PRISMA 6 (2023): 745-750
PRISMA, Prosiding Seminar Nasional Matematika
https://journal.unnes.ac.id/sju/index.php/prisma/
ISSN 2613-9189

Gaussian Mixture Model dengan Algoritme *Expectation Maximization* untuk Pengelompokan Data Distribusi Air Bersih di Jawa Barat

Rizqi Ummami[a,*], Bowo Winarno[b]

[a, b] *Universitas Sebelas Maret, Jebres, Surakarta, 57126, Indonesia*

* *Alamat Surel: rizqi_ummami5@students.uns.ac.id*

Abstrak

Air bersih menjadi sesuatu yang tidak dapat dipisahkan dari kehidupan sehari-hari. Ketersediaan air bersih harus tetap terjaga agar kehidupan masyarakat sejahtera dan tidak terjadi krisis air bersih. Namun, meskipun jumlah air relatif tetap menurut ruang dan waktu, kebutuhan terhadap air bersih semakin tinggi akibat pertumbuhan jumlah penduduk dan taraf hidup yang semakin meningkat. Dari data yang ada, kemudian akan diolah menggunakan *Gaussian Mixture Model* (GMM) dengan algoritme *Expectation Maximization* (EM) untuk mengelompokan data distribusi air bersih di Jawa Barat. Penelitian ini diharapkan dapat menjadi acuan dalam menentukan kebijakan atau upaya untuk memaksimalkan distribusi air bersih di Jawa Barat. Pada data tersebut dapat dikelompokkan menjadi beberapa *cluster* yang dihitung menggunakan *Bayesian Information Criterion* (BIC). Dari hasil *clustering* menunjukan bahwa, ada beberapa kabupaten/kota di Jawa Barat belum sepenuhnya memperoleh distribusi air bersih dari PDAM secara maksimal seperti Kabupaten Purwakarta dan Kota Banjar.

Kata kunci:
Air bersih, *clustering*, *Expectation Maximization*, *Gaussian Mixture Model*

1. Pendahuluan

Air merupakan salah satu kebutuhan hidup manusia yang sangat penting. Akses air minum layak adalah air minum yang berasal dari sumber air terlindung seperti sumur bor atau sumur pompa, air ledeng (keran), keran umum, hydrant umum, terminal air, penampungan air hujan (PAH) atau mata air dan sumur terlindung, yang jaraknya minimal 10meter dari pembuangan kotoran, pembuangan sampah dan penampungan limbah (Badan Pusat Statistik, 2018).

Ketersediaan air relatif tetap walaupun bervariasi menurut ruang dan waktu, sedangkan kebutuhan air cenderung terus meningkat karena jumlah penduduk dan taraf hidup yang meningkat (Nugroho, 2018). Keadaan jumlah penduduk yang terus meningkat, akan menimbulkan ketimpangan di bebagai daerah tidak terkecuali di Jawa Barat. Untuk mengetahui kondisi distribusi air bersih di wilayah Jawa Barat, dapat menggunakan teknik *data mining*. Teknik *data mining* yang dapat dimanfaatkan untuk memperoleh informasi dari data distribusi air bersih di Jawa Barat yaitu *clustering.*

Pada penelitian ini, akan dilakukan pengelompokan distribusi air bersih kabupaten/kota di Jawa Barat ke dalam beberapa *cluster* dengan *Gaussian Mixture Model* (GMM) dan algoritme *Expectation Maximization* (EM). Penelitian dengan menggunakan *Gaussian Mixture Model* pernah dilakukan oleh Putra & Afifah (2018) untuk menghitung tingkat kebersihan sungai. Jurnal tersebut membahas mengenai bagaimana memperoleh informasi tingkat kebersihan sungai dengan metode *Gaussian Mixture Model* (GMM) berbasis pengolahan citra.

Clustering untuk data distribusi air bersih pada penelitian ini menggunakan 3 variabel bebas dan 1 variabel terikat. Hasil dari pengelompokan atau *clustering* tersebut dapat digunakan untuk menentukan

To cite this article:
Ummami, R., & Winarno, B. (2023). *Gaussian Mixture Model* dengan Algoritme *Expectation Maximization* untuk Pengelompokan Data Distribusi Air Bersih di Jawa Barat. *PRISMA, Prosiding Seminar Nasional Matematika* 6, 745-750.

kebijakan atau upaya yang tepat agar distribusi air bersih oleh PDAM merata dan maksimal untuk setiap kabupaten/kota di Jawa Barat.

2. Metode

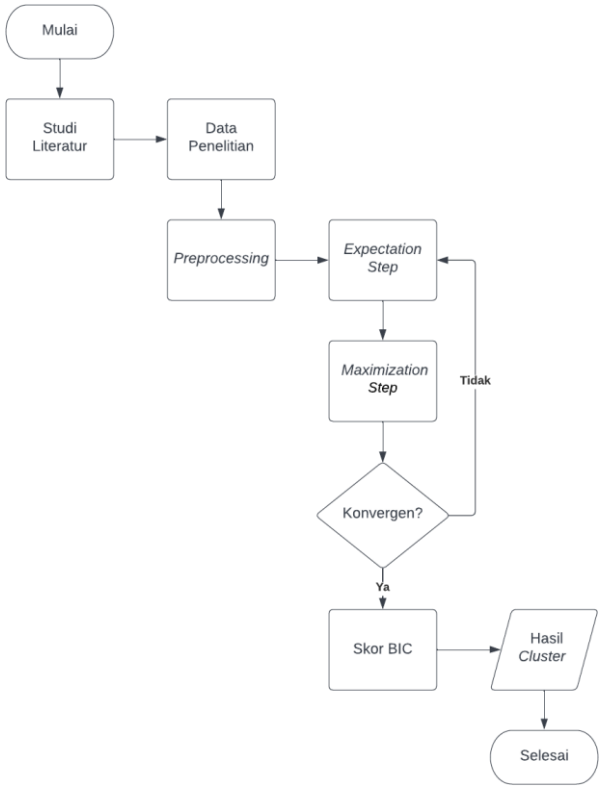
Metodologi penelitian dibagi menjadi dua bagian yaitu data penelitian dan langkah penelitian. Bagian pertama adalah data penelitian yang berisi jenis penelitian, sumber data, dan identifikasi variabel.

2.1. *Data Penelitian*

Penelitian ini merupakan penelitian terapan mengenai *Gaussian Mixture Model* (GMM) dengan algoritme *Expectation Maximization* (EM). Data yang digunakan adalah data sekunder dari Badan Pusat Statistik (BPS) Provinsi Jawa Barat (2021). Data tersebut merupakan data statistik air bersih di Jawa Barat tahun 2020 dari hasil survei perusahaan air bersih. Variabel-variabel yang digunakan dalam penelitian ini yaitu variabel bebas dan variabel tidak bebas atau terikat. Variabel bebas yang digunakan terdiri dari kapasitas potensial, kapasitas efektif, dan jumlah pekerja. Sedangkan, variabel terikatnya yaitu volume air bersih yang disalurkan.

2.2. *Langkah Penelitian*

Berikut merupakan langkah penelitian yang dilakukan untuk mencapai tujuan penelitian.



Gambar 1. Flowchart Langkah Penelitian

(1) Mempelajari literatur-literatur tentang *Gaussian Mixture Model* dan algoritme *Expectation Maximization*.
(2) Mengumpulkan data penelitian.
(3) Melakukan *preprocessing* data.
(4) Menghitung estimasi parameter dengan *Expectation Step* (E-Step).
(5) Menghitung estimasi parameter dengan *Maximization* Step (M-Step).

(6) Jika memenuhi syarat konvergen, maka lanjut ke langkah selanjutnya, jika tidak maka berulang ke langkah 4.
(7) Menentukan skor BIC untuk menentukan jumlah *cluster*
(8) Melakukan *clustering* dan analisis *cluster* data yang telah diperoleh.

3. Hasil dan Pembahasan

Data mining adalah proses yang menggunakan suatu teknik untuk mengekstraksi dan mengidentifikasi informasi dan pola pengetahuan baru yang tidak diketahui sebelumnya pada data (Stanton *et al.*, 2019). Teknik dalam data mining dapat diklasifikasikan menjadi *association, classification, prediction, clustering, outliner*, serta *trend* dan *evolution* (Widya & Sudarma, 2019).

Penelitian ini menggunakan teknik *data mining*, yaitu *clustering*. *Clustering* adalah suatu teknik untuk mengidentifikasi objek yang serupa dengan memperhatikan beberapa kriteria tertentu dan kemudian dikelompokkan menjadi beberapa *cluster* (Sari, 2020). *Clustering* didasarkan pada kesamaan data karena pola dalam satu *cluster* memiliki lebih banyak kemiripan daripada pola yang tidak berada dalam *cluster* yang sama (Sirait dkk., 2015).

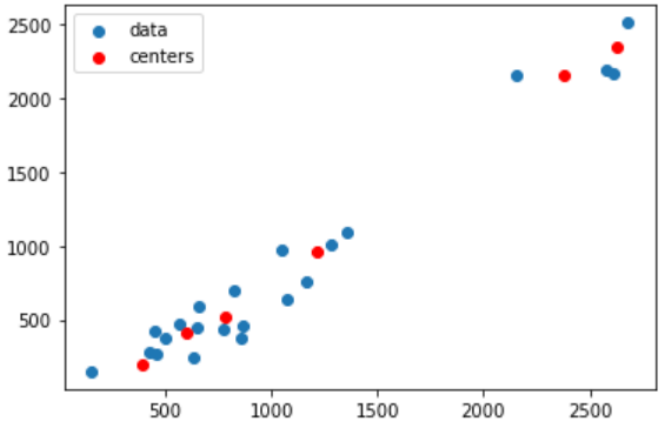
3.1. *Gaussian Mixture Model (GMM)*

Gaussian Mixture Model (GMM) termasuk dalam model berbasis *clustering* atau model distribusi (Deofanny dkk., 2022). Menurut Lin *et al.* (2019), parameter GMM diestimasikan menggunakan algoritme *Expectation Maximization* (EM) secara iteratif sehingga dapat terklasterkan dengan karakter yang serupa. Untuk *Gaussian Mixture Model*, diasumsikan bahwa sampel x yang diberikan adalah realisasi dari vektor acak yang distribusinya merupakan campuran dari beberapa distribusi, yaitu

$$P(X|\theta) = \sum_{k=1}^{K} \eta_k g(x_n|\theta_k) \quad (1)$$

dengan
η_k = peluang prior (awal) *cluster* ke-j
$g(x\_n|\theta\_k)$= fungsi kepadatan peluang dari variabel yang digunakan
k = banyaknya *cluster* yang terbentuk
θ = himpunan parameter

Gaussian Mixture Model memiliki tiga parameter, yaitu rata-rata, ragam, dan bobot campuran dari semua komponen GMM. Parameter tersebut akan diestimasi dan dimaksimalkan menggunakan algoritme *Expectation Maximization* (EM) dengan dua tahap, yaitu *Expectation Step* (*E-step*) dan *Maximization Step* (*M-Step*). Berikut merupakan visualisasi titik pusat cluster dengan menggunakan GMM.



Gambar 2 Visualisasi Titik Pusat *Cluster* dengan Menggunakan GGM

3.2. *Expectation Maximization (EM)*

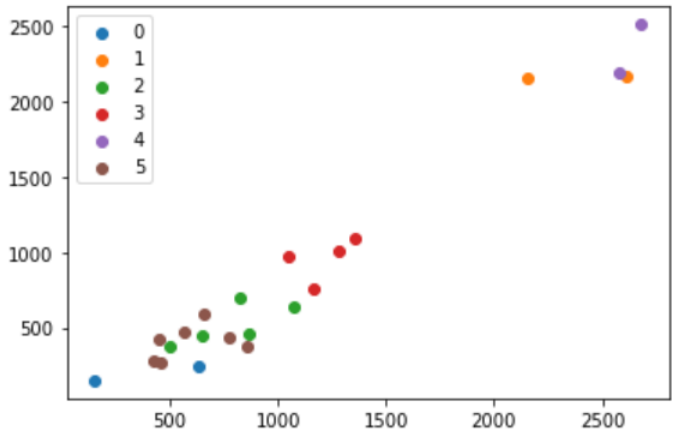
Clustering adalah suatu teknik untuk mengidentifikasi objek yang serupa dengan memperhatikan beberapa kriteria tertentu dan kemudian dikelompokkan menjadi beberapa *cluster* (Sari, 2020). Proses ini akan secara iteratif memberi skor ulang atau mengestimasi parameter terhadap kepadatan campuran yang dibuat oleh vektor parameter. Kemudian, estimasi parameter yang telah diketahui tersebut digunakan untuk memperbarui nilai parameter dengan memaksimalkannya.

Menurut Imro'ah dkk (2022), setiap iterasi algoritme EM terdiri dari dua proses, yaitu *Expectation Step* (*E-Step*) dan *Maximization Step* (*M-Step*). Proses *E-Step* digunakan untuk mencari suatu fungsi yaitu ekspetasi dari fungsi *log-likelihood* yang dinotasikan dengan:

$$E\left[\log[L(\theta)]\,\middle|\,x_n, \hat{\theta}^{r-1}\right] \tag{2}$$

Dimana $r-1$ menandakan parameter sebelum diestimasi.

Persamaan *log-likelihood* pada proses *E-Step* diturunkan terhadap $\hat{\theta}$ sehingga nilai taksiran dari η_k^r, μ_{ik}^r, dan σ_{ik}^r diperoleh pada masing-masing parameter. Berikut merupakah hasil dari perhitungan dengan algoritme *Expectation Maximization* (EM):



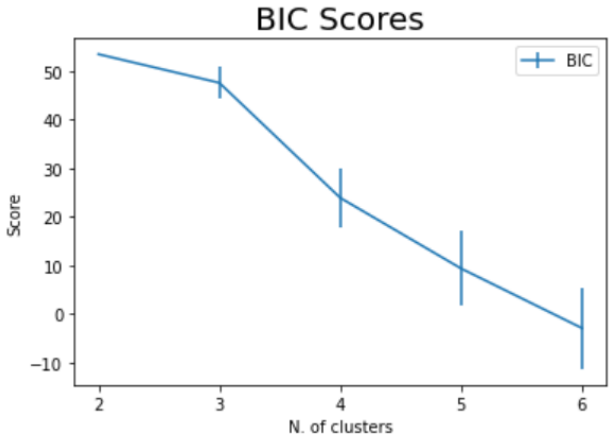
Gambar 3 Visualisasi *Clustering*

3.3. *Bayesian Information Criterion (BIC)*

Bayesian Information Criterion (BIC) menjadi salah satu metode untuk menentukan jumlah *cluster*. Nilai BIC yang terkecil atau terendah menjadi jumlah *cluster* terbaik (Sundari dkk., 2021). Persamaan BIC yaitu sebagai berikut (Webster, 2022)

$$BIC = 2\log\hat{l} + k\log n\ \ n \tag{3}$$

dengan $\hat{l}$ merupakan fungsi *log-likehood*, k adalah jumlah parameter dan n adalah jumlah data. Berikut merupakan hasil dari proses BIC untuk menentukan jumlah *cluster:*



Gambar 4 Nilai BIC untuk Menentukan Jumlah *Cluster*

Terlihat bahwa nilai *cluster* terendah atau terkecil berada di angka 6, yang berarti jumlah *cluster* terbaik dalam penelitian ini yaitu 6 *cluster*.

3.4. *Distribusi Air Bersih*

Berdasarkan data yang diperoleh dari Badan Pusat Statistik (BPS) Provinsi Jawa Barat (2021), yaitu publikasi Statistik Air Bersih Provinsi Jawa Barat 2020/2021, perusahaan air bersih memiliki beberapa kriteria utama untuk distribusi air bersih. Pengelompokan berdasarkan kriteria utama perusahaan air bersih dan air yang disalurkan ini menghasilkan 6 *cluster*. Hasil dari *clustering* distribusi air bersih di Jawa Barat yaitu sebagai berikut:

a. *Cluster* 1 : Kabupaten Purwakarta dan Kota Banjar
b. *Cluster* 2 : Kabupaten Bogor dan Kabupaten Bekasi
c. *Cluster* 3 : Kabupaten Garut, Kabupaten Kuningan, Kabupaten Cirebon, Kota Sukabumi, dan Kota Bekasi
d. *Cluster* 4 : Kabupaten Karawang, Kota Cirebon dan Kota Cimahi
e. *Cluster* 5 : Kabupaten Bogor dan Kota Bandung
f. *Cluster* 6 : Kabupaten Sukabumi, Kabupaten Cianjur, Kabupaten Ciamis, Kabupaten Majalengka, Kabupaten Sumedang, Kabupaten Subang, dan Kota Tasikmalaya

4. Simpulan

Berdasarkan hasil dan pembahasan diperoleh simpulan bahwa *Gaussian Mixture Model* menggunakan algoritme *Expectation Maximization* untuk mengestimasi dan memaksimalkan parameter. Parameter dari GMM yaitu rata-rata, ragam dan bobot campuran. Dari proses *clustering* tersebut diperoleh 6 *cluster*, dimana terlihat bahwa pada cluster 1, yaitu Kabupaten Purwakarta dan Kota Banjar merupakan kabupaten/kota yang perlu diperhatikan dan prioritaskan dalam pendistribusian air bersih oleh PDAM setempat.

Dengan mengetahui kondisi distribusi air bersih di kabupaten/kota di Jawa Barat, pihak PDAM ataupun pemerintah setempat dapat mengambil langkah atau upaya agar pendistribusian air bersih di Jawa Barat merata dan mudah untuk diakses.

Daftar Pustaka

BPS Provinsi Jawa Barat. (2021). Statistik Air Bersih Provinsi Jawa Barat 2020/2021. (Online). (https://jateng.bps.go.id/publication/2021/12/30/35d62d308e18472d5b41a7c0/statistik-air-bersih-provinsi-jawa-tengah-2020.html, diakses 01 September 2022)

Badan Pusat Statistik. (2018). Indikator Perumahan dan Kesehatan Lingkungan 2018. (*Online*). (https://www.bps.go.id/publication/2018/11/23/7a89433186e6103fa7c15b92/indikator-perumahan-dan-kesehatan-lingkungan-2018.html, diakses 25 September 2022)

Deofanny, N. F., Rohmawati, A. A., & Indwiarti, I. (2022). Model Gaussian Mixture Pada Distribusi Kecepatan Angin Dengan Algoritma Em. *EProceedings of Engineering*, *9*(3), 1978–1984.

Imro'ah, Nurfitri dkk. (2022). Implementasi Metode Latent Class CLuster Analysis dalam Pengelompokan Wilayah Berdasarkan Indikator Indeks Pembangunan Manusia. *Buletin Ilmiah Math. Stat dan Terapannya (Bimaster)*, 11(2), 213-220.

Lin, X., Yang, X., & Li, Y. (2019). A Deep Clustering Algorithm Based on Gaussian Mixture Model. Journal of Physics: Conference Series, 1302(3), 1-9.

Nugroho, S. P. (2018). Evaluasi Keseimbangan Air Di Provinsi Jawa Tengah. *Jurnal Air Indonesia*, 3(2), 175–181.

Putra, B. C & Afifah, Y. N. (2018). Gaussian Mixture Model untuk Penghitungan Tingkat Kebersihan Sungai Berbasis Pengolahan Citra. *Teknika : Engineering and Sains Journal*, 2(1), 53-58.

Sari, R. M. (2020). Implementation of Data Mining Using Clustering Methods for Analysis of Dangerous Disease Data. *International Journal of Research and Review*, 7(4), 237-242.

Sirait, R. E., Darwianto, E., Dwi, D., & Suwawi, J. (2015). *Implementasi dan AnalisisAlgoritma Clustering Expectation-maximization (EM) Pada Data Tugas Akhir Universitas Telkom*. *e-Proceeding of Engineering*, 2(2), 6711-6717.

Stanton, Jeffrey & Robert W. De Graaf. (n.d.). Version 3: An Introduction to Data Science. New York: Creative Commons.

Sundari, O., Bataradewa, S., & Matulessy, E. R. (2021). Penerapan Latent Class Cluster Analysis (LCCA) Pada Pengelompokan Kabupaten/Kota Di Provinsi Papua Barat Berdasarkan Indikator Kesejahteraan Rakyat. *Jurnal Natural*, *17*(2), 165–174.

Webster, Anthony J. (2022). Bayesian Information Criteria for Clustering Normally Distributed Data. *Nuffield Department of Population Health, Big Data Institute, Old Road Campus, University of Oxford*, 2022.

Widya, P. A., & Sudarma, M. (2019). Implementation of EM Algorithm in Data Mining for Clustering Female Cooperative. *IJEET International Journal of Engineering and Emerging Technology*, 3(1), 75–79.