

Clustering Harga Rumah: Perbandingan Model K-Means dan Gaussian Mixture Model

1st Rizky Rahmattullah

Fakultas Informatika

Universitas Telkom

Bandung, Indonesia

rrahmattullah@student.telkomuni
ty.ac.id

2nd Indwiarti

Fakultas Informatika

Universitas Telkom

Bandung, Indonesia

indwiarti@telkomuniversity.ac.id

3rd Aniq Atiqi Rohmawati

Fakultas Informatika

Universitas Telkom

Bandung, Indonesia

aniqatiqi@telkomuniversity.ac.id

Abstrak—Rumah merupakan kebutuhan primer manusia sebagai tempat bernaung, berlindung, dan beristirahat. Sebagai kebutuhan primer, seluruh manusia berhak untuk mencari tempat tinggalnya masing-masing dengan keluarganya. Seiring berjalannya waktu, kebutuhan akan tempat tinggal semakin meningkat dan mempengaruhi harga jual rumah. Maka dilakukan clustering mengenai harga rumah dengan menggunakan metode *K-Means* dan *Gaussian Mixture Model*. Pada penelitian ini menggunakan data hargarumah di wilayah Kabupaten Bogor yang dihimpun dari website olx.co.id. *Silhouette Score* digunakan sebagai pembandingan dari dua metode *Clustering* yang digunakan. Hasil dari penelitian ini, *K-Means* memiliki *Silhouette Score* sebesar 0.63516 lebih besar dari *Gaussian Mixture Model* yang memiliki *Silhouette Score* sebesar 0.62723 menjadikan kualitas cluster pada *K-Means* lebih baik daripada *Gaussian Mixture Model* pada penelitian ini.

Kata kunci—rumah, *clustering*, *gaussian mixture model*, *K-Means*

Abstract—The house is a primary human need as a place of shelter, and rest. As a primary need, all humans have the right to find their own place to live with their families. Over time, the need for housing increases and affects the selling price of the house. Then the clustering of house prices is carried out using the *K-Means* method and the *Gaussian Mixture Model*. In this study, data on house prices in the Bogor Regency area were collected from the olx.co.id website. *Silhouette Score* is used as a comparison of the two *Clustering* methods used. The results of this study, *K-Means* has a *Silhouette Score* of 0.63516 which is greater than the *Gaussian Mixture Model* which has a *Silhouette Score* of 0.62723 making the cluster quality in *K-Means* better than the *Gaussian Mixture Model* in this study.

Keywords—house, *clustering*, *gaussian mixture model*, *K-Means*

I. PENDAHULUAN

A. Latar Belakang

Rumah merupakan salah satu bentuk investasi yang menarik. Permintaan akan tempat tinggal di Kabupaten Bogor masih terus meningkat akibat pertumbuhan penduduk akibat urbanisasi dan migrasi dari luar kota. 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 Kabupaten Bogor, dan mengakibatkan harga rumah menjadi fluktuatif.

Clustering merupakan proses *machine learning* yang berfungsi untuk mengelompokkan sekumpulan data yang memiliki kesamaan karakteristik menjadi satu cluster. Penelitian ini bertujuan untuk menguji performa dua model, yaitu *clustering* dengan menggunakan *K-Means* dan *Gaussian Mixture Model* dan akan dibandingkan hasil proses dari kedua metode tersebut. Pada penelitian sebelumnya, ada penelitian yang menggunakan *Gaussian Mixture Model* untuk mengidentifikasi kepadatan kendaraan di jalan tol, dan menghasilkan tingkat akurasi sebesar 91%[1]. Lalu ada identifikasi area kanker ovarium pada citra CT Scan, yang memperoleh hasil tidak cukup bagus dengan persentasi *true positive* sebesar 45% lebih kecil dari *false positive* sebesar 55%[2]. Hasil dari penelitian tersebut memberi informasi tentang prediksi jumlah total kasus COVID-19 di seluruh dunia dan prediksi tanggal berakhirnya pandemic COVID-19. Lalu terdapat penelitian yang melakukan perbandingan *clustering cloud workloads* dengan menggunakan dua metode yaitu *K-Means* dan *Gaussian Mixture Model*[3]. Dari penelitian tersebut menghasilkan *Gaussian Mixture Model* memberikan cluster yang lebih baik, dan lebih rinci daripada *K-Means*, namun dibutuhkan waktu yang lebih lama dibanding proses *clustering* pada *K-Means*.

B. Topik dan Batasannya

1. Pengujian kualitas dua metode *clustering* yaitu *K-Means* dan *Gaussian Mixture Model*
2. Data rumah yang dipakai pada penelitian ini adalah data rumah di Kabupaten Bogor
3. Atribut yang digunakan untuk penelitian adalah harga rumah, karena tujuan penelitian ini hanya menguji performansi dari dua model, bukan untuk membuat informasi data rumah secara detail.
4. Untuk membandingkan kualitas dari hasil *clustering K-Means* dan *Gaussian Mixture Model* dilakukan penghitungan *Silhouette Score*.

C. Tujuan

1. Membuat *clustering* harga rumah menggunakan *K-Means* dan *Gaussian Mixture Model*
2. Menguji performansi dan kualitas *cluster* dari dua metode *clustering*.

II. KAJIAN TEORI

A. Clustering

Clustering merupakan proses *machine learning* yang mengelompokkan sekumpulan data ke dalam kelompok-kelompok berdasarkan kemiripan karakteristik antar data satu dengan yang lain. Metode *clustering* bertujuan untuk mengelompokkan data yang belum memiliki kelas ke dalam kelompok, dan mengambil informasi dari setiap data untuk selanjutnya dianalisis[4].

$$d_{sq} = \sum_{i=1}^D (x_i - y_i)^2 \quad (1)$$

Keterangan:

x,y = titik data

D = ruang dimensi

B. Elbow Method

Elbow Method adalah metode yang digunakan untuk menentukan jumlah cluster yang optimal, yaitu dengan menghitung SSE dari tiap *cluster*. SSE diperoleh dari perhitungan jarak setiap data dengan pusat

$$SSE = \sum_{i=1}^n \sum_{j=1}^k w_{ij} \|x_i - c_j\|^2 \quad (2)$$

Keterangan:

x = data

c = centroid atau pusat data = *cluster*

w = indikator

w akan bernilai 1 apabila x anggota suatu *cluster*, dan w akan bernilai 0 apabila x bukan anggota suatu *cluster*.

C. Gaussian Mixture Model

$$p(x) = \sum_{j=1}^K w_j P(X|\mu_j, \sigma_j) \quad (3)$$

A. K-Means

K-Means adalah metode *clustering* yang digunakan untuk mengelompokkan data dengan cara menemukan pusat cluster atau centroid. Kemiripan data dengan data lain didasarkan pada seberapa dekat data tersebut dengan pusat atau centroid cluster. Proses *K-Means* akan berhenti ketika jumlah iterasi yang ditentukan telah tercapai dan data telah berhasil dikelompokkan. Secara umum pengelompokkan data menggunakan *K-Means* menggunakan jarak Euclidean kuadrat sebagai ukuran kesamaan untuk keanggotaan cluster[3]:

data, semakin jaraknya berjauhan, maka semakin tinggi nilai SSE. *Cluster* yang mengalami penurunan nilai SSE yang paling besar akan membentuk siku pada grafik, maka *cluster* tersebut ideal untuk selanjutnya digunakan pada proses *clustering*[5]. Rumus SSE untuk *K-Means*:

Gaussian Mixture Model (GMM) merupakan teknik clustering yang membentuk cluster berdasarkan *probability density function*. Setiap *cluster* memiliki tiga parameter, yaitu bobot, mean, dan variansi. Formula GMM sebagai berikut[3]:

Keterangan:

K = jumlah *cluster*

w = bobot

X = data harga rumah μ = mean

σ = variansi

Dalam proses *clustering* GMM dilakukan

$$\log[p(X|\theta)] = \sum_{i=1}^N \log \left(\sum_{j=1}^K w_j P(X|\mu_j, \sigma_j) \right) \tag{4}$$

Keterangan:

N = banyaknya data K = jumlah *cluster*

θ = parameter

w = bobot atau peluang mixing X = data harga rumah

μ = mean

σ = variansi

Namun pada metode MLE terdapat kekurangan, yaitu sering terjadi masalah dalam menentukan nilai yang optimal dari ketiga parameter GMM, maka digunakan algoritma *Expectation-Maximization* untuk memaksimalkan nilai dari ketiga parameter tersebut.

D. Algoritma *Expectation-Maximization*

Algoritma *Expectation-Maximization* (EM) merupakan algoritma yang dapat digunakan untuk memaksimalkan estimasi parameter GMM pada metode MLE. Pada setiap iterasi

proses mengestimasi ketiga parameter pembentuk PDF. Metode yang paling umum dalam mengestimasi parameter GMM adalah *Maximum Log-Likelihood Estimation* (MLE). Target dari metode tersebut adalah mengestimasi parameter yang dapat digunakan pada GMM. MLE dijelaskan oleh rumus sebagai berikut[3]:

algoritma EM terdapat dua langkah untuk melakukan proses, yaitu langkah *Expectation* dan *Maximization*. Algoritma EM memiliki proses sebagai berikut:

1. Inialisasi ketiga parameter GMM yaitu bobot(w), mean(μ), dan variansi(σ) secara acak lalu dilakukan evaluasi fungsi *log-likelihood* menggunakan ketiga parameter tersebut yang memenuhi persamaan(4)
2. Lalu dilanjutkan dengan proses *Expectation* dilakukan penghitungan mengikuti iterasi yang ada pada probabilitas dengan menggunakan rumus:

$$P(j|X_i) = \frac{w_j P(X_i|\mu_j, \sigma_j)}{\sum_{j=1}^K w_j P(X_i|\mu_j, \sigma_j)} \tag{5}$$

Keterangan:

w = bobot atau peluang mixing X = data harga rumah

N = banyaknya data μ = mean

σ = variansi

3. Langkah selanjutnya yaitu proses *Maximization* parameter akan terus diperbarui selama iterasi.

4. Pada Langkah terakhir, log-likelihood dievaluasi dan kriteria rumus berikut diperiksa.

$$(\| \log[p(X|\theta_{t+1})] - \log[p(X|\theta_t)] \| < \epsilon) \tag{6}$$

Keterangan:

$\log[p(X|\theta_{t+1})]$ = log-likelihood pada iterasi t+1

$\log[p(X|\theta_t)]$ = log-likelihood pada iterasi

ϵ = batas error prediksi

Jika kondisi tersebut belum terpenuhi, maka proses diulang dari tahap expectation. Dimana ϵ menjadi batas error prediksi.

- E. Bayesian Information Criterion
Bayesian Information Criterion (BIC) adalah

$$BIC = -2 \log p(X|\theta) + K \log(N) \tag{7}$$

Keterangan:

$\log p(X|\theta)$ = log-likelihood
 K = banyaknya cluster

N = banyaknya data

Hasil penghitungan BIC untuk masing-masing model kemudian diambil nilai BIC terkecil untuk digunakan pada penentuan jumlah cluster pada clustering GMM.

metode yang digunakan untuk seleksi model dengan tujuan mengetahui jumlah cluster optimal pada GMM. BIC memanfaatkan nilai log-likelihood dalam penghitungannya. Semakin kecil nilai BIC, maka cluster yang terbentuk akan semakin baik. Rumus penghitungan BIC[6]:

- F. Silhouette Score
Silhouette Score adalah nilai untuk mengukur

$$s(j) = \frac{b(j) - a(j)}{\max\{a(j), b(j)\}} \tag{8}$$

Keterangan:

s = nilai silhouette

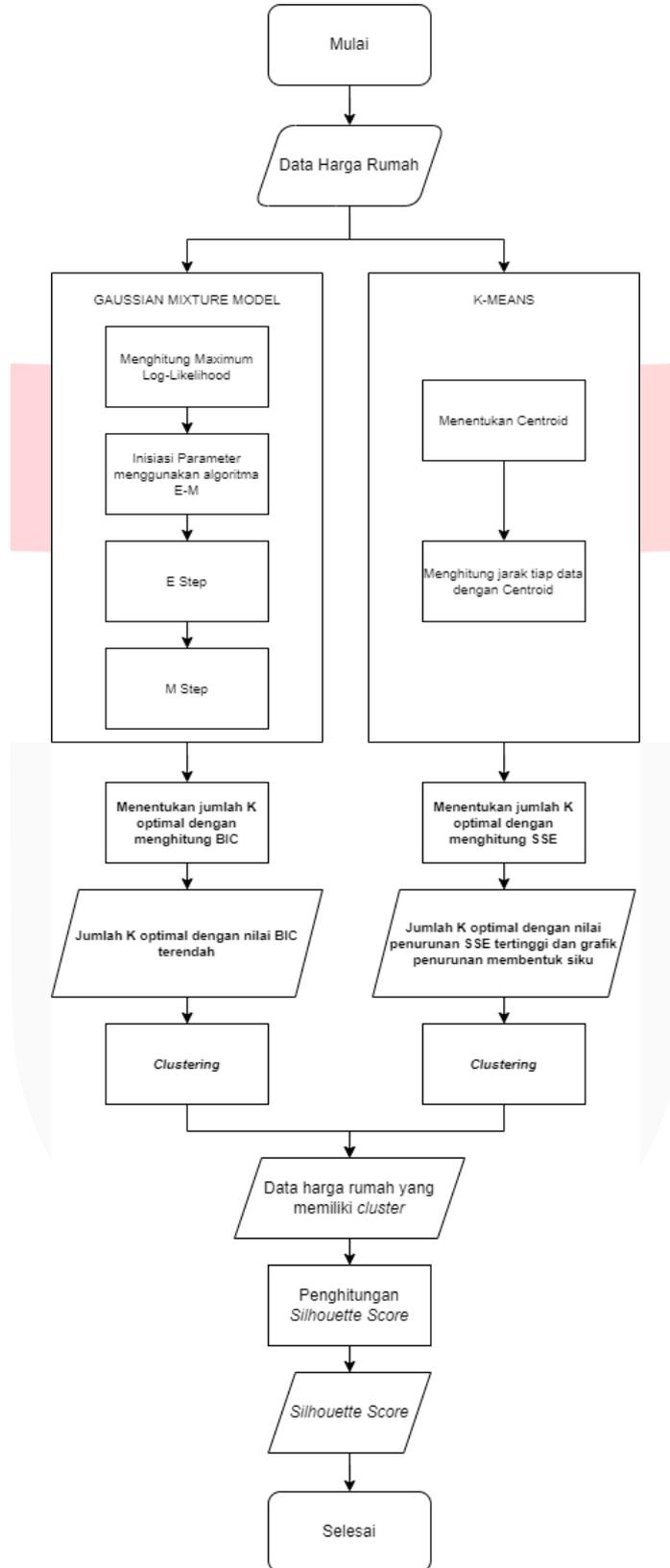
a(j) = rata-rata jarak dari objek j dengan objek yang berada di cluster berbeda

b(j) = rata-rata jarak dari objek j dengan seluruh objek yang berada di cluster yang sama

kualitas suatu cluster pada proses clustering. Silhouette akan menghitung nilai rata-rata semua data dalam setiap cluster, nilai yang dihitung merupakan selisih antara nilai separation dan compactness, dibagi dengan nilai maksimum antara kedua nilai[7]. Rumus untuk mencari Silhouette Score[5]:

Silhouette Score memiliki rentang nilai antara 0 sampai 1 dimana performa clustering dihitung dari nilai yang semakin mendekati 1 menjadi cluster yang paling baik.

III. METODE



GAMBAR 1 FLOWCHART SISTEM

A. Pengumpulan Dataset

Dataset yang digunakan pada penelitian ini berisi data harga jual rumah di wilayah Kabupaten Bogor. Data dihimpun dari website olx.co.id dengan menggunakan *tools Instant Data Scraper* yang tersedia pada ekstensi

Google Chrome. *Tools* ini mengambil data spesifikasi rumah seperti harga, luas bangunan, jumlah kamar tidur, dan jumlah kamar mandi.

TABEL 1
DATA MENTAH YANG TERHIMPUN

Harga	Spesifikasi
800000000	4 KT – 3 KM – 99 m2
700000000	3 KT – 2 KM – 82 m2
270000000	2 KT – 1 KM – 55 m2
410000000	4 KT – 2 KM – 69 m2
1600000000	3 KT – 2 KM – 110 m2
425000000	2 KT – 1 KM – 70 m2
...	...

Data diatas kemudian dilakukan *pre-processing* data dengan menghilangkan atribut yang tidak akan digunakan pada proses *clustering*. Data juga disaring dengan menghilangkan data yang terduplikat dan mengandung simbol-simbol yang dapat

mengganggu proses *clustering*. Berikut adalah dataset yang telah dilakukan *pre-processing* dan atribut yang akan digunakan hanya harga dan luas bangunan.

TABEL 2
DATA HASIL PRE-PROCESSING

Luas Bangunan	Harga
800000000	4 KT – 3 KM – 99 m2
700000000	3 KT – 2 KM – 82 m2
270000000	2 KT – 1 KM – 55 m2
410000000	4 KT – 2 KM – 69 m2
1600000000	3 KT – 2 KM – 110 m2
425000000	2 KT – 1 KM – 70 m2
...	...

B. Skenario Penelitian

Penelitian dimulai dari proses pada *K-Means* dengan mencari jumlah K menggunakan *Elbow Method*. Pada proses *Elbow Method* dilakukan proses iterasi sampai angka 15, lalu akan didapatkan jumlah K yang optimal dengan melihat grafik yang membentuk siku dan selanjutnya K tersebut digunakan untuk proses *clustering K-Means*. Lalu pada proses *Gaussian Mixture Model* penentuan jumlah K

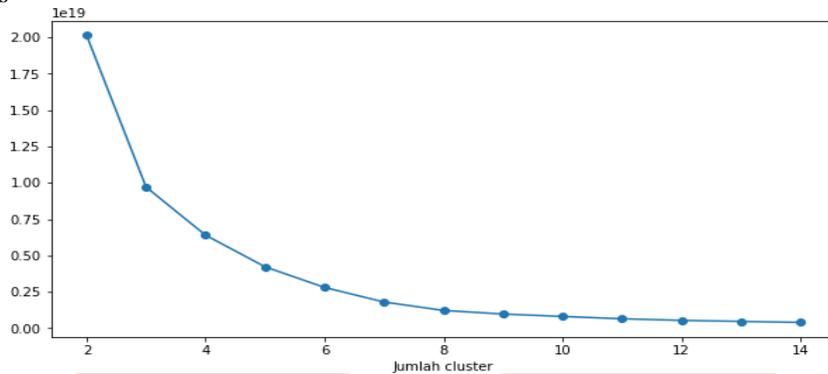
dilakukan dengan menghitung nilai BIC, dan dicari nilai BIC terendah dari iterasi sampai 15, lalu K dengan nilai BIC terendah akan digunakan untuk proses *clustering Gaussian Mixture Model*.

IV. HASIL DAN PEMBAHASAN

A. Hasil Pengujian

1. K-Means

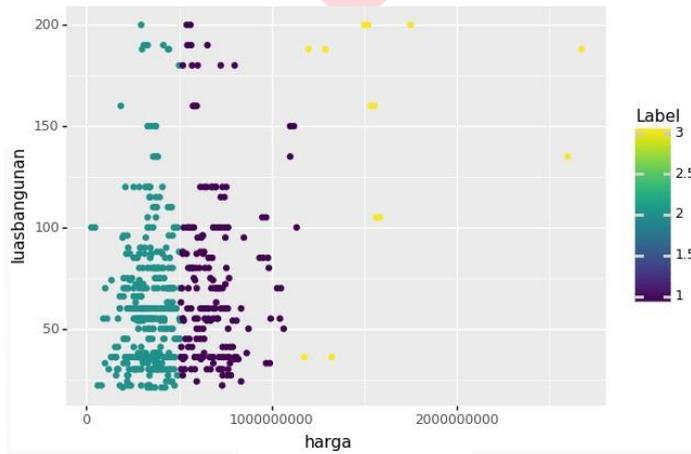
Proses *K-Means* diawali dengan mencari jumlah *K* dengan menggunakan *Elbow Method*.



GAMBAR 2
GRAFIK PENURUNAN NILAI SSE PADA *ELBOW METHOD*

Pada proses *Elbow Method*, dilakukan iterasi pada range *K* dari 2 sampai 14, proses iterasi akan berhenti pada saat iterasi telah

sampai pada jumlah *K* = 14. Penurunan nilai SSE terbesar terlihat pada *K*=3 grafik membentuk *Elbow* atau menyiku

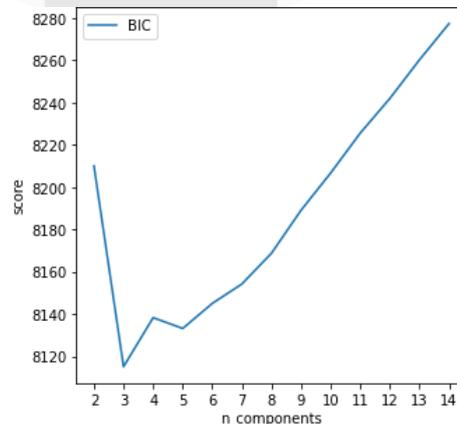


GAMBAR 3
PLOT HASIL *CLUSTERING K-MEANS*

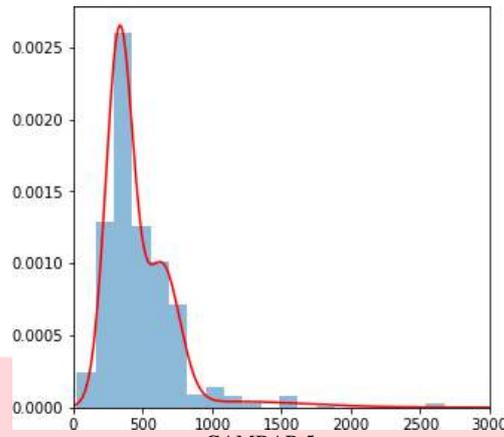
Dapat dilihat pada gambar 3, hasil *clustering K-Means* terbagi menjadi 3 *cluster*, dan masing-masing *cluster* memiliki jumlah anggota yang beragam. *Cluster 1* memiliki

396 anggota, *Cluster2* memiliki 190 anggota, dan *cluster 3* memiliki 13 anggota.

2. Gaussian Mixture Model



GAMBAR 4.
PENGHITUNGAN BIC



GAMBAR 5
PLOT PROBABILITY DENSITY FUNCTION

Dilihat pada gambar 4, nilai BIC terendah terdapat pada $K = 3$, lalu selanjutnya digunakan pada proses *clustering Gaussian Mixture Model*.

Pada gambar 5 menunjukkan jumlah PDF sebanyak 3, sesuai dengan

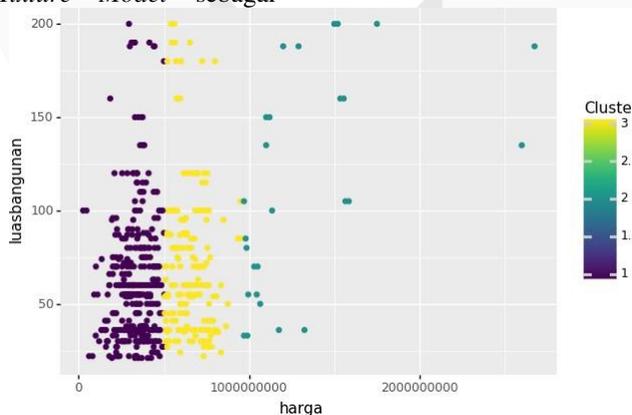
jumlah K yang telah dicari pada proses BIC. Pemodelan diatas terbentuk dari ketiga parameter yang dimiliki oleh masing-masing cluster, yaitu bobot(w), mean(μ), dan variansi(σ). Dengan rincian nilai parameter dari masing-masing cluster sebagai berikut:

TABEL 3
PARAMETER MASING-MASING CLUSTER *GAUSSIAN MIXTURE MODEL*

Parameter	Cluster 1	Cluster 2	Cluster 3
Bobot	0.63942	0.05349	0.30707
Mean	$3.3179985864 \times 10^{10}$	$1.17449695842 \times 10^{11}$	$6.3867111393 \times 10^{10}$
Variansi	$9.72034323353 \times 10^{11}$	$2.736697318582 \times 10^{12}$	$1.621595315794 \times 10^{12}$

Lalu didapat plot hasil *clustering Gaussian Mixture Model* sebagai

berikut:



GAMBAR 6
PLOT HASIL CLUSTERING *GAUSSIAN MIXTURE MODEL*

Dapat dilihat pada gambar 6, hasil *clustering Gaussian Mixture Model* terbagi menjadi 3 cluster, dengan rincian Cluster 1 memiliki 393 anggota, Cluster 2 memiliki 27 anggota, dan cluster 3

memiliki 179 anggota.

3. Penghitungan *Silhouette Score*
Setelah dilakukan proses *clustering*, lalu dilakukan penghitungan *Silhouette Score* untuk

membandingkan kualitas cluster dari dua proses *clustering* yang telah dilakukan. Dari penghitungan tersebut menghasilkan metode *K-Means* memperoleh *Silhouette Score* sebesar 0.63516 dengan waktu proses *clustering* 1.97 detik, sedangkan GMM memperoleh *Silhouette Score* sebesar 0.62723 dengan waktu proses *clustering* 7.01 detik.

B. Analisis Hasil Pengujian

Dari hasil *clustering* yang telah dilakukan dengan dua metode yaitu *K-Means* dan GMM, penentuan jumlah K pada masing-masing metode menghasilkan jumlah K yang sama. Dapat dilihat pada gambar 2, metode *K-Means* dengan menggunakan *Elbow Method* pada titik K = 3 mengalami penurunan nilai SSE yang signifikan sehingga membentuk siku, dan pada gambar 4 untuk penghitungan jumlah K GMM dengan menggunakan BIC menghasilkan nilai BIC terendah ada pada K = 3. Lalu dilakukan proses *clustering* dengan menggunakan jumlah K yang telah ditentukan. Hasil *clustering* kemudian diuji kualitas *clusternya* dengan menghitung *Silhouette Score* dan waktu yang dibutuhkan untuk proses *clustering*. *K-Means* mendapat *Silhouette Score* sebesar 0.63516 dan GMM mendapat *Silhouette Score* sebesar 0.62723. Dari pengujian tersebut menghasilkan *K-Means* unggul dari kedua aspek, dari segi nilai *Silhouette* unggul tipis 0.01, yang menjadikan kualitas *cluster* pada *K-Means* lebih baik daripada GMM, dan dari segi waktu *K-Means* unggul cukup jauh dengan selisih 5 detik.

V. KESIMPULAN

Proses *clustering* dengan GMM diawali dengan inialisasi parameter secara acak, lalu dilakukan optimasi pada ketiga parameter dengan menggunakan algoritma EM yang prosesnya berulang sampai nilai log-likelihood konvergen. Hasil dari kedua proses *clustering* menunjukkan bahwa *K-Means* memberikan performa yang lebih baik dari *Gaussian Mixture Model*, dilihat dari 2 aspek yaitu nilai *Silhouette* dan waktu yang dibutuhkan dalam proses *clustering*. *K-Means* mendapat nilai *Silhouette* 0.63516 dengan waktu proses 1.97 detik, sedangkan *Gaussian Mixture Model* mendapat nilai *Silhouette* 0.62723 dengan waktu proses 7.01 detik. Dengan demikian disimpulkan bahwa pada penelitian kali ini, proses *clustering* dengan *K-Means* mendapat hasil lebih baik daripada *Gaussian Mixture Model*.

Adapun saran dari peneliti untuk penelitian selanjutnya, dapat dilakukan perbandingan dengan

metode *clustering* lain, lalu lingkup wilayah penelitian bisa diperluas lagi, dan jumlah record dataset bisa diperbanyak.

REFERENSI

- [1] Yuwono, Muhammad J.. 2015. Penghitungan Kepadatan Kendaraan Di Jalan Tol Menggunakan Metode *Gaussian Mixture Model* dan Kalman Filter. Telkom University.
- [2] L. Handayani. 2012. Identifikasi Area Kanker Ovarium pada Citra CT Scan Abdomen Menggunakan Metode Expectation Maximization. UINSUSKA
- [3] Patel, Eva., Kushwaha, Dharmender S.. 2020. Clustering Cloud Workloads: K-Means vs Gaussian Mixture Model. Elsevier.
- [4] L. Rokach. 2005. "Clustering Methods", Data Mining and Knowledge Discovery Handbook, pp 331-352, Springer.
- [5] Izzadin, F.M. 2020. Optimasi Jumlah Cluster K-Means dengan Metode Elbow dan Silhouette pada Produktivitas Tanaman Pangan di Provinsi Jawa Tengah Tahun 2019. Universitas Islam Indonesia.
- [6] Watanabe, S. 2012. "A widely applicable Bayesian information criterion," arXiv Prepr. arXiv1208.6338.
- [7] A. Aditya, I. Jovian, and B. N. Sari. 2020. "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.