**Segmentasi Harga Jual Rumah Menggunakan Gaussian Mixture Model-Based Clustering**

**Proposal Tugas Akhir**

**Kelas MK Penulisan Proposal (CCH4A3)**

**1301184079**

**Muhammad Hafidh Raditya**

****

**Program Studi Sarjana Informatika**

**Fakultas Informatika**

**Universitas Telkom**

**Bandung**

**< 2021>**

# **Lembar Persetujuan**

**Segmentasi Harga Jual Rumah Menggunakan Gaussian Mixture Model-Based Clustering**

**House Price Segmentation Using Gaussian Mixture Model-Based Clustering**

**NIM :1301184079**

**Muhammad Hafidh Raditya**

Proposal ini diajukan sebagai usulan pembuatan tugas akhir pada

Program Studi Sarjana Informatika

Fakultas Informatika Universitas Telkom

Bandung,<16/12/2021>

Menyetujui

|  |  |  |
| --- | --- | --- |
| Calon Pembimbing 1 |  | Calon Pembimbing 2 |
|  |  |  |
|  |  |  |
| <Dra. Indwiarti, M.Si> <98690022 > |  | <Nama Lengkap dengan Gelar> <NIP> |

# 

# **ABSTRAK**

Rumah merupakan tempat tinggal bagi manusia. Rumah juga merupakan salah satu kebutuhan pokok yang dimiliki 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 sehingga penelitian ini dapat menghasilkan informasi yang berguna bagi calon pembeli rumah. Salah satu cara melakukan proses segmentasi adalah menggunakan Teknik *clustering*. Teknik *clustering* pada penelitian ini akan menggunakan metode *Gaussian Mixture Model* (GMM) karena teknik *clustering* menggunakan metode GMM dapat menghasilkan bentuk *cluster* yang lebih fleksibel.

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

# **DAFTAR ISI**

[**Lembar Persetujuan** 1](#_Toc93922360)

[**ABSTRAK** i](#_Toc93922361)

[**DAFTAR ISI** ii](#_Toc93922362)

[**1.** **PENDAHULUAN** 1](#_Toc93922363)

[**1.1.** **Latar Belakang** 1](#_Toc93922364)

[**1.2.** **Perumusan Masalah** 2](#_Toc93922365)

[**1.3.** **Tujuan** 2](#_Toc93922366)

[**1.4.** **Rencana Kegiatan** 2](#_Toc93922367)

[**1.5.** **Jadwal Kegiatan** 3](#_Toc93922368)

[**2.** **KAJIAN PUSTAKA** 4](#_Toc93922369)

[**2.1.** ***Clustering*** 4](#_Toc93922370)

[**2.2.** ***Univariate Tukey Boxplot*** 4](#_Toc93922371)

[**2.3.** **Deteksi Data Pencilan Menggunakan *Interquartile Range*** 4](#_Toc93922372)

[**2.4.** ***Gaussian Mixture Model*** 5](#_Toc93922373)

[**2.5.** ***Expectation-Maximization*** 5](#_Toc93922374)

[**2.6.** **Koefisien Korelasi Pearson** 6](#_Toc93922375)

[**2.7.** ***Bayesian Information Criterion*** 7](#_Toc93922376)

[**2.8.** ***Silhouette*** 7](#_Toc93922377)

[**3.** **Metodologi Penelitian** 8](#_Toc93922378)

[**3.1.** **Identifikasi Masalah** 8](#_Toc93922379)

[**3.2.** **Studi Literatur** 8](#_Toc93922380)

[**3.3.** **Pengumpulan Data** 9](#_Toc93922381)

[**3.4.** **Eksplorasi dan Praproses Data** 9](#_Toc93922382)

[**3.5.** **Proses Clustering** 10](#_Toc93922383)

[**3.6.** **Eksplorasi Data Hasil Clustering** 10](#_Toc93922384)

[**3.7.** **Kesimpulan dan Saran** 10](#_Toc93922385)

[**DAFTAR PUSTAKA** 11](#_Toc93922386)

[**LAMPIRAN** 12](#_Toc93922387)

# 

# **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 statistic 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*. Berikut adalah tabel perbandingan antara GMM dan *hierarchical clustering* menggunakan metode *silhouette*.

Tabel 1.1 Perbandingan GMM dan *hierarchical clustering* menggunakan *silhouette*

|  |  |  |
| --- | --- | --- |
| **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*.

## **Perumusan Masalah**

Berikut adalah perumusan masalah yang akan diangkat pada penelitian ini:

1. Bagaimana cara optimalisasi parameter dari GMM?
2. Bagaimana hasil segmentasi harga rumah menggunakan *Gaussian Mixture Model-Based Clustering* dan informasi apa saja yang bisa didapat?

## **Tujuan**

Berikut adalah tujuan yang ingin dicapai dalam penulisan proposal:

1. Mengidentifikasi cara melakukan optimalisasi parameter dari GMM
2. Mengidentifikasi hasil segmentasi harga jual rumah menggunakan GMM-Based Clustering dan menganalisis hasil yang didapat.

## **Rencana Kegiatan**

Rencana kegiatan yang akan dilakukan adalah sebagai berikut:

1. Studi literatur
2. Identifikasi masalah
3. Persiapan dan praproses data
4. Proses segmentasi
5. Analisis hasil proses segmentasi
6. Publikasi dan pelaporan

## **Jadwal Kegiatan**

Laporan proposal tugas akhir ini akan dijadwalkan sesuai dengan tabel berikut:

Tabel 1.2 Jadwal kegiatan proposal tugas akhir

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Kegiatan** | **Bulan** | | | | | |
| 1 | 2 | 3 | 4 | 5 | 6 |
| Studi literatur |  |  |  |  |  |  |
| Identifikasi masalah |  |  |  |  |  |  |
| Persiapan dan praproses data |  |  |  |  |  |  |
| Proses segmentasi |  |  |  |  |  |  |
| Analisis hasil proses segmentasi |  |  |  |  |  |  |
| Publikasi |  |  |  |  |  |  |

# **KAJIAN PUSTAKA**

## ***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].

## ***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 berikut adalah contoh dari penggunaan *univariate Tukey boxplot*.

Chart

Description automatically generated

Gambar 2. 1 *Univariate Tukey Boxplot*

## **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; . 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. *lower bound* dan *upper bound* memiliki persamaan sebagai berikut:

## ***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:

Dimana merupakan banyaknya *cluster*.

Yang jadi permasalahan utama pada GMM adalah bagaimana menentukan nilai optimal dari ketiga parameter tersebut. Untuk itu diperlukan algoritma *Expectation-Maximization* (EM) dalam upaya optimalisasi ketiga parameter tersebut.

## ***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 berat () secara acak lalu evaluasi fungsi ­*log-likelihood* menggunakan ketiga parameter tersebut yang memenuhi persamaan [12]:

Dimana adalah banyaknya data dan adalah banyaknya *cluster*.

1. E-*step* (*Expectation)*

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

1. M-*step* (­*Maximization*)

Tahap ini melakukan pembaruan pada ketiga parameter tersebut dengan menggunakan total nilai probabilitas () pada tiap iterasinya. Persamaan yang dipakai adalah:

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

## **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 [13]:

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

1. Range nilai dari adalah .
2. Jika , dan berkorelasi positif.
3. Jika , tidak terdapat korelasi linier antara dan
4. Jika , dan berkorelasi negatif.
5. Semakin besar nilai maka tingkat korelasi linier antara dan juga semakin besar.

## ***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. BIC memiliki persamaan:

Dimana adalah nilai log­-*likelihood* dari model yang dimiliki, adalah banyaknya data, dan adalah banyaknya *cluster*.

## ***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 memiliki persamaan [15]:

Dimana adalah banyaknya data, adalah banyaknya *cluster* dan adalah nilai dari *sum square error*.

## ***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 yang dipakai pada metode *silhouette* adalah sebagai berikut [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

# **Metodologi Penelitian**

Metodologi penelitian merupakan tahapan-tahapan dalam penyelesaian masalah pada penelitian ini. Gambar berikut merupakan akan menjelaskan tahapan-tahapan dalam penelitian ini dalam bentuk diagram alir.

Diagram

Description automatically generated

Gambar 3. 1 Diagram alir metodologi penelitian

## **Identifikasi Masalah**

Pada tahap ini dilakukan proses identifikasi dan pemahaman terhadap masalah yang dihadapi. Proses identifikasi masalah berdasar dari tujuan dari penelitian ini. Berdasarkan tujuan penelitian ini, maka dapat ditentukan kalau masalah yang akan dipecahkan adalah bagaimana hasil dari segmentasi harga rumah jika menggunakan metode *gaussian mixture model-based clustering*.

## **Studi Literatur**

Studi literatur merupakan tahapan yang dilakukan untuk mengumpulkan dan mengkaji Pustaka yang berkaitan dengan proses penelitian dan penyelesaian masalah. Konsep dan metode yang digunakan adalah *clustering*, *Univariate Tukey Boxplot,* Deteksi data pencilan menggunakan *Interquartile* Range, Gaussian *Mixture Model, Expectation-Maximization,* Koefisien Korelasi Pearson, *Bayesian Information Criterion*, dan *silhouette*.

## **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.

Dataset ini memiliki beberapa atribut antara lain:

Tabel 3. 1 Beberapa atribut yang terdapat dalam 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

## **Eksplorasi dan Praproses Data**

Pada tahap ini akan dilakukan proses eksplorasi data, yaitu proses yang dilakukan untuk memahami data sebelum dilakukan praproses data. Praproses data adalah proses pengelohan data mentah menjadi data yang siap pakai untuk dilakukan proses *modelling.* Tahapan yang dilakukan pada praproses data antara lain:

1. Berikut adalah bentuk dari dataset yang masih 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 diubah menjadi tipe *float* agar mempermudah pemrosesan data. Sebagai contoh atribut harga pada awalnya bertipe *string* sehingga harus diubah menjadi tipe *float*. Selain itu peneliti juka melakukan pemisahan pada atribut lokasi menjadi tiga atribut baru yaitu Kecamatan, Kotamadya, dan Provinsi. Berikut adalah hasil dari dataset yang telah dilakukan proses-proses yang telah disebutkan:



1. Penanganan *missing values*. Pada tahap ini semua *missing values* akan langsung dihapus baris datanya.
2. Penanganan *outlier* pada atribut target yaitu atribut harga menggunakan *univariate tukey boxplot* dan konsep *interquartile range*.
3. Melakukan pengecekan atribut numerikal apa yang paling berpengaruh terhadap atribut harga menggunakan koefisien korelasi Pearson. Berikut adalah hasilnya:



Setiap atribut berkorelasi positif terhadap atribut harga. Atribut luas\_bangunan merupakan atribut yang paling berpengaruh, diikuti oleh luas\_tanah. Sedangkan lantai merupakan atribut yang pengaruhnya paling kecil diantara atribut lain.

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

## **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). Berikut adalah hasil identifikasinya dengan jumlah *cluster* dari 1-10.

Chart, line chart

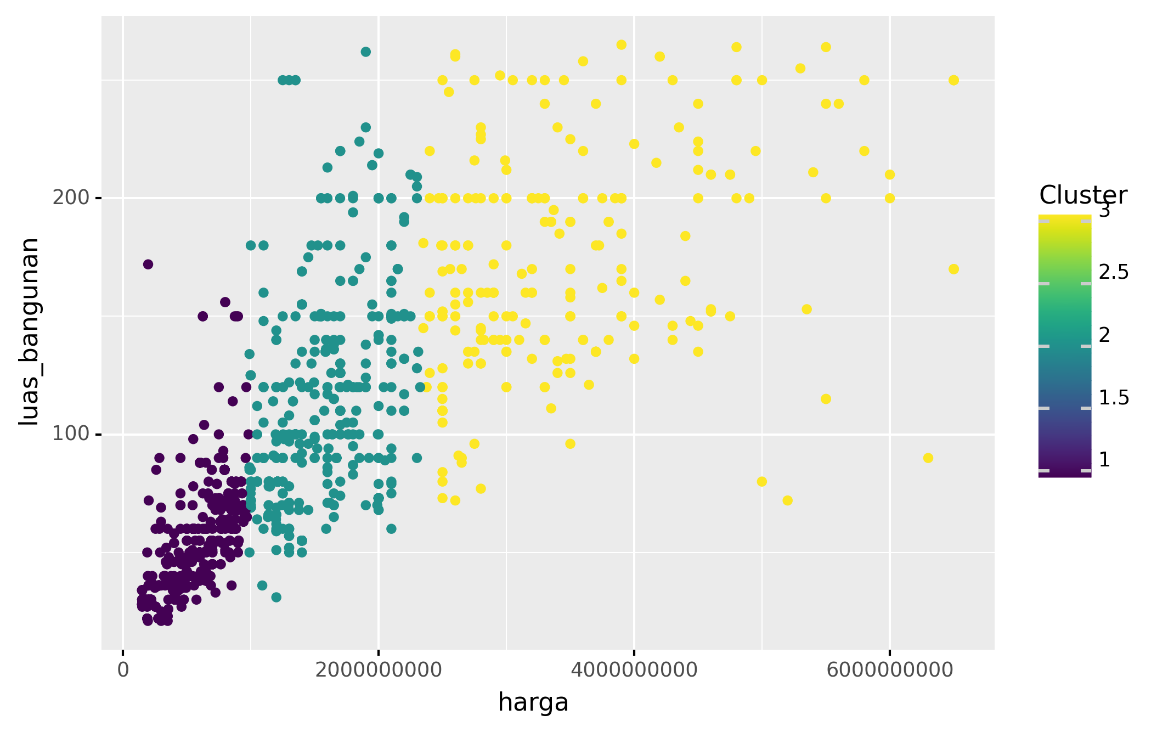
Description automatically generated

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, 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*. Berikut adalah hasilnya.

Chart, pie chart

Description automatically generated



*Cluster* 1 memiliki 785 anggota, *cluster* 2 memiliki 529 anggota, dan *cluster* 3 memiliki 234 anggota. Rincian tiga parameter *Gaussian Mixture Model*nyaadalah sebagai berikut.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **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.

# **Hasil dan Evaluasi**

## **Hasil**

Graphical user interface, chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

Chart

Description automatically generated

Chart

Description automatically generated

Chart, waterfall chart

Description automatically generated

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

## **Evaluasi**

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 tersebar merata pada *cluster* 1, 2, dan 3.

# **KESIMPULAN**

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 konvergen.

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.

# **DAFTAR PUSTAKA**

|  |  |
| --- | --- |
| [1] | M. Yazdani, "House Price Determinants and Market Segmentation in Boulder, Colorado: A Hedonic Price Approach," *General Economics,* p. 45, 2021. |
| [2] | T. S. Madhulatha, "An Overview On Clustering Methods," *IOSR Journal of Engineering,* vol. 2, no. 4, p. 7, 2012. |
| [3] | Y. Miin-Shen, "A robust EM clustering algorithm for Gaussian mixture models," *Pattern Recognition,* vol. 45, no. 11, p. 12, 2012. |
| [4] | L. Haitian and Z. Kunping, "Predicting Precipitation Events Using Gaussian Mixture Model," *Journal of Data Analysis and Information Processing,* vol. 5, no. 4, p. 9, 2017. |
| [5] | L. Kehua, M. Zhenjun, R. Duane and M. Jun, "Identification of typical building daily electricity usage profiles using Gaussian mixture model-based clustering and hierarchical clustering," *Applied Energy,* vol. 231, p. 12, 2018. |
| [6] | S. Amit, P. Mukesh, G. Akshansh, B. Neha, P. Prakash, T. Aruna, E. Meng Joo, W. Ding and L. Chin-Teng, "A review of clustering techniques and developments," *Neurocomputing,* vol. 267, p. 34, 2017. |
| [7] | G. Shevlyakov, A. Klinton, L. Choudur, P. Smirnov, A. Ulanov and N. Vassilieva, "Robust versions of the Tukey boxplot with their application to detection of outliers," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013. |
| [8] | H. P. Vinutha, B. Poornima and B. M. Sagar, "Detection of Outliers Using Interquartile Range Technique from Intrusion Dataset," *Information and Decision Sciences,* vol. 701, pp. 511-518, 2018. |
| [9] | D. Reynolds, "Gaussian Mixture Models," in *Encyclopedia of Biometrics*, 2009, pp. 659-663. |
| [10] | H. Xiaofei, IEEE, C. Deng, S. Yuanlong, B. Hujun and H. Jiawei, "Laplacian Regularized Gaussian Mixture Model," *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING,* vol. 23, no. 9, p. 13, 2011. |
| [11] | S. Nima, Z. Yujia and H. Biao, "A review of the Expectation Maximization algorithm in data-driven," *Journal of Process Control,* vol. 73, p. 14, 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," p. 120, 2018. |
| [13] | D. Jixiang, D. Yong and C. Kang Hao, "Combining Conflicting Evidence Based on Pearson Correlation Coefficient and Weighted Graph," *International Journal of Intelligent Systems,* vol. 36, no. 12, p. 18, 2021. |
| [14] | S. Watanabe, "A Widely Applicable Bayesian Information Criterion," *Journal of Machine Learning Research,* vol. 14, p. 31, 2014. |
| [15] | J. E. Cavanaugh, "The Akaike information criterion: Background, derivation, properties, application, interpretation, and refinements," *WIREs Computational Statistics,* vol. 11, no. 3, 2019. |
| [16] | A. Agil, J. Ivan and S. Betha Nurina, "Implementasi K-Means Clustering Ujian Nasional Sekolah Menengah," *JURNAL MEDIA INFORMATIKA BUDIDARMA,* vol. 4, no. 1, p. 8, 2020. |
| [17] | J. Peter Rousseuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics,* vol. 20, p. 13, 1987. |

# **LAMPIRAN**