# CLUSTERING KASUS PENCEMARAN LINGKUNGAN HIDUP DI INDONESIA MENGGUNAKAN ALGORITMA *FUZZY C-MEANS*

Inu Sinthia[1], Ginanjar Abdurrahman[2], Hardian Oktavianto[3]
Program Studi Teknik Informatika, Fakultas Teknik,
Universitas Muhammadiyah Jember
e-mail: inusintia266@gmail.com[1]

## ABSTRAK

Lingkungan hidup merupakan lingkungan utama yang sangat dekat dengan manusia yang dapat memberikan dampak positif jika dirawat dengan baik dan sebaliknya akan memberikan dampak negatif jika dibiarkan tercemar begitu saja. Saat ini kondisi hampir seluruh lingkungan hidup di dunia berada pada tingkat pencemaran yang menghawatirkan.Salah satunya Negara Indonesia yang telah mengalami pencemaran lingkungan seperti pencemaran tanah, pencemaran udara, pencemaran air. Kajian ini bertujuan untuk memberikan masukan bagi pemerintah dalam rangka pemerintah segera mengatasi pencemaran lingkungan hidup yang semakin meningkat. Metode yang digunakan untuk mengelompokkan provinsi di Indonesia berdasarkan banyaknya desa/lurah menurut jenis pencemaran lingkungan hidup pada tahun 2018 dengan menggunakan metode *clustering* yaitu algoritma *Fuzzy C-Means.*Untuk mengukur *cluster* optimum dalam menentukan *cluster* terbaik, metode yang digunakan adalah metode *Elbow.* Data yang digunakan yaitu banyaknya desa/ lurah di 34 provinsi di Indonesia tahun 2018. Dari banyaknya pengujian mulai dari 2 *cluster* sampai 10 *cluster*, hasil *cluster* optimum berada pada 2 *cluster* berdasarkan jarak SSE (*Sum of Squares Error*) pada metode *Elbow*. Pada *cluster* 1 terdiri 29 anggota provinsi dan *cluster* 2 terdiri dari 5 anggota provinsi. Berdasarkan hasil karakteristik data, cluster 1 memiliki jumlah kasus pencemaran lingkungan lebih rendah dibandingkan dengan *cluster* 2.

**Kata Kunci** : Pencemaran Lingkungan Hidup, *clustering, fuzzy c-means, elbow.*

# CLUSTERING CASES POLLUTION OF THE LIVING IN INDONESIA ALGORITHM *C-MEANS FUZZY*


## ABSTRACT

The main environment is very close to you can have a positive impact if are well cared for and instead will affect negative if its left just tainted. Now the condition almost all environment in the world be on a level pollution is worrying .One of them country which has experienced environmental pollution as pollution land , air pollution , water pollution . So this study attempts to inform the government that the government immediately reduce the increasing pollution of the living . Methods used to classify provinces in indonesia to the village / heads by the type of environmental pollution in 2018 by using the method is *fuzzy c-means clustering* algorithms. For measuring *clusters* in determining the optimum, *clusters* methods used is the method. Elbow the data used them is many / head of village indonesia 34 provinces in 2018.Many testing from 2 to 10 clusters clusters, the cluster steady is at 2 cluster based on the distance sse ( *Sum Of Squares Error* ) on method elbow. In clusters 1 29 consisting of provincial and clusters of 2. provincial 5 consisting of membersBased on the characteristics of, data clusters of 1 having the number of cases of pollution were lower than. 2 clusters.



**Keyword :** *Pollution of the living , Clustering, Fuzzy C-Means, Elbow*

## 1. PENDAHULUAN

Lingkungan merupakan kesatuan ruang lingkup sesuatu yang berada di sekitar mkhluk hidup yang memiliki pengaruh dan timbal balik terhadap kelestarian lingkungannya. Terdapat beberapa macam komponen lingkungan hidup antara lain manusia, hewan, tumbuhan, air dan juga komponen lainnya. Makhluk hidup akan selalu membutuhkan suatu interaksi dengan lingkungan sekitar tempat tinggalnya. Pengertian lingkungan kehidupan juga ditegaskan berdasarkan UU No 32 Tahun 2009, lingkungan merupkan suatu ke-satuan ruang dari semua benda, daya, kondisi, makhluk hidup, seperti manusia dan makhluk hidup lainnya. Berdasarkan pengertian di atas, dapat disimpulkan secara sederhana bahwa lingkungan hidup merupakan suatu kesatuan yang mencakup segala aspek kehidupan. Kondisi lingkungan memiliki peranan yang amat penting dan berkesinambungan dalam kelestarian suatu ekosistem. Dengan terciptanya suatu lingkungan hidup yang berkualitas, maka akan menjadikan manusia sehat. Banyak manfaat yang diperoleh dari lingkungan yang baik, contohnya yaitu pemanfaatan lingkungan hidup bagi makhluk hidup terutama manusia seperti udara untuk keperluan bernafas karena manusia tidak dapatt bertahan hidup tanpa adanya udara, air untuk minum dan juga dipergunakan untuk sumber pembangkit tenaga listrik. Menurut Anjelita M, tingginya jumlah penduduk menyebabkan terjadinya peningkatan akan bermacam-macam kebutuhaan. Dampak dari adanya kebutuhan tersebut, yakni dapat mengakibatka suatu kerusakan lingkungan di berbagai belahan, tidak terkecuali indonesia. Dari data PBB pada Hari Penduduk Dunia jatuh pada tanggal 11 Juli dengan populasi penduduk terus meningkat di setiap tahunnya. Data Perserikatan Bangsa-Bangsa (PBB) menyebutkan bahwa jumlah penduduk dunia pada tahun 2017 tercatat sebesar 7,6 miliar dan akan meningkat menjadi 8,6 milliar pada tahun 2030, 9,8 milliar pada tahun 2050. Data The Spector Index ditahun 2018 dari 20 negara dengan penduduk terbanyak di urutan nomer empat di dunia berdasarkan data Worldometers, Indonesia di tahun 2019 jumlah penduduk mencapai 269 juta jiwa atau 3,49% dari total populasi dunia hal tersebut menyebabkan kerusakan lingkungan yang terjadi di Indonesia. Salah satu kerusakan yang diakibatkan adalah pencemaran air di lingkungan. Air merupakan hal penting dari kehiidupan sehari-hari, tak hanya untuk manusia, namun bagi hewan dan tumbuhan juga.

Pada penelitian sebelumnya yang dilakukan oleh Anjelita (2019) dengan studi kasus "Pengembangan Data Mining Klastering Pada Kasus Pencemaran Lingkungan Hidup" Data yang digunakan adalah data pencemaran lingkungan hidup pada tahun 2018, Peneliti menggunakan dua metode data mining *clustering* yaitu metode pengelompokan *k-means* dengan pengelompokan *medoid k-means* menggunakan *software rapidminer* versi 5.3. Pada penelitian tersebut peneliti tidak menggunakan optimasi *cluster* dalam melakukan *clustering* dan telah berhasil mengumpulkan data pencemaran lingkungan hidup pada tahun 2018 menggunakan algoritma *K-means clustering* menjadi 2 *cluster*, yaitu *cluster* 1 berjumlah 4 provinsi, dan 30 lainnya berada di *cluster* 2. penelitian terhadap pencemaran lingkungan hidup di Indonesia ditahun 2018 dari data BPS karena jumlah pencemaran lingkungan hidup masih menunjukan banyak terjadinya pencemaran lingkungan oleh karna itu penulis melakukan penelitian data pencemaran lingkungan dengan judul " Clustering Kasus Pencemaran Lingkungan Hidup Di Indonesia Menggunakan Lagoritma *Fuzzy C-Means*". Pada penelitian ini terdapat 4 atribut yang di gunakan yaitu, pencemaran air, pencemaran tanah, pencemaran udara, tidak ada pencemaran. Selain menggunakan *Fuzzy C-Means* penulis juga melakukan kombinasi menggunakan dengan metode *Elbow*. Metode ini adalah metode yang digunakan untuk menghasilkan informasi dalam menentukan jumlah *cluster* terbaik dengan melihat presentase hasil perbandingan antara jumlah *cluster* yang akan membentuk siku pada suatu titik.

## 2. TINJAUAN PUSTAKA

### 2.1 *Data Mining*

*Data mining* adalah aktivitas yang menggambarkan sebuah proses analisis yang terjadi secara interaktif pada database yang besar, dengan tujuan mengestrak informasi informasi dan *knowledge* yang akurat dan berpotensial berguna untuk *knowledge workes* yang berhubungan dengan pengambilan keputusan dan pemecahan masalah (Vercellis, 2009).

Menurut Laronse & Daniel *data mining* dilakukan dengan *tool* khusus, yang mengeksekusi operasi *data mining* yang telah didenifikasi berdasarkan model analisis. *Data mining* merupakan proses analisis terhadap data dengan penekanan menemukan informasi yang tersembunyi pada jumlah data besar disimpan ketika bisnis perusahaan. Kemajuan Yang luar biasa terus berlanjut di bidang data mining yang didorong oleh beberapa faktor antara lain (Hidayahdarmin 2015):

1. Pertumbuhan pesat dalam kumpulan data
2. Penyimpanan data di data *warehouse*, sehingga semua perusahaan memiliki akses ke *database*.

3. Adanya peningkatan akses data melalui navigasi web dan internet.
4. Tekanan persaingan usaha untuk meningkatkan pangusaha pasar dalam globalisasi ekonomi.
5. Pengembangan teknologi perangkat lunak untuk data mining (ketersediaan teknologi).
6. Perkembangan besar dalam kapabilitas komputasi dan perkembangan kapasitas media penyimpanan.

*Data mining* adalah suatu yang dipakai untuk menggambarkan penemuan pengetahuan didalam basis data *Data mining* adalah proses yang memakai teknik statistik, matematika, kecerdasan buatan, dan *machine learning* untuk mengeksekusi dan mengindentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai basis data besar (Turban, dkk. 2005).

## 2.2 *Clustering*

Menurut Nugraha, dkk. (2014) *Clustering* adalah salah satu teknik *data mining* yang digunakan untuk mendapatkan kelompok – kelompok dari objek – objek yang mempunyai karakteristik yang umum di data yang cukup besar. Tujuan utama dari metode *clustering* adalah pengelompokan sejumlah data atau objek kedalam *cluster* atau grup sehingga dalam setiap *cluster* akan berisi data yang semirip mungkin. *Clustering* melakukan pengelompokan data yang didasarkan pada kesamaan antar objek oleh karena itu klasterisasi digolongkan sebagai metode *unsupervised learning*.

Pada dasarnya *clustering* merupakan suatu metode untuk mencari dan mengelompokkan data yang memilikikemiripan karakteristik (*similarity*) antara satu data dengan data yang lain. *Clustering* merupakan salah satu metode *data mining* yang bersifat tanpa arahan (*unsupervised*); maksudnya metode ini digunakan tanpa adanya latihan (*training*) dan tanpa ada guru (*teacher*) serta tidak memerlukan target *output*. Dalam *data mining* ada dua jenis metode *clustering hierarchiacal* dan *non-hierarchiacal clustering* (Santosa, 2007).

## 2.3 *Fuzzy C-Means*

*Fuzzy C-Means* pertama kali diperkenalkan oleh Jim Bezdek tahun 1981. *Fuzzy C-means* merupakan metode clustering dengan pendekatan *fuzzy*, artinya setiap data yang di *cluster* memungkinkan menjadi anggota lebih dari satu *cluster* (Kusumadewi & Purnomo, 2010). Konsep dasar *Fuzzy C-Means* adalah menentukan pusat *cluster*, pada kondisi awal pusat cluster ini masih belum akurat dan setiap data memiliki derajat keanggotaan untuk

tiap-tiap *cluster*. Dengan cara memperbaiki pusat *cluster* dan nilai keanggotaan tiap fungsi obyektif yang menggambarkan jarak dari titik data tertentu ke pusat cluster yang ditimbang oleh derajat keanggotaan titik data tersebut, ada beberapa algoritma *clustering* data, salah satu diantaranya adalah *Fuzzy C-means* adalah salah satu teknik klastering data di mana keberadaan tiap-tiap data dalam suatu *cluster* ditentukan oleh derajat keanggotaan. Teknik ini pertama kali diperkenalkan oleh Jim Bezdek pada tahun 1981. *Output* dari *Fuzzy C-Means* bukan *sistem inferensi fuzzy*, tetapi deretan pusat *cluster* dan beberapa derajat keanggotaan untuk setiap titik data. Informasi ini dapat digunakan untuk membangun *sistem inferensi fuzzy* (Kusumadewi & Purnomo, 2010). Langkah – langkah Algoritma *Fuzzy C- Means* adalah sebagai berikut:

**1.** Input data yang akan di cluster X, berupa matriks berukuran n x m (n = jumlah sampel data, m = tiap atribut data0. $X_{ij}$=sampel data ke-i (i=1,2,.....,n), atribut ke-j(1,2,.....,n).

**2.** Menentukan beberapa masukan yang dibutuhkan dalam perhitungan *Fuzzy C-Means* yaitu:

- Jumlah cluster (c)
- Pangkat (w)
- Iterasi maksimum (MaxIter)
- *Kesalahan terkecil (ξ)*
- Fungsi tujuan awal ($P_0 = 0$)
- Iterasi awal (t = 1)

**3.** Menentukan bilangan random µik, i=1,2,.....,n; k=1,2,.....,c; sebagai elemen-elemen matrik partisi awal U Hitung Jumlah setiap kolom

$$Qi = \sum_{k=1}^{c} \mu ik$$

$Q_1$ ialah jumlah setiap kolom dari nilai *random* sebuah matrik, jumlah Q tergantung dari beberapa jumlah kriteria penilaian.

**4.** Menghitung pusat cluster ke-k: $V_{kj}$, di mana k = 1,2, ....., c; dan j = 1,2, ....., m

$$V_{kj} = \frac{\sum_{i-1}^{n}\left((\mu_{ik})^{w} \times X_{ij}\right)}{\sum_{i-1}^{n}(\mu_{ik})^{w}}$$

$i$ = iterasi

$\mu_{ik}$ = perubahan matriks

$X_{ij}$ = atribut

$V_{kj}$ merupakan titik pusat dari setiap cluster, banyaknya $V_{kj}$ tergantung dari berapa cluster yang akan dibentuk.

**5.** Menghitung fungsi objectif pada iterasi ke-t$P_t$

$$P_t = \sum_{i-1}^{n}\sum_{k=1}^{c}\left(\left[\sum_{j=1}^{m}\left(X_{ij} - V_{kj}\right)^2\right](\mu_{ik})^w\right)$$

$P_t$ = fungsi objectif

$\sum_{i-1}^{n}$ = jumlah data di*cluster*

$\sum_{k-1}^{n}$ = jumlah perhitungan *cluster* awal

t merupakan iterasi terhitung, jika iterasi dimulai dari 1 maka pada awal perhitungan nilai t adalah 1. Iterasi tersebut akan berulang sesuai dengan ketentuan iterasi yang sedang berjalan. Hitung perubahan dalam matriks partisi.

$$\mu_{ik(t)} = \frac{\left[\sum_{j=1}^{m}\left(X_{ij} - V_{kj}\right)^2\right]^{\frac{-1}{w-1}}}{\sum_{k=1}^{c}\left[\sum_{j=1}^{m}\left(X_{ij} - V_{kj}\right)^2\right]^{\frac{-1}{w-1}}}$$

Iterasi akan terus berulang jika nilai atau kondisi tertentu belum tercapai, adapun kondisinya adalah jika (| Pt-1 | <ξ) atau (t> MaxIter) maka berhenti dimana Pt adalah pusat dari iterasi cluster ke t kurang dari nilai kesalahan yang diharapkan atau jika t (jumlah iterasi) lebih besar dari iterasi maksimum. Namun jika iterasi diulangi dengan t + 1 maka akan mengulangi proses ke-4 atau menghitung kembali pusat cluster (Kusumadewi & Purnomo, 2010).

### 2.4 Metode *Elbow*

Menurut Merliana, Emawati, & Santoso (2015), metode *Elbow* merupakan suatu metode penentuan jumlah *cluster* optimum atau terbaik untuk menghasilkan suatu informasi dengan melihat presentase perbandingan jumlah *cluster* yang akan membentuk siku pada suatu titik. Dengan metode ini memberikan gambaran dengan memilih nilai cluster kemudian menambahkan nilai cluster terbaik. Selain itu hasil persentase perhitungan yang dihasilkan akan menjadi perbandingan antara jumlah cluster yang ditambahkan, hasil presentase yang berbeda dari setiap nilai cluster dapat ditampilkan dengan menggunakan grafik sebagai sumber informasinya. Jika nilai *cluster* pertama dengan nilai *cluster* kedua memberikan sudut pada grafik atau nilai mengalami penurunan paling signifikan atau terbesar, maka nilai *cluster* tersebut adalah yang terbaik.Secara tabel, jarak antara 2 titik *cluster* dapat dihitung dengan cara mengurangi nilai SSE (*Sum of*

*Squares Error*) antara 2 titik *cluster*. Berikut ini merupakan rumus SSE.

$$\text{SSE} = \sum_{k-1}^{k} \sum_{xi \in Sk} \|Xi - Ck\|_2^2$$

Keterangan :

$X_1$ = fitur atau atribut dari data ke – *i*

$C_k$ = fitur atau atribut pusat *cluster* ke *i*

Algoritma metode *Elbow* dalam menentukan nilai K pada *K-Means* (Bholowalia, dkk. 2014 dalam Alatubir, 2017):

1. Inisialisasi awal nilai K
2. Naikkan nilai K
3. Hitung hasil *sum of square error* dari tiap nilai K
4. Melihat hasil *sum of square error* yang turun secara drastis
5. Tetapkan nilai K yang berbentuk siku.

Metode *Elbow* memberikan sebuah gagasan dengan cara memilih nilai *cluster* yang kemudian menambahkan nilai *cluster* terbaiknya. Presentase perhitungan yang diperoleh dalam perhitungan menjadi sebuah pembanding dimana jumlah *cluster* yang di tambah, hasil presentase yang berbeda dari tiap nilai *cluster* dapat ditunjukan dengan sebuah grafik atau dimana nilainya mengalami penurunan paling signifikan atau paling besar, maka dapat disimbulkan bahwa nilai *cluster* tersebut adalah terbaik. Secara tabel, jarak antara 2 titik *cluster* dapat dihitung dengan cara mengurangi nilai SSE (*Sum of Squares Error* ) antara 2 titik *cluster* (Merliana, Ernawati & Santoso, 2015).

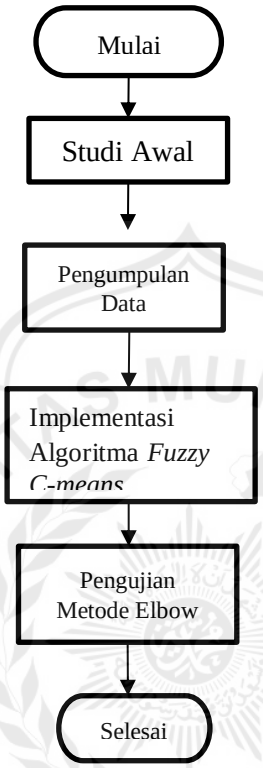
## 2.5 Rstudio

Rstudio adalah *itergrated development environment* (IDE) khusus untuk sebuah bahasa pemrograman R. *Software* yang menyediakan *R console, code editor* dengan *syntax highlighting, code completion, direct execution* dan banyak fitur-fitur lainnya. Rstudio juga mempunyai dua versi, yaitu *versi open source* (gratis) dan *commercial edition* (berbayar). R-studio tidak hanya terrbatas dengan aplikasi desktop saja tetapi ada juga versi Rstudio servernya yaitu Rstudio dengan aksesnya melalui browser yang terkoneksi dengan jaringan komputer(Putra,2018).

## 3. METODOLOGI

### 3.1 Tahapan Penelitian

Tahapan penelitian yang dilakukan dalam dalam menggunakan algoritma *Fuzzy C-Means* untuk mengelompokkan provinsi di Indonesia berdasarkan Pencemaran Lingkungan Indonesia, mempunyai tahapan seperti berikut:




Gambar 1. Diagram Alur Penelitian

### 3.2 Studi Awal

Studi awal pada penelitian ini adalah mencari dan mempelajari masalah pencemaran lingkungan hidup yang ada di masyarakat, kemudian menentukan ruang lingkup masalah, latar belakang, serta melakukan observasi dengan cara membaca beberapa artikel dan berita yang berkaitan dengan permasalahan pencemaran lingkungan hidup di masyarakat. Untuk mencapai tujuan, penulis mencari dan mempelajari metode pengelompokan data dengan membaca beberapa paper, artikel ilmiah dan makalah, kemudian peneliti menggunakan metode pengelompokan data sebagai salah satu solusi permasalahan pencemaran lingkungan hidup yang selama ini ada di masyarakat.

### 3.3 Pengumpulan Data

Dalam penelitian, data yang digunakan didapat melalui situs resminya Badan Pusat Statistik yang berjudul PENCEMARAN LINGKUNGAN HIDUP 2018 diakses pada link:

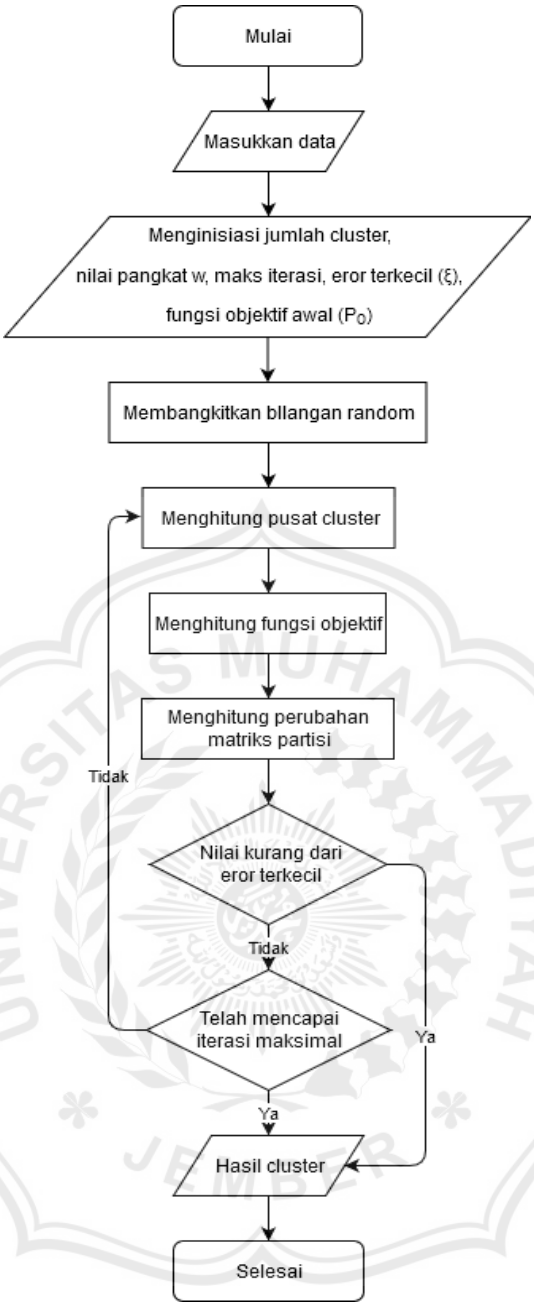
https://www.bps.go.id/dynamictable/2019/02/07/1578/banyaknya-desa-kelurahan-menurut-jenis-pencemaran-lingkungan-hidup-2014-2018.html. Data yang digunakan dalam penelitian ini adalah data Jenis Pencemaran Lingkungan di setiap Provinsi di Indonesia pada tahun 2018 terdiri dari 34 provinsi.

### 3.4 Dataset

Dataset pada penelitian ini adalah data PENCEMARAN LINGKUNGAN tahun 2018 yang terdiri dari 136 data dari 34 provinsi di Indonesia seperti pencemaran air, pencemaran tanah, pencemaran udara. Data yang digunakan hanya 15 data sampel untuk mewakili data yang dihitung.

### 3.5 Implementasi Algoritma *Fuzzy C-Means*

Diagram alur algoritma *Fuzzy C-Means* yang digunakan berdasarkan data Pencemaran Lingkungan, pada umumnya alur diagram *Fuzzy C-Means* adalah sebagai berikut:




Gambar 2. *Flowchart Fuzzy C-Means*

## 4. HASIL DAN PEMBAHASAN

### 4.1 Data Pengujian

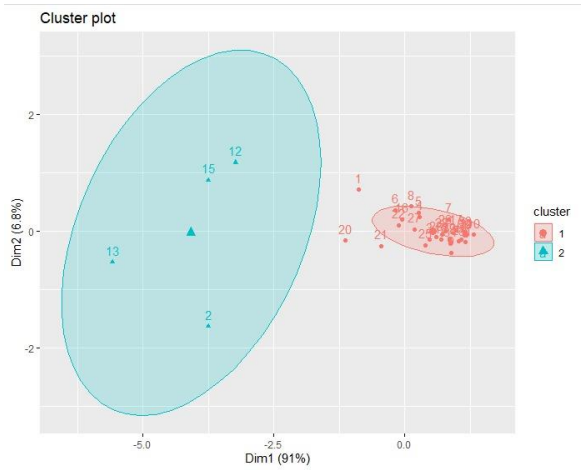
Berdasarkan hasil penelitian yang sudah di dapat dari data Pencemaran Lingkungan yang telah dikumpulkan akan di *cluster* menggunakan algoritma *Fuzzy C-Means* kemudian diolah untuk mendapatkan hasil cluster optimum atau cluster terbaik dengan menggunakan metode Elbow. Data yang digunakan merupakan sebuah data pencemaran lingkungan tahun 2018 yang terdiri 136 data yang terdiri dari 34 provinsi.

| No | Provinsi | Pencemaran Air | Pencemaran Tanah | Pencemaran Udara |
|---|---|---|---|---|
| 1 | ACEH | 0.365 | 0.132 | 0.468 |
| 2 | SUMATERA UTARA | 0.623 | 0.976 | 0.594 |
| 3 | SUMATERA BARAT | 0.143 | 0.127 | 0.098 |
| 4 | RIAU | 0.216 | 0.079 | 0.166 |
| 5 | JAMBI | 0.303 | 0.074 | 0.093 |
| 6 | SUMATERA SELATAN | 0.335 | 0.122 | 0.209 |
| 7 | BENGKULU | 0.125 | 0.013 | 0.080 |
| 8 | LAMPUNG | 0.280 | 0.066 | 0.183 |
| 9 | KEP. BANGKA BELITUNG | 0.056 | 0.119 | 0.013 |
| 10 | KEP. RIAU | 0.000 | 0.000 | 0.005 |
| 11 | DKI JAKARTA | 0.038 | 0.021 | 0.008 |
| 12 | JAWA BARAT | 0.995 | 0.376 | 0.642 |
| 13 | JAWA TENGAH | 1.000 | 1.000 | 1.000 |
| 14 | DI YOGYAKARTA | 0.024 | 0.045 | 0.033 |
| 15 | JAWA TIMUR | 0.861 | 0.481 | 0.876 |
| ... | ... | ... | ... | ... |
| 34 | PAPUA | 0.105 | 0.093 | 0.003 |

Tabel 1. Data Pencemaran Lingkungan Hidup

### 4.2 Fuzzy C-Means Pada Rstudio

Data diolah menggunakan Rstudio yang di *cluster* meggunakan algoritma *Fuzzy C-Means* dari 2 *cluster* sampai dengan 10 *cluster*. *Output* yang dapat dihasilkan dari eksekusi perintah pada Rstudio adalah jumlah iterasi, pusat *cluster*, fungsi objektif dan derajat keanggotaan setiap objek terhadap tiap *cluster*. *Cluster* yang dihasilkan dari perintah *Fuzzy C-Means* pada Rstudio ditampilkan ke dalam *plot* dari setiap *cluster* yang terbentuk. Berikut ini adalah contoh *plot* pada 2 *cluster* hasil *Fuzzy C-Means* di Rstudio;




Gambar 3. Plot 2 *cluster Fuzzy C-Means*

- *Cluster* 1 terdapat 30 provinsi yaitu:
  1. Provinsi Aceh
  2. Sumatra Barat
  3. Riau
  4. Jambi
  5. Sumatera Selatan
  6. Bengkulu
  7. Lampung
  8. Kepulauan Bangka Belitung,
  9. Kepulauan Riau
  10. DKI Jakarta
  11. DI Yogyakarta
  12. Banten
  13. Bali
  14. Nusa Tenggara Barat
  15. Nusa Tenggara Timur
  16. Kalimantan Barat
  17. Kalimantan Tengah
  18. Kalimantan Selatan
  19. Kalimantan Timur
  20. Kalimantan Utara
  21. Sulawesi Utara
  22. Sulawesi Tengah
  23. Sulawesi Selatan
  24. Sulawesi Tenggara
  25. Gorontalo
  26. Sulawesi Barat
  27. Maluku
  28. Maluku Utara
  29. Papua Barat
  30. Papua

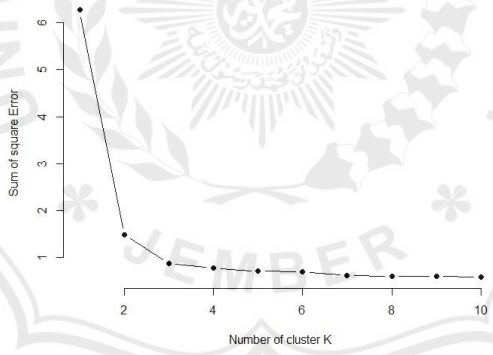
- *cluster* 2 terdapat 4 provinsi:
  1. Sumatra Utara
  2. Jawa Barat
  3. Jawa Tengah
  4. Jawa Timur.

## 4.3 Penentuan Jumlah *Cluster Optimum*

Setelah melakukan proses *cluster* menggunakan *Fuzzy C-Means*, kemudian melakukan proses metode *Elbow* untuk menentukan *cluster* optimum atau *cluster* terbaik. Hasil perhitungan metode *Elbow* pada Rstudio ditampilkan dalam nilai SSE (*Sum Of Squares Error*) dan grafik yang terdiri dari sumbu x dan y. Nilai pada sumbu x adalah jumlah yang dihasilkan dari pusat setiap *cluster*. Berikut adalah hasil metode *Elbow*.

| C | SSE | Jarak | Keterangan |
|---|---|---|---|
| 1 | 6.078833 | - | - |
| **2** | 1.49627 | 4.582563 | Jarak C1 ke C2 |
| 3 | 0.886134 | 0.610136 | Jarak C2 ke C3 |
| 4 | 0.570775 | 0.31536 | Jarak C3 ke C4 |
| 5 | 0.524775 | 0.046 | Jarak C4 ke C5 |
| 6 | 0.317883 | 0.206892 | Jarak C5 ke C6 |
| 7 | 0.251352 | 0.066531 | Jarak C6 ke C7 |
| 8 | 0.229574 | 0.021777 | Jarak C7 ke C8 |
| 9 | 0.188959 | 0.040615 | Jarak C8 ke C9 |
| 10 | 0.165146 | 0.023813 | Jarak C9 ke C10 |

Tabel 2. Hasil Nilai *Elbow*




Gambar 4. Plot Metode Elbow

## 4.4 Profiling Cluster Optimum

Pada metode *Elbow*, nilai *cluster* yang diambil sebagai *cluster* optimum atau terbaik adalah titik yang membentuk siku. Penjelasan pada titik yang membentuk siku terdapat dititik yang mengalamii penuruna signifikan diantara 2 titik *cluster* dan kemudian di ikuti dengan nilai-nilai yang relatif konstan. Dapat dilihat pada tabel 4.2, menunjukan nilai SSE *(Sum of Squares Error)* dengan jarak paling signifikan atau paling besar terdapat pada 2 *cluster* dengan jarak 1 *cluster* ke 2 *cluster*. Nilai jarak 1 *cluster* ke 2 *cluster*

tersebut merupakan nilai jarak yang mengalami penurunan paling signifikan atau paling besar dan kemudian diikuti oleh nilai jarak yang relatif konstan, sehingga 2 *cluster* merupakan *cluster* optimum atau terbaik. Dapat dilihat juga pada gambar 4.2, bahwa titik yang membentuk siku terdapat pada titik 2 *cluster*, dimana dari 1 *cluster* ke 2 *cluster* mengalami penurunan yang signifikan dibandingkan dengan yang lainnya. Kemudian dari titik 2 *cluster* ke titik selanjutnya diikuti nilai yang relatif konstan. Jadi, *cluster* optimum terdapat pada 2 *cluster*. Dari proses *clustering* menggunakan algoritma *Fuzzy C-Means* danMencari *cluster* optimum menggunakan metode *Elbow*, diketahui *cluster* optimum berada di 2 *cluster*. *cluster* 1 dan *cluster* 2

## 5. KESIMPULAN DAN SARAN

### 5.1 Kesimpulan

Berdasarkan hasil dari pembahasan permasalah pada bab sebelumnya, bahwa penerapan *Fuzzy C-Means* pada Pencemaran Lingkungan untuk mengelompokkan provinsi di Indonesian menghasilkan 2 *cluster* optimum dengan jarak *Sum of Square* (SSE) antara titik 1 *cluster* ke titik 2 *cluster* pada metode *Elbow* yaitu 6.078833 Jarak 1 *cluster* ke 2 *cluster* tersebut merupakan nilai jarak yang mengalami penurunan paling signifikan / paling besar dan kemudian diikuti oleh nilai jarak yang relatif konstan, sehingga 2 *cluster* merupakan optimum atau terbaik.

*Cluster* 1 terdapat 30 provinsi yaitu Aceh, Sumatra Barat, Riau, Jambi, Sumatera Selatan, Bengkulu, Lampung, Kepulauan Bangka Belitung, Kep. Riau, DKI Jakarta, Daerah Istimewa Yogyakarta, Banten, Bali, Nusa Tenggara Barat, Nusa Tenggara Timur, Kalimantan Barat, Kalimantan Tengah, Kalimantan Timur, Kalimantan Selatan, Kalimantan Utara, Sulawesi Tengah, Sulawesi Utara, Sulawesi Selatan, Sulawesi Tenggara, Gorontalo, Sulawesi Barat, Maluku, Maluku Utara, Papua Barat, dan Papua. Pada *cluster* 2 terdapat 4 provinsi terdiri dari Sumatra Utara, Jawa Barat, Jawa Tengah, Jawa Timur. Berdasarkan hasil dari karakteristik *cluster* 1 memiliki jumlah kasus pencemaran lingkungan lebih rendah dibandingkan dengan *cluster* 2.

## 5.2 Saran

Saran pada penelitian ini yaitu:

1. Perhitungan manual cluster Fuzzy C-Means di Microsoft Excel digunakan sebagai perbandingan hasil *cluster* dari Rstudio. Dan perhitungan manual di Microsoft Excel dapat menggunakan bilangan acak untuk menentukan cluster awal yang diambil dari Rstudio agar menghasilkan pengelompokan yang serupa.

2. Dan untuk mencari cluster optimum , bisa menggunakan metode alternatif lainnya seperti contoh *Silhouette, Gap Statistic, Davies Bouildin Index*, dll.

## 6. DAFTAR PUSTAKA


Amanda Pratama Putra, "Belajar Data Science: Memahami Layout Rstudio", 25 Mei 2018. <https://medium.com/@mandes95/belajar-data-science-memahami-layout-rstudio-d3d46f9f955c>. Diakses pada 22 Oktober 2020.

Anjelita Mawaddah , Widarto Perdana Agus , Hartama Dedy . 2019. " Pemanfaatan DataMining Pada Pengelompokkan Provinsi Terhadap Pencemaran Lingkungan Hidup.KOMIK (Konferensi Nasional Teknologi Informasi dan Komputer".

Arikunto, S. 1997 *Prosedur Penelitian : Suatu Pendekatan Praktik, Edisi Revisi*. Jakarta: Rineka Cipta

Bwolowalia, P. & Kumar, A. 2014. RBK-Means: *A Clustering Techiniques based on Elbow Method and K-Means in WSN. International Journal of Computer Aplication* (0975-8887), IX (105), 17-24.

Kusumadewi, S., & Purnomo, H. 2010. *Aplikasi Logika Fuzzy untuk Pendukung Keputusan edisi 2*. Yogyakarta: Graha Ilmu.

Nugraha, D. D. C., Naimah, Z., Fahmi, M. & Setiani, N. 2014. *Klasterisasi Judul Buku dengan Menggunakan Metode K-Means.* Seminar Nasional Aplikasi Teknologi Informasi. ISSN: 1907-5022. G02.

Merliana, N. P. E., Ermawati, & Santoso, A. J. 2015. *Analisa Penentuan Jumlah Cluster Terbaik pada Metode K-Means Clustering*. Yogyakarta: Program Studi Magister Teknik Informatika Universitas Atma Jaya.

Putri Ulfiyah Ananda, "Manfaat lingkungan bagi Manusia", 23 Juli 2013 <https://mediacenter.malangkota.go.id/2013/07/manfaat-lingkungan-bagi-manusia/>. Di akses pada 29 September 2020 .

Santosa, B.2007. *Teknik Pemanfaatan Data untuk Keperluan Bisnis.* Yogyakarta: Graha Ilmu.

Santoso, S.2010. *Statistik Multivariant*. Jakarta: Elex Media Komputindo.

Saputra Bintang Denny, Riksakomara. 2018 " Implementasi Fuzzy C-means dan Model RFM untuk Segmentasi Pelanggan (Studi Kasus: PT. XYZ) Jurnal Teknik ITS.

Turban, E., Ramesh, S., & Dursun, D. 2005. *Decision Sopport System and Intelligent Sytem* (Sistem Pendukung Keputusan Dan Sistem Cerdas). Yogyakarta:ANDI.

Vercellis, C. 2009. Business Intelligence: *Data Mining and Optimization for Decision Making*. United Kingdom: Wiley.

Wang, Lin Xin. 1997. A *Course in Fuzzy System and Control*, Prentice-Hall inc, New Jersey.

Welianto Ari, "Interasi Makhluk Hidup dengan Lingkungan ", 20 Januari 2020 https://www.kompas.com/skola/read/2020/01/20/120000069/interaksi-makhluk-hidup-dengan-lingkungan?page=all>. Di akses pada 19 September 2020.