambar 12. Diagram batang atribut kotamadya

Chart, bar chart

Description automatically generated

Gambar 13. Diagram batang atribut kecamatan

Setelah ditampilkan hasilnya, yang selanjutnya dilakukan adalah menganalisis hasil dari histogram dan diagram batang yang telah ditampilkan. Berikut adalah hasil analisis per atributnya:

1. Harga

*Cluster* 1 merupakan *cluster* dengan harga paling rendah yaitu berada pada rentang Rp 150.000.000 - Rp 982.300.000. Namun jika dilihat dari persebaran datanya melalui histogram, rumah pada *cluster* 1 paling banyak berada pada rentang harga Rp 450.000.000 - Rp 650.000.000.

*Cluster* 2 merupakan *cluster* dengan harga menengah yaitu berada pada rentang Rp 990.000.000 - Rp 2.325.000.000. Namun jika dilihat dari persebaran datanya melalui histogram, rumah pada *cluster* 2 paling banyak berada pada rentang harga Rp 990.000.000 - Rp 1.800.000.000.

*Cluster* 3 merupakan *cluster* dengan harga paling tinggi yaitu berada pada rentang Rp 2.350.000.000 - Rp 6.500.000.000. Namun jika dilihat dari persebaran datanya melalui histogram, rumah pada *cluster* 3 paling banyak berada pada rentang harga Rp 2.350.000.000 - Rp 4.000.000.000.

Jika dilihat dari harganya, dapat disimpulkan kalau *cluster* 1 merupakan *cluster* dengan harga rumah yang terjangkau, *cluster* 2 merupakan *cluster* dengan harga rumah kelas menengah, dan *cluster* 3 merupakan *cluster* dengan harga rumah yang tinggi.

1. Luas tanah

*Cluster* 1 merupakan *cluster* dengan luas tanah terkecil dengan rentang luas tanah 21 m2 - 151 m2. Namun jika dilihat sebaran datanya melalui histogram, rumah pada *cluster* 1 paling banyak berada pada rentang luas tanah 45 m2 - 75 m2.

*Cluster* 2 merupakan *cluster* dengan luas tanah menengah dengan rentang luas tanah 31 m2 - 200 m2. Namun jika dilihat dari sebaran datanya melalui histogram, rumah pada *cluster* 2 paling banyak berada pada rentang luas tanah 60 m2 - 100 m2.

*Cluster* 3 merupakan *cluster* dengan luas tanah tertinggi dengan rentang luas tanah 48 m2 - 203 m2. Namun jika dilihat dari sebaran datanya melalui histogram, rumah pada *cluster* 3 paling banyak berada pada rentang luas tanah 80 m2 - 140 m2.

1. Luas bangunan

*Cluster* 1 merupakan *cluster* dengan luas bangunan terkecil dengan rentang luas bangunan 21 m2 - 172 m2. Namun jika dilihat sebaran datanya melalui histogram, rumah pada *cluster* 1 paling banyak berada pada rentang luas tanah 21 m2 - 50 m2.

*Cluster* 2 merupakan *cluster* dengan luas tanah menengah dengan rentang luas tanah 31 m2 - 262 m2. Namun jika dilihat dari sebaran datanya melalui histogram, rumah pada *cluster* 2 paling banyak berada pada rentang luas tanah 75 m2 - 125 m2.

*Cluster* 3 merupakan *cluster* dengan luas tanah tertinggi dengan rentang luas tanah 72 m2 - 265 m2. Namun jika dilihat dari sebaran datanya melalui histogram, rumah pada *cluster* 3 paling banyak berada pada rentang luas tanah 110 m2 - 225 m2.

1. Kamar tidur

Mayoritas rumah pada *cluster* 1, yaitu sekitar 700 rumah, memiliki kamar tidur dengan rentang 2-4 kamar tidur. Sedangkan sangat sedikit rumah pada *cluster* 1 yang memiliki dibawah 2 kamar tidur dan diatas 4 kamar tidur.

Mayoritas rumah pada *cluster* 2, yaitu sekitar 400 rumah, memiliki kamar tidur dengan rentang 2-4 kamar tidur. Lalu sekitar 100 rumah memiliki kamar tidur dengan rentang 4-6 kamar tidur. Sedangkan sangat sedikit rumah pada *cluster* 2 yang memiliki dibawah 2 kamar tidur dan diatas 6 kamar tidur.

Mayoritas rumah pada pada *cluster* 3, yaitu sekitar 160 rumah, memiliki kamar tidur dengan rentang 4-6 kamar tidur. Sekitar 60 rumah memiliki kamar tidur dengan rentang 2-4 kamar tidur. Sedangkan sangat sedikit rumah pada *cluster* 3 yang memiliki dibawah 2 kamar tidur dan diatas 6 kamar tidur.

1. Kamar mandi

Mayoritas rumah pada *cluster* 1, yaitu sekitar 300-400 rumah memiliki kamar mandi dengan rentang 1-4 kamar mandi. Terdapat beberapa rumah pada *cluster* 1 yang memiliki kamar mandi dengan rentang 4-6 kamar mandi namun sangat sedikit. Sedangkan tidak ada rumah pada *cluster* 1 yang memiliki diatas 6 kamar mandi.

Mayoritas rumah pada *cluster* 2, yaitu lebih dari 400 rumah, memiliki kamar mandi dengan rentang 2-4 kamar mandi. Terdapat beberapa rumah pada *cluster* 2 yang memiliki 1-2 kamar mandi dan 4-6 kamar mandi, namun sangat sedikit. Sedangkan tidak ada rumah pada *cluster* 2 yang memiliki diatas 6 kamar mandi.

Mayoritas rumah pada *cluster* 3, yaitu sekitar 140 rumah, memiliki kamar mandi dengan rentang 2-4 kamar mandi. Lalu sekitar 80 rumah memiliki kamar mandi dengan rentang 4-6 kamar mandi. Namun sangat sedikit rumah pada *cluster* 3 yang memiliki dibawah 2 kamar mandi dan diatas 6 kamar mandi.

1. Lantai

Mayoritas rumah pada *cluster* 1 memiliki lantai dengan rentang 1-4 lantai. Sedangkan sangat sedikit rumah yang memiliki lantai diluar rentang 1-4 lantai.

Mayoritas rumah pada *cluster* 2 memiliki lantai dengan rentang 2-3 lantai. Beberapa rumah memiliki rentang 1-2 lantai dan 3-4 lantai. Sedangkan sangat sedikit rumah pada *cluster* 2 yang memiliki diatas 4 lantai.

Rumah pada *cluster* 3 memiliki ciri yang cukup serupa dengan *cluster* 2 jika dilihat dari sebaran atribut lantai.

1. Lokasi

Rumah pada *cluster* 1 paling banyak berlokasi di Jakarta Selatan dan Jakarta Pusat sedangkan paling sedikit berlokasi di Jakarta Utara. Jika dilihat dari kecamatannya, rumah pada *cluster*1 paling banyak berlokasi di Kecamatan Cempaka Putih, Jakarta Pusat.

Rumah pada *cluster* 2 paling banyak berlokasi di Jakarta Selatan dan Jakarta Timur sedangkan paling sedikit berlokasi di Jakarta Utara. Jika dilihat dari kecamatannya, rumah pada *cluster* 2 paling banyak berlokasi di Kecamatan Duren Sawit, Jakarta Timur.

Rumah pada *cluster* 3 paling banyak berlokasi di Jakarta Selatan dan Jakarta Barat sedangkan paling sedikit berlokasi di Jakarta Pusat. Namun jika dilihat dari Kecamatannya, rumah pada *cluster*3 paling banyak berlokasi di Kecamatan Kelapa Gading, Jakarta Utara.

Jika dilihat pola persebarannya berdasarkan lokasi, Jakarta Utara merupakan Kawasan elit dimana harga rumah sangat tinggi yaitu paling banyak berada di *cluster* 3 sedangkan sangat sedikit rumah di Jakarta Utara yang berada di *cluster* 1 dan *cluster* 2. Jakarta Pusat merupakan lokasi dengan harga rumah yang cukup terjangkau karena rumah di Jakarta Pusat paling banyak berada di *cluster* 1 sedangkan sangat sedikit rumah di Jakarta Pusat yang berada di *cluster* 3. Jakarta Selatan merupakan lokasi yang memiliki rumah dengan harga sangat beragam yang karena tersebar merata pada *cluster* 1, 2, dan 3.

1. **Kesimpulan dan Saran**

Pada proses *clustering* dengan GMM, parameter GMM pada awalnya akan diinisialisasikan secara acak lalu dilakukan optimasi menggunakan algoritma EM yang memanfaatkan fungsi *log-likelihood* yang prosesnya akan berulang sampai nilai *log­-likelihood* yang diperoleh mencapai konvergensi.

Pada penelitian ini, jumlah *mixture model* atau *cluster* yang diperoleh adalah sebanyak 3 *cluster*. *Cluster* 1 merupakan *cluster* dengan profil rumah kelas rendah yang mayoritas berada dikisaran harga Rp 450.000.000 – Rp 650.000.000 dan paling banyak berlokasi di kecamatan Cempaka Putih, Jakarta Pusat. *Cluster* 2 merupakan *cluster* dengan profil rumah kelas menengah yang mayoritas berada dikisaran harga Rp 990.000.000 – Rp 1.800.000.000 dan paling banyak berlokasi di Kecamatan Duren Sawit, Jakarta Timur. Sedangkan *cluster* 3 merupakan *cluster* dengan profil rumah kelas atas yang mayoritas berada dikisaran harga Rp 2.350.000.000 – Rp 4.000.000.000 dan paling banyak berlokasi di kecamatan Kelapa Gading, Jakarta Utara.

Adapun saran yang dapat diberikan peneliti untuk para calon pembeli rumah di Provinsi DKI Jakarta yaitu jika calon pembeli memiliki *budget* yang tergolong rendah maka disarankan untuk mencari rumah di daerah Jakarta Pusat. Jika calon pembeli memiliki *budget* yang tinggi maka disarankan untuk mencari rumah di daerah Jakarta Utara. Namun jika calon pembeli masih belum menetapkan berapa *budget* yang dimiliki dan masih belum melakukan survey yang mendalam, disarankan untuk mencari rumah di daerah Jakarta Selatan karena rumah-rumah di daerah Jakarta Selatan memikiki harga yang sangat beragam.

**Daftar Pustaka**

[1] M. Yazdani, “House Price Determinants and Market Segmentation in Boulder, Colorado: A Hedonic Price Approach,” *arXiv Prepr. arXiv2108.02442*, 2021.

[2] T. S. Madhulatha, “An overview on clustering methods,” *arXiv Prepr. arXiv1205.1117*, 2012.

[3] M.-S. Yang, C.-Y. Lai, and C.-Y. Lin, “A robust EM clustering algorithm for Gaussian mixture models,” *Pattern Recognit.*, vol. 45, no. 11, pp. 3950–3961, 2012.

[4] H. Ling and K. Zhu, “Predicting Precipitation Events Using Gaussian Mixture Model,” *J. Data Anal. Inf. Process.*, vol. 5, no. 04, p. 131, 2017.

[5] K. Li, Z. Ma, D. Robinson, and J. Ma, “Identification of typical building daily electricity usage profiles using Gaussian mixture model-based clustering and hierarchical clustering,” *Appl. Energy*, vol. 231, pp. 331–342, 2018.

[6] A. Saxena *et al.*, “A review of clustering techniques and developments,” *Neurocomputing*, vol. 267, pp. 664–681, 2017.

[7] G. Shevlyakov, K. Andrea, L. Choudur, P. Smirnov, A. Ulanov, and N. Vassilieva, “Robust versions of the Tukey boxplot with their application to detection of outliers,” in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 6506–6510.

[8] H. P. Vinutha, B. Poornima, and B. M. Sagar, “Detection of outliers using interquartile range technique from intrusion dataset,” in *Information and decision sciences*, Springer, 2018, pp. 511–518.

[9] D. A. Reynolds, “Gaussian mixture models.,” *Encycl. biometrics*, vol. 741, no. 659–663, 2009.

[10] X. He, D. Cai, Y. Shao, H. Bao, and J. Han, “Laplacian regularized gaussian mixture model for data clustering,” *IEEE Trans. Knowl. Data Eng.*, vol. 23, no. 9, pp. 1406–1418, 2010.

[11] N. Sammaknejad, Y. Zhao, and B. Huang, “A review of the expectation maximization algorithm in data-driven process identification,” *J. Process Control*, vol. 73, pp. 123–136, 2019.

[12] S. F. Qonita, “Segmentasi Citra MRI Tumor Otak Menggunakan Gaussian Mixture Model dan Hybrid Gaussian Mixture Model-Spatially Variant Finite Mixture Model dengan Algoritma Expectation-Maximization.” Institut Teknologi Sepuluh Nopember, 2018.

[13] J. Deng, Y. Deng, and K. H. Cheong, “Combining conflicting evidence based on Pearson correlation coefficient and weighted graph,” *Int. J. Intell. Syst.*, vol. 36, no. 12, pp. 7443–7460, 2021.

[14] S. Watanabe, “A widely applicable Bayesian information criterion,” *arXiv Prepr. arXiv1208.6338*, 2012.

[15] J. E. Cavanaugh and A. A. Neath, “The Akaike information criterion: Background, derivation, properties, application, interpretation, and refinements,” *Wiley Interdiscip. Rev. Comput. Stat.*, vol. 11, no. 3, p. e1460, 2019.

[16] A. Aditya, I. Jovian, and B. N. Sari, “Implementasi K-Means Clustering Ujian Nasional Sekolah Menengah Pertama di Indonesia Tahun 2018/2019,” *J. Media Inform. Budidarma*, vol. 4, no. 1, pp. 51–58, 2020.

[17] P. J. Rousseeuw, “Silhouettes: a graphical aid to the interpretation and validation of cluster analysis,” *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, 1987.

[18] S. A. Alasadi and W. S. Bhaya, “Review of data preprocessing techniques in data mining,” *J. Eng. Appl. Sci.*, vol. 12, no. 16, pp. 4102–4107, 2017.

Lampiran