Imple*mentas*i *machine learning* dalam pengelompokkan provinsi di indonesia berdasarkan data pencemaran lingkungan hidup

Annisa Nurul Azmi; Arsyka Laila Oktalia Siregar; Faqih Indra Lesmana; Andi Ardiansyah Nasir, Fitri Kartiasih*

Politeknik Statistika STIS

**E-mail korespodensi:* fkartiasih@stis.ac.id

Abstract

Environmental pollution is a crucial issue that needs serious attention. The increasing world population will also increase the level of environmental pollution (Mittal & Mittal, 2013) especially in developing countries (Remilekun Adeuti, 2020) This is also the case in Indonesia. Therefore, this research aims to find out which provinces in Indonesia have a high level of pollution by clustering provinces based on environmental pollution data. The methods used in this research are K-Medoids, K-Means, and Fuzzy C- Means as well as Complete Linkage and Ward's Linkage for Agglomerative Hierarchy. The results show that the K-Medoids method is the best method produces 3 clusters, namely clusters with high average pollution of 11 provinces, clusters with average pollution of 12 provinces, and clusters with low average pollution of 11 provinces.

Keywords: *pollution; clustering, environment.*

Abstrak

Pencemaran lingkungan merupakan isu krusial yang perlu mendapat perhatian serius. Bertambahnya populasi dunia juga akan meningkatkan tingkat pencemaran lingkungan (Mittal & Mittal, 2013) apalagi pada negara berkembang (Remilekun Adeuti, 2020) seperti halnya di Indonesia juga demikian. Maka dari itu, penelitian ini bertujuan untuk mengetahui provinsi di Indonesia yang memiliki tingkat pencemaran yang tinggi dengan melakukan pengelompokan provinsi berdasarkan data pencemaran lingkungan hidup. Metode yang digunakan dalam penelitian ini, yaitu K-Medoids, K-Means, dan Fuzzy C-Means serta Complete Linkage dan Ward's Linkage untuk Aglomerative Hierarchy. Hasilnya menunjukkan bahwa metode K-Medoids adalah metode terbaik dengan 3 cluster yang dihasilkan, yaitu cluster dengan rata-rata pencemaran tinggi sebanyak 11 provinsi, cluster dengan rata-rata pencemaran sedang sebanyak 12 provinsi, dan cluster dengan rata-rata pencemaran rendah sebanyak 11 provinsi.

Kata kunci: pencemaran, clustering, lingkungan.

PENDAHULUAN

Pencemaran lingkungan merupakan isu krusial yang perlu mendapat perhatian serius. Menurut KBBI, Pencemaran adalah proses atau perbuatan mencemari maupun mencemarkan udara dan lingkungan. Sedangkan pencemaran lingkungan adalah masuk atau dimasukkannya makhluk hidup, zat, energi, dan/atau komponen lain ke dalam air/udara, dan/atau berubahnya tatanan (komposisi) air/udara oleh kegiatan manusia dan proses alam, sehingga kualitas air/udara menjadi kurang atau tidak dapat berfungsi lagi

sesuai dengan peruntukannya (Kementerian Lingkungan Hidup, 1988). Untuk mengurangi terjadinya pencemaran lingkungan diperlukan pengendalian dengan menetapkan baku mutu lingkungan. Baku mutu lingkungan adalah batas kadar yang diperkenankan bagi zat atau bahan pencemar terdapat di lingkungan dengan tidak menimbulkan gangguan terhadap makhluk hidup, tumbuhan atau benda lainnya.

Dari berbagai jenis pencemaran, terdapat tiga pencemaran utama yang mempengaruhi lingkungan yaitu pencemaran air, pencemaran udara, dan pencemaran tanah (Ukaogo, Ewuzie, & Onwuka, 2020). Pertama, Pencemaran air disebabkan oleh Penggunaan air seluruh dunia terus meningkat. Namun, belum terdapat pengolahan limbah yang memadai dari kegiatan industri, irigasi pertanian, maupun kebutuhan rumah tangga sehingga air terkontaminasi dengan zat berbahaya (Lin, Yang, & Xu, 2022). Dengan adanya pencemaran air dapat menyebabkan berbagai penyakit, seperti diare, kelainan kulit, maupun kanker (Lin et al., 2022). Hal tersebut dapat terjadi saat mengonsumsi atau menggunakan air yang terpapar zat kimia berbahaya (Petrisor, n.d.). Kedua, Pencemaran udara seringkali tidak dapat dirasakan maupun terlihat. Berdasarkan AQI dan polusi udara PM2.5 di dunia, pada tahun 2023 Indonesia menempati peringkat 14 wilayah paling tercemar di dunia. Pencemaran udara disebabkan oleh asap kendaraan , kebakaran, industri, dan letusan gunung berapi yang mengeluarkan zat-zat berbahaya ke udara sehingga berdampak terhadap kesehatan pernapasan(Rice et al., 2021). Tidak hanya itu, rumah tangga juga menyumbang emisi gas rumah kaca sebesar 8% (Hartono et al., 2023). Selanjutnya, pencemaran tanah diakibatkan penggunaan zat kimia secara sembarangan seperti pupuk, pestisida, sampah yang tidak terurai, dan limbah industri(Bell, 1997). Pencemaran tanah mengakibatkan keracunan pada tanaman baik sayur-sayuran maupun buah-buahan yang dikonsumsi manusia sehingga menyebabkan masalah kesehatan seperti mual, sakit kepala, ruam kulit, maupun kulit(Petrisor, n.d.).

Pencemaran lingkungan baik tanah, air, maupun udara sangat berbahaya untuk kehidupan di bumi. Bukan hanya manusia, semua organisme yang ada dalam rangkaian rantai yang harmonis akan mengalami ketidakseimbangan yang menyengsarakan. WHO menyebutkan, bahwa 7 juta manusia meninggal setiap tahun yang disebabkan oleh polusi udara (WHO dateboks, 2019). Di habitat lain seperti laut, UNEP menyebutkan sampah mikro dari konsumsi rumah tangga menyebabkan Penyu, burung laut, dan binatang lainnya mengira itu sebuah makanan karena wujud dan aromanya. Bukan hanya itu, UNEP juga mengatakan bahwa penyebab utama kematian Paus Kanan Atlantik Utara, salah satu paus yang terancam punah di dunia adalah terjerat dalam alat penangkap jenis jaring hantu. Tentunya hal tersebut mengancam keberlanjutan kehidupan di dunia (Harris et al., 2021).

Bertambahnya populasi dunia juga akan meningkatkan tingkat pencemaran lingkungan (Mittal & Mittal, 2013) apalagi pada negara berkembang (Remilekun Adeuti, 2020) seperti halnya di Indonesia juga demikian. Hal ini disebabkan karena kebutuhan manusia meningkat. Dengan demikian sampah yang dihasilkan rumah tangga juga akan meningkat yang berakibat pada pencemaran. Pencemaran lingkungan menyebabkan *global warming* yang selanjutnya berdampak pada perubahan iklim yang drastis. Dengan demikian pencegahan pencemaran lingkungan menjadi hal yang mendesak.

Urgensi penanganan pencemaran lingkungan hidup sejalan dengan SDGs ke 6 yakni yang bertujuan untuk mendapatkan ketersediaan air bersih dan sanitasi layak. Hal ini karena pencemaran lingkungan khususnya pencemaran air menyebabkan kebutuhan akan air yang layak menjadi berkurang(KILIÇ, 2021). Pencemaran air juga berdampak pada Kesehatan (Sompotan & Sinaga, 2022) seperti penyakit diare bahkan malnutrisi (Lin et al., 2022). Selain SDGs ke 6, urgensi penanganan pencemaran lingkungan hidup juga

sejalan dengan SDGs ke-13 yang bertujuan untuk penanganan perubahan iklim, yang indikator didalamnya mencakup upaya-upaya untuk mengurangi emisi gas rumah kaca.

Penelitian ini bertujuan untuk mengetahui provinsi di Indonesia yang memiliki tingkat pencemaran yang tinggi dengan melakukan analisis *cluster* atau dilakukan pengelompokan provinsi berdasarkan data pencemaran lingkungan hidup. Dengan mengetahui provinsi atau wilayah mana yang memiliki tingkat pencemaran yang tinggi maka pemerintah dapat melakukan upaya pencegahan pencemaran lingkungan. Pemerintah memiliki tanggung jawab untuk mengatasi masalah pencemaran lingkungan di Indonesia. Pengendalian pencemaran lingkungan hidup yang dilakukan oleh pemerintah dilaksanakan dalam rangka pelestarian lingkungan hidup meliputi pencegahan, penanggulangan, dan pemulihan (Sompotan & Sinaga, 2022). Pemerintah harus mengetahui provinsi mana saja yang memiliki tingkat pencemaran lingkungan yang tinggi.

Analisis *cluster* atau *clustering* merupakan proses pengelompokan data ke dalam beberapa kelompok atau *cluster* sehingga data dalam satu *cluster* memiliki kemiripan yang tinggi satu sama lain, tetapi sangat berbeda dengan data di *cluster* yang lain. *Clustering* mencakup analisis *cluster* hierarki dan non hierarki. Pada analisis hierarki tidak perlu menentukan jumlah *cluster* pada awal tahapannya, sedangkan pada analisis non-hirarki perlu menentukan jumlah *cluster* pada awal tahapannya (Syafiyah, Puspitasari, Asrafi, Wicaksono, & Sirait, 2022). Pada penelitian ini metode yang digunakan dalam analisis *cluster* adalah K-Means, Fuzzy C-Means, K-Medoids, dan Hierarki *Ward*.

Penelitian dengan menggunakan analisis *cluster* pernah dilakukan oleh Herman et al., 2022 dalam mengevaluasi kinerja perusahaan ritel makanan di Hongaria dan Rumania. Mereka menggunakan metode K-Means dan K-Medoids berdasarkan rasio keuangan seperti ROS (*Return on Sales*), ROA (*Return on Assets*), dan ROE (*Return on Equity*). Dalam penelitian tersebut, metode K-Means menghasilkan variasi kelompok yang lebih banyak, sedangkan kelompok dan hasil yang diperoleh dengan metode K-Medoids lebih seimbang. Dalam penelitian lain yang dilakukan oleh Matlis et al., 2024 yang melakukan penelitian tentang analisis pengelompokan untuk pemeringkatan akademik untuk merangking universitas dengan melakukan *clustering* menggunakan beberapa metode seperti K-Means, GMM, Agglomerative, dan Fuzzy C-Means didapatkan bahwa algoritma Fuzzy C-Means merupakan metode yang paling baik dalam meng-*cluster* kan untuk tujuh *cluster* tanpa bobot hal ini ditunjukkan dengan Rand Index (RI) tertinggi. Kemudian, penelitian Septianingsih, 2022 melakukan pemetaan kabupaten/kota di provinsi Jawa Timur. Pemetaan dilakukan berdasarkan tingkat kasus penyakit dengan metode Agglomerative Hierarchical *clustering* dengan empat metode yaitu *Average Linkage, Complete Linkage, Single Linkage,* dan *Ward Linkage*. Hasil optimal yang diperoleh dengan korelasi *cophenetic* yaitu metode Average Linkage. Berdasarkan uji validasi *cluster* menggunakan uji indeks validitas *connectivity*, indeks *dunn* dan *silhouette* jumlah *cluster* optimal terbentuk yaitu 4 *cluster*.

Adapun kontribusi dari penelitian ini yang belum ada pada penelitian sebelumnya adalah memeriksa tingkat pencemaran serta sumber utama pencemaran lingkungan hidup pada 34 provinsi di Indonesia sehingga dapat memberikan gambaran keadaan lingkungan provinsi tersebut guna untuk mewujudkan misi ke-5 Indonesia Emas 2045 yakni menuju *net zero emision* dengan cara mengelompokkan provinsi tercemar di Indonesia dengan metode K-Medoids, K-Means, Fuzzy C-Means serta *Complete Linkage* dan *Ward's Linkage* untuk *Aglomerative Hierarchy*. Dengan mengetahui tingkat pencemaran pada

provinsi tersebut dapat memberikan rekomendasi kepada pemerintah selaku pemangku kebijakan.

METODE

Data dan Sumber Data

Dalam penelitian ini, data yang digunakan adalah jumlah desa dengan jenis pencemaran lingkungan hidup menurut 34 provinsi di Indonesia pada tahun 2021. Selain itu, digunakan juga jumlah desa tercemar menurut sumber utama pencemaran di tiap jenis pencemaran lingkungan hidup(tanah, air, dan udara) menurut provinsi di Indonesia tahun 2021. Beberapa data tersebut diperoleh dari publikasi BPS [Statistik Lingkungan Hidup Indonesia 2023]. Rincian variabelnya dapat dilihat pada tabel 1.

Tabel 1. Keterangan variabel

Variabel	Keterangan	Tipe
Provinsi	Nama Provinsi	Karakter
Tanah	Jumlah desa dengan polusi Tanah	Numerik
Air	Jumlah desa dengan polusi Air	Numerik
Udara	Jumlah desa dengan polusi Udara	Numerik
Tidak Polusi	Jumlah desa yang Tidak Polusi	Numerik
T.RT	Jumlah desa dengan Tanah tercemar yang bersumber utama dari Rumah Tangga	Numerik
T.P	Jumlah desa dengan Tanah tercemar yang bersumber utama dari Pabrik	Numerik
T.L	Jumlah desa dengan Tanah tercemar yang bersumber utama dari Lainnya	Numerik
A.RT	Jumlah desa dengan Air tercemar yang bersumber utama dari Rumah Tangga	Numerik
A.P	Jumlah desa dengan Air tercemar yang bersumber utama dari Pabrik	Numerik
A.L	Jumlah desa dengan Air tercemar yang bersumber utama dari Lainnya	Numerik
U.RT	Jumlah desa dengan Udara tercemar yang bersumber utama dari Rumah Tangga	Numerik
U.P	Jumlah desa dengan Udara tercemar yang bersumber utama dari Pabrik	Numerik
U.L	Jumlah desa dengan Udara tercemar yang bersumber utama dari Lainnya	Numerik

Preprocessing data

Sebelum dilakukan analisis lebih lanjut, data jumlah desa akan diubah menjadi proporsi desa di setiap provinsi, dengan cara membagi jumlah desa pada masing-masing variabel dengan total jumlah desa di setiap provinsinya. Hal ini dilakukan agar tinggi rendahnya nilai pada data bisa berfokus pada seberapa tercemar wilayah tersebut tanpa terpengaruh dengan perbedaan jumlah desa antar provinsi yang mungkin saja bisa berbeda di tiap wilayahnya.

K-Medoids

K-Medoids merupakan salah satu pembelajaran mesin tanpa pengawasan yang digunakan untuk menangani masalah pengelompokan. K-Medoids juga dikenal sebagai partisi di sekitar medoids (PAM). Metode K-Medoids lebih tahan terhadap pencilan (*outlier*) dan *noise* dibandingkan dengan K-Means. Perbedaan utama antara K-Medoids dan K-Means terletak pada penentuan pusat *cluster*-nya. K-Means menggunakan rata-rata, sedangkan K-Medoids menggunakan median (Medoids). Hal ini membuat K-Medoids lebih tahan terhadap outlier (Ikhsanudin & Wijayanto, 2024). Meskipun waktu komputasi algoritma K-Medoids lebih lama dibandingkan dengan K-Means karena terdapat perbedaan kompleksitas dalam komputasinya, tidak ada perbedaan yang jelas antara waktu komputasi model dan waktu komputasi K-Means(Shang, Yu, & Xie, 2022).

K-Means

K-Means memiliki kelebihan dalam proses penghitungan yang cukup efektif dan efisien. Dengan menghitung jarak sebuah karakteristik objek dari titik pusat sebuah kelompok dan mengulanginya, memberikan K-Means kecepatan dalam melakukan pekerjaan yang diberikan meski dalam dataset yang cukup besar. Penggunaan jarak sebagai pengelompokan, membuat K-Means memiliki hasil yang terukur dan relatif lebih mudah untuk mengevaluasi seberapa baik *cluster* yang didapat dengan menghitung SSE. Namun, K-Means memiliki kekurangan dalam pengelompokannya yang kaku. Hal ini terjadi karena, pada metode K-Means sebuah objek hanya dikelompokkan dalam satu *cluster* saja. Dalam pengukuran jarak K-Means pada penelitian ini digunakan metode Euclidian (Anton & Rorres, 2018).

Fuzzy C-Means

Algoritma fuzzy *clustering* yang paling terkenal dan sering digunakan adalah Fuzzy C-Means. Fuzzy C-Means (FCM) salah satu metode *soft clustering*(Ikhsanudin & Wijayanto, 2024). FCM dapat mengatasi kekurangan metode K-Means *Clustering* (KCM)(Siringoringo & Jamaludin, 2019) yang mana sebuah objek data hanya di-*cluster* kan pada satu *cluster* saja sehingga bersifat kaku dan tegas. Pada FCM suatu objek data dapat di-*cluster* kan lebih dari satu *cluster* berdasarkan nilai keanggotaan data tersebut(Ikhsanudin & Wijayanto, 2024). Nilai keanggotaan bernilai antara 0 hingga 1, semakin mendekati satu nilai keanggotaannya, maka semakin mirip data tersebut dengan suatu kelompok.

Hierarki *Clustering*

Salah satu algoritma dalam hierarki adalah *Agglomerative* yaitu secara agregasi (Vijaya, Sharma, & Batra, 2019). Cara kerja algoritma ini dengan membandingkan data maupun *cluster* kemudian mencari kesamaan atau perbedaan. Sehingga data atau *cluster* yang paling mirip digabung membentuk *cluster* baru. Hal ini terus berulang hingga semua data atau *cluster* membentuk satu *cluster* yang mencakup seluruh data. Untuk menentukan metode *linkage* diukur dengan koefisien *cophenetic* terbesar.

Metode *Agglomerative* yang digunakan dalam penelitian ini yaitu metode *Ward Linkage* dan *Complete Linkage*. Dalam metode *Ward Linkage*, sebelum menggabungkan *cluster* terdapat pertimbangan agar meminimalkan informasi yang hilang. Hal ini dapat dilihat dengan menghitung *Sum Of Squares* (SSE), semakin kecil nilai SSE maka semakin kecil informasi yang hilang. Sedangkan *Complete Linkage* menggabungkan objek atau *cluster* berdasarkan jarak maksimum atau jarak terjauh antara dua objek atau cluster.

Evaluasi model

Internal

Validitas internal merupakan metode untuk mengevaluasi kualitas dan kecocokan hasil pengelompokan data menggunakan informasi yang terkandung dalam data itu sendiri (Brock, Pihur, Datta, & Datta, 2008). Terdapat beberapa metrik yang umum digunakan dalam validasi internal, antara lain:

1. **Indeks *Connectivity***

   Nilai indeks ini berkisar antara 0 hingga tak terhingga (inf). Semakin rendah nilainya, semakin baik hasil pengelompokan yang diperoleh. Indeks *Connectivity* digunakan untuk mengukur seberapa terhubung observasi dalam satu kelompok. Nilai yang rendah menunjukkan observasi dalam kelompok yang sama memiliki hubungan yang kuat satu sama lain, sementara observasi antar kelompok memiliki hubungan yang lemah.

2. **Indeks *Dunn***

   Metrik ini membandingkan jarak terdekat antara dua observasi dari kelompok yang berbeda dengan jarak terjauh dalam masing-masing kelompok. Semakin besar nilainya, semakin baik hasil pengelompokan. Indeks *Dunn* digunakan untuk mengevaluasi seberapa terpisah kelompok tersebut satu sama lain dan seberapa kompak observasi dalam masing-masing kelompok. Nilai yang tinggi mengindikasikan bahwa kelompok-kelompok tersebut terpisah dengan baik dan observasi dalam setiap kelompok sangat mirip satu sama lain.

3. **Indeks *Silhouette***

   Metrik ini mengukur tingkat kepercayaan terhadap hasil pengelompokan dengan nilai antara -1 hingga 1. Seperti halnya Indeks *Dunn*, semakin besar nilainya, semakin baik hasil pengelompokan. Indeks *Silhouette* mempertimbangkan seberapa mirip suatu observasi dengan kelompoknya sendiri dibandingkan dengan kelompok lainnya.

Stabilitas

Validitas stabilitas merupakan ukuran untuk membandingkan hasil *clustering* dengan data lengkap dan pengelompokan dengan penghapusan satu kolom pada satu waktu (Brock et al., 2008). Untuk menguji stabilitas dapat dengan empat jenis ukuran yaitu sebagai berikut.

1. ***Average Proportion of Non-Overlap* (APN)**

   APN berfungsi untuk mengukur proporsi rata-rata observasi yang tidak ditempatkan dalam *cluster* yang sama saat melakukan pengelompokan. Nilai ini berkisar pada interval [0,1] dimana semakin mendekati nol maka semakin konsisten.

2. ***Average Distance* (AD)**

   AD berfungsi untuk menghitung jarak rata-rata antara observasi pada *cluster* yang sama dengan rata-rata *cluster*-nya. Semakin kecil nilai AD maka semakin bagus hasil yang diperoleh.

3. ***Average Distance Between Means* (ADM)**

   ADM berfungsi untuk menghitung jarak rata-rata *cluster* untuk observasi yang ada pada *cluster* yang sama dengan mengelompokkan berdasarkan data lengkap dan mengelompokkan berdasarkan data dengan satu kolom dihapus. Semakin kecil nilai ADM maka semakin bagus hasil yang diperoleh.

4. *Figure of Merit* (FOM)

FOM berfungsi untuk mengukur rata-rata varians intra-*cluster* dari observasi di kolom yang dihapus, di mana pengelompokan didasarkan pada sampel yang tersisa (tidak terhapuskan). Ini memperkirakan kesalahan rata-rata menggunakan prediksi berdasarkan rata-rata *cluster*. Semakin kecil nilai FOM maka semakin bagus hasil yang diperoleh

HASIL DAN PEMBAHASAN

Eksplorasi data

Sebelum dilakukan konversi ke proporsi, dengan menjumlahkan desa yang terkena pencemaran di Indonesia menurut jenisnya dapat juga diketahui nilai proporsinya terhadap total desa di Indonesia. Pencemaran air merupakan jenis pencemaran tertinggi dengan proporsi total di Indonesia sebesar 12.70%. Disusul dengan udara sebesar 6.71% dan juga tanah sebesar 1.78%.

Setelah dilakukan konversi proporsi, dapat juga dilakukan eksplorasi dengan melihat minimum, maksimum, dan rata-rata proporsi pada setiap variabel. Tabel 2 merupakan tabel hasil eksplorasi setelah dilakukan konversi ke proporsi.

Tabel 2. Statistik Deskriptif

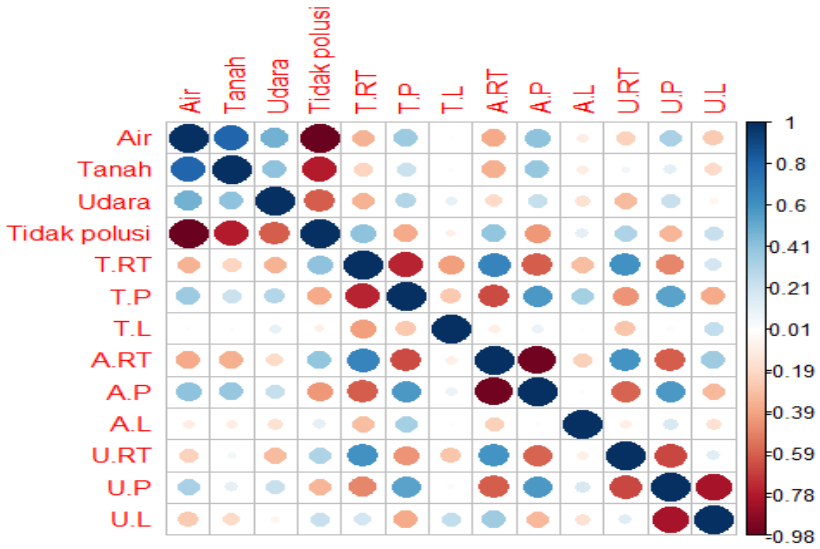
Variabel	Minimum	Maksimum	Rata-rata
Tanah	0.3531	7.9315	2.0516
Air	1.964	38.706	**14.115**
Udara	0.8056	15.7303	6.5384
Tidak Polusi	58.76	97.08	81.99
T.RT	0	92.63	41.65
T.P	4.211	100	42.707
T.L	0	62.857	15.642
A.RT	30.4	91.44	**59.27**
A.P	8.562	69.6	39.926
A.L	0	25.8065	0.8084
U.RT	2.247	66.667	13.56
U.P	8	88.73	**56.72**
U.L	5.238	70.686	29.717

Didapatkan nilai rata-rata proporsi variabel terbesar yakni “Tidak Polusi” dengan nilai 81.99% artinya rata-rata proporsi desa di tiap provinsi di Indonesia yang tidak tercemar nilainya cukup besar, namun hal tersebut tidak menunjukkan bahwa provinsi di Indonesia bebas dari pencemaran. Nilai rata-rata terkecil adalah variabel “A.L” dengan nilai 0.8084% artinya rata-rata jumlah desa di tiap provinsi di Indonesia sumber utama pencemaran air selain pabrik dan rumah tangganya kecil.

Selain itu, dapat dilihat juga rata-rata proporsi desa tercemar di tiap provinsi berdasarkan jenis pencemaran, rata-rata proporsi desa dengan pencemaran air merupakan yang tertinggi dibandingkan dengan udara dan tanah, yaitu sebesar 14.11%. Bukan hanya itu, dari setiap jenis pencemaran dapat juga dilihat berdasarkan sumber utamanya. Pada pencemaran air, sumber utama rata-rata proporsi desa tercemar air di tiap provinsi terbesar yaitu berasal dari rumah tangga dengan nilai sebesar 59.27%. Rata-rata proporsi desa tercemar tanah tiap provinsi di Indonesia menurut sumber utama pencemarannya

yang paling tinggi yaitu berasal dari pabrik dengan proporsi 42.71%. Sedangkan pada jenis pencemaran udara, rata-rata proporsi desa tercemar tanah tiap provinsi di Indonesia menurut sumber utama pencemarannya yang paling tinggi yaitu dari pabrik juga dengan nilai 56.72%.

Setelah dilihat rata-ratanya di setiap variabel, dapat juga dilihat hubungan antara variabel satu dengan yang lainnya. Gambar 1 merupakan visualisasi plot korelasi antar variabel.

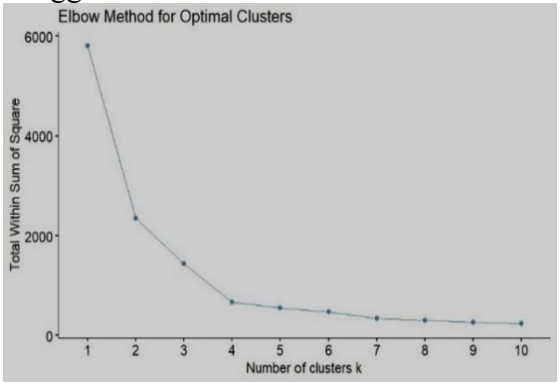


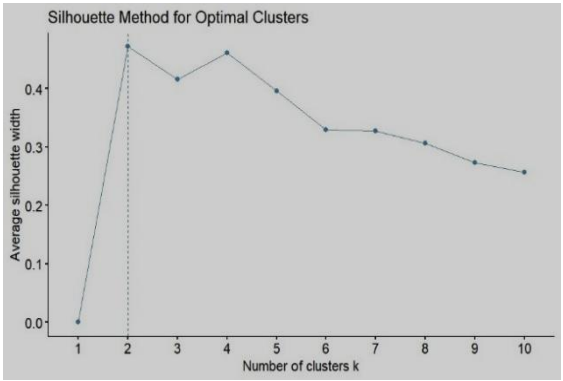
Gambar 1. Korelasi antar variabel

Dapat dilihat, beberapa variabel memiliki hubungan yang sangat kuat. Pada variabel "Tidak Polusi" korelasinya kuat dan arahnya berbanding terbalik. Hal ini terjadi karena pengubahan data menjadi proporsi dan juga memasukkan variabel tidak polusi yang menjadi salah satu pembentuk/sisa dari variabel jenis pencemaran yang berbentuk proporsi juga. Selain itu, beberapa variabel seperti "Air" dan "Tanah" yang memiliki hubungan searah yang cukup kuat sepertinya tidak memiliki penjelasan secara khusus. Meskipun beberapa variabel memiliki korelasi yang tinggi, pada penelitian ini variabel tersebut masih dimasukkan untuk menambah informasi dalam proses *clustering*.

Hasil *Clustering*

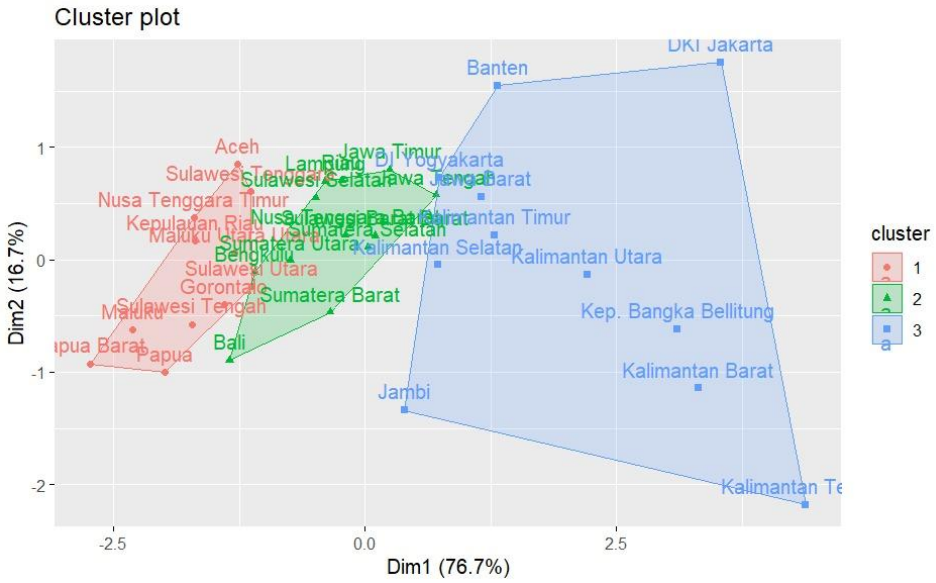
Pada pengelompokan K-Medoids, langkah pertama yang dilakukan adalah menentukan jumlah *cluster* yang akan dibentuk. Pembentukan *cluster* ini dilakukan menggunakan metode *silhouette* dan *elbow*.





Gambar 2. Grafik *elbow* dan *silhouette* pada metode K-Medoids

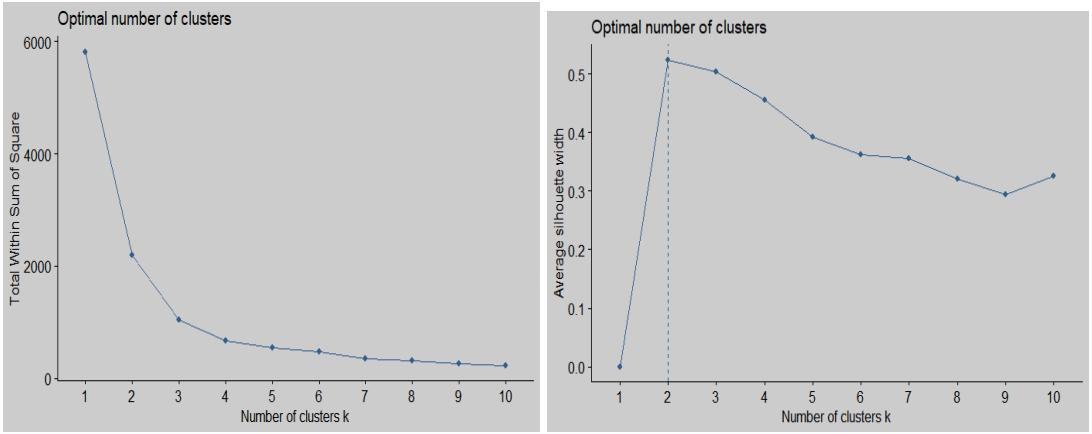
Berdasarkan Gambar 2 metode *silhouette*, didapatkan hasil bahwa terdapat dua *cluster* optimum yang terbentuk. Sedangkan untuk metode *elbow*, dapat dilihat bahwa penurunan nilai total dari WSS mulai berkurang ketika jumlah *cluster*-nya tiga sehingga jumlah optimum *cluster* adalah tiga untuk metode *elbow*.



Gambar 3. Hasil visualisasi *clustering* menggunakan Metode K-Medoids

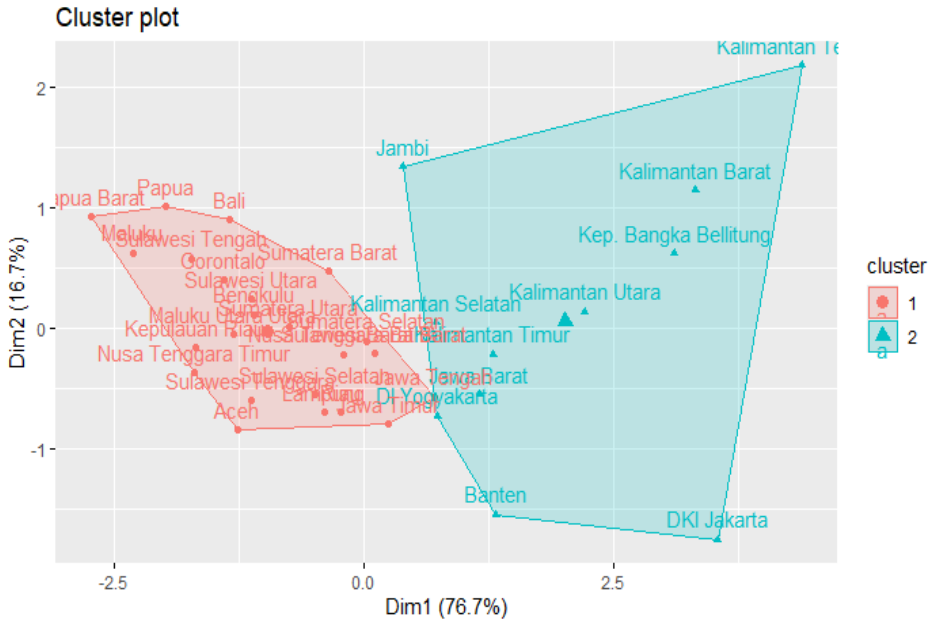
Setelah mengecek banyaknya *cluster* yang terbentuk menggunakan metode *elbow* dan *silhouette* didapatkan 3 *cluster* yang optimal. Setelah itu dilakukan visualisasi *cluster* seperti gambar 3 diatas. Hasilnya *cluster* pertama terdiri dari 11 provinsi, *cluster* dua terdiri dari 12 provinsi dan *cluster* tiga terdapat 11 provinsi.

Selanjutnya dilakukan langkah yang sama untuk membuat *cluster* menggunakan metode K-Means.



Gambar 4. Grafik *elbow* dan *silhouette* pada metode K-Means

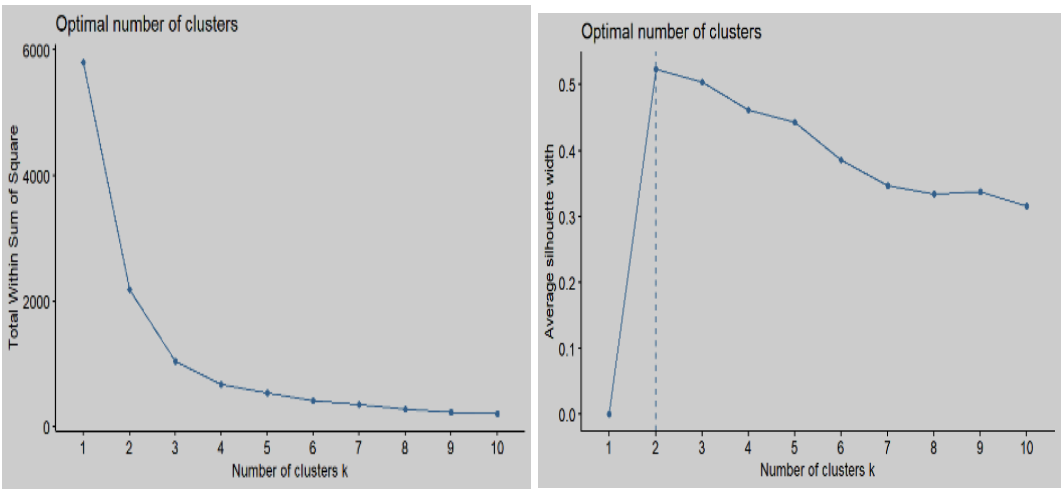
Pada Gambar 4 *elbow* dan *silhouette* yang berbasis K-Means, dapat dilihat bahwa *cluster* yang optimal pada data adalah 2 *cluster* . Pada *elbow*, ketika *cluster* nya berjumlah 3 perubahannya tidak terlalu signifikan. Sama seperti pada grafik *elbow*, grafik *silhouette* yang menunjukkan perubahan yang tidak terlalu signifikan pada jumlah *cluster* 3. Hal ini membuat *cluster* sebanyak 2 lebih disarankan/optimal untuk dipilih.



Gambar 5. Hasil visualisasi *clustering* menggunakan metode K-Means

Setelah dilakukannya pengecekan *elbow* dan *silhouette* dengan 2 *cluster* sebagai keputusannya, Dilakukanlah metode K-Means menggunakan library "Factoextra" pada Rstudio. Setelah itu, hasil pengelompokan divisualisasikan menjadi gambar di atas yang menunjukkan hasil dari pengelompokan dengan menggunakan metode K-Means. Dari visualisasi tersebut, dapat dilihat bahwa sebanyak 23 provinsi untuk *cluster* 1 dan 11 provinsi untuk *cluster* 2.

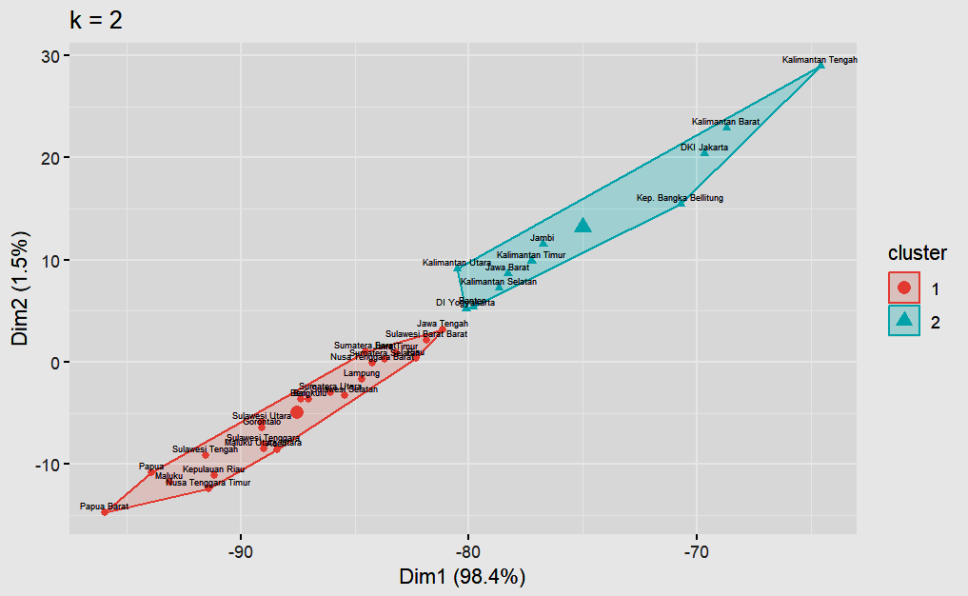
Penentuan jumlah *cluster* optimum juga dilakukan pada metode Fuzzy C-Means (FCM). Tahapannya dilakukan dengan menggunakan metode *elbow* dan *silhouette.*



Gambar 6. Grafik *elbow* dan *silhouette* pada metode Fuzzy C-Means

Pada Gambar 6, grafik *elbow* terlihat membentuk siku atau bend signifikan pada *cluster* 2. Sedangkan pada grafik *silhouette* terlihat titik optimumnya adalah 2 maka jumlah *cluster* optimumnya adalah 2. Selanjutnya dilakukan pemodelan FCM menggunakan fungsi fcm() pada R dengan k=2. Gambar 7 merupakan hasil visualisasi

clustering menggunakan metode FCM dengan *cluster* 1 terdiri dari 23 provinsi dan *cluster* 2 terdiri dari 11 provinsi.



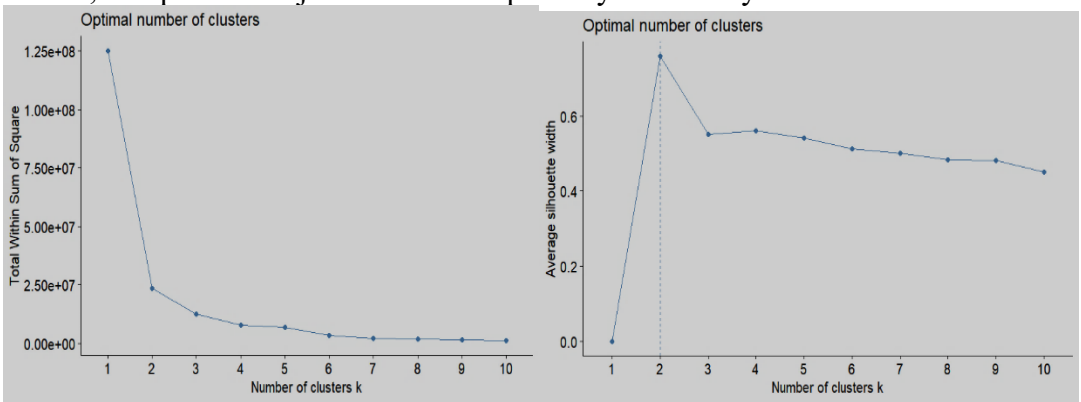
Gambar 7. Hasil visualisasi *clustering* menggunakan metode Fuzzy C-Means

Pada metode hierarki, hal yang pertama dilakukan adalah menentukan metode *linkage* terbaik berdasarkan koefisien korelasi *cophenetic*. Nilai korelasi dari kedua *cluster* dapat dilihat pada tabel 3.

Tabel 3. Perbandingan nilai korelasi *Cophenetic*

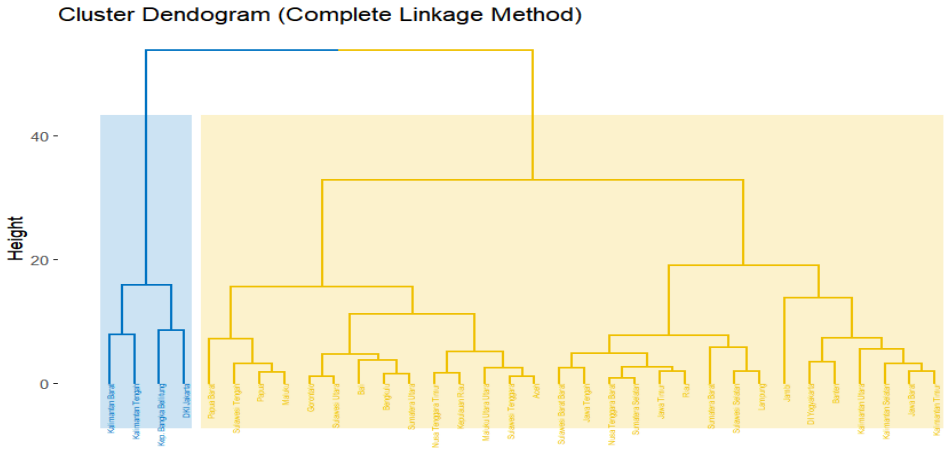
Metode	Korelasi *Cophenetic*
Complete Linkage	0.6423208
Ward's Linkage	0.5227623

Nilai korelasi *Cophenetic* tertinggi diperoleh dengan metode *complete linkage*. Sehingga metode hierarki yang digunakan untuk pembuatan *cluster* dapat dilakukan dengan metode *complete linkage*. Untuk menentukan jumlah *cluster* yang baik maka dapat dilihat berdasarkan grafik *elbow* dan grafik *silhouette*. Berdasarkan kedua grafik tersebut, didapat bahwa jumlah *cluster* optimal yaitu sebanyak 2 *cluster*.



Gambar 8. Grafik *elbow* dan *silhouette* pada metode hierarki dengan *Complete Linkage*

Hasil *clustering* hierarki metode *complete linkage* dengan jumlah *cluster* 2 dapat dilihat pada visualisasi *dendogram* pada gambar 9. Dari visualisasi tersebut, dapat dilihat bahwa sebanyak 4 provinsi untuk *cluster* 1 dan 30 provinsi untuk *cluster* 2.



Gambar 9. Dendogram visualisasi *clustering* menggunakan metode *Complete Linkage*

Validasi

Setelah masing-masing metode ditentukan jumlah *cluster* terbaik yang digunakan. Maka Selanjutnya membandingkan validasi tiap metode untuk memilih yang terbaik dalam membuat *cluster*.

Tabel 4. Uji validasi pada jumlah desa berdasarkan jenis pencemaran menurut provinsi

Metode	*Cluster*	Internal			Stabilitas			
		IC	ID	IS	APN	AD	ADM	FOM
K-Means	2	6.593	0.086	0.52	0.038	9.695	0.951	3.829
K-Medoids	3	10.884	0.086	0.41	**0.024**	**7.052**	**0.321**	**3.220**
Fuzzy C-Means	2	9.940	0.025	0.45	0.05	9.802	1.129	3.966
Complete Linkage	2	**5.060**	**0.282**	**0.57**	0.028	10.929	0.7185	4.154

Dari hasil uji validasi, terlihat bahwa metode K-Medoids unggul dalam empat indikator validasi stabilitas. Sedangkan metode *complete linkage* unggul dalam tiga indikator validasi internal. Sehingga pada penelitian ini diambil metode terbaik yang digunakan yaitu K-Medoids dengan 3 *cluster*. Dengan menggunakan metode K-Medoid k=3 hasil *cluster*-nya dapat dilihat pada tabel 5.

Tabel 5. Hasil *Cluster* dengan metode K-Medoid dengan K=3

Cluster	**Provinsi**
Cluster 1	Papua Barat, Papua, Maluku, Nusa Tenggara Timur, Sulawesi Tengah, Kepulauan Riau, Maluku Utara, Sulawesi Tenggara, Gorontalo, Sulawesi Utara, Aceh

Cluster 2	Bali, Bengkulu, Sumatera Utara, Sulawesi Selatan, Lampung, Sumatera Barat, Nusa Tenggara Barat, Sumatera Selatan, Jawa Timur, Riau, Sulawesi Barat, Jawa Tengah.
Cluster 3	DI Yogyakarta, Kalimantan Utara, Banten, Kalimantan Selatan, Jawa Barat, Kalimantan Timur, Jambi, Kep. Bangka Belitung, DKI Jakarta, Kalimantan Barat, Kalimantan Tengah

Profiling* Hasil Analisis *Cluster

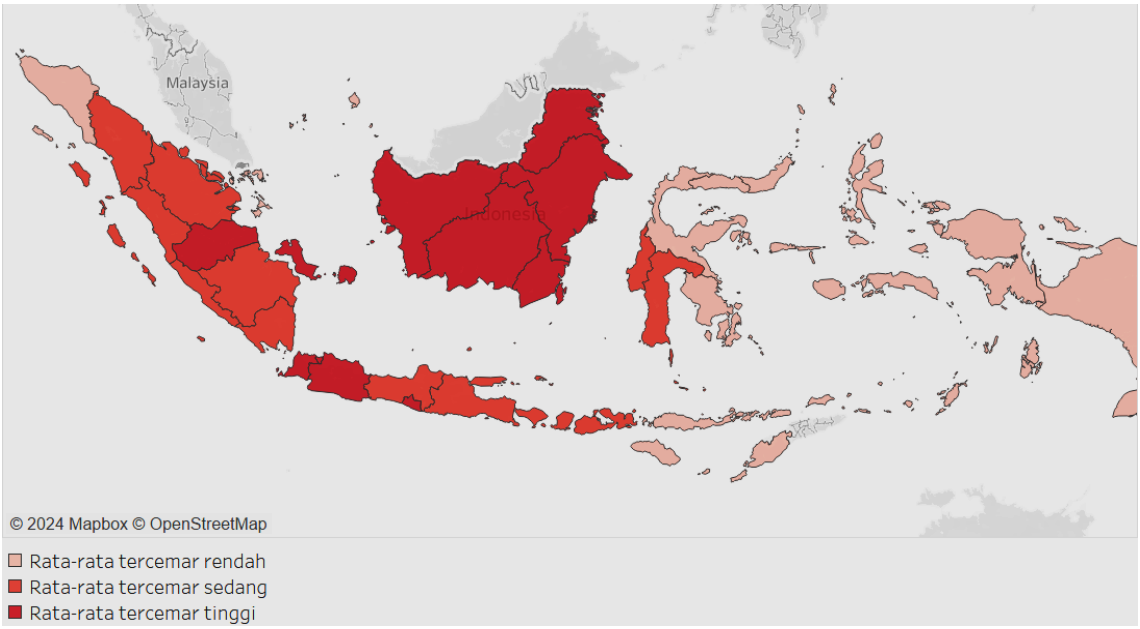
Metode terbaik yang terpilih adalah metode K-Medoids yang menghasilkan 3 cluster. Setiap *cluster*-nya dihitung rata-rata di setiap variabelnya untuk mendapatkan karakteristik tiap *cluster*. Karakteristik tiap *cluster*-nya dapat dilihat pada tabel 6.

Tabel 6. Karakteristik Hasil *Clustering* K-Medoid dengan K=3

Cluster	**Air**	**Tanah**	**Udara**
Cluster 1	5.297	0.997	4.203
Cluster 2	12.810	1.367	6.620
Cluster 3	24.357	3.852	8.783

Dari tabel 6 terlihat rata-rata *cluster* 3 paling besar di semua variabel dan *cluster* 1 nilai rata-ratanya paling kecil di semua variabel, sedangkan *cluster* 2 nilainya antara *cluster* 1 dan 2. Dengan demikian proporsi desa di provinsi yang masuk *cluster* 1 cenderung memiliki tingkat pencemaran lingkungan hidup yang rendah, proporsi desa di provinsi yang masuk *cluster* 2 cenderung memiliki tingkat pencemaran lingkungan hidup yang sedang, dan proporsi desa di provinsi yang masuk *cluster* 3 cenderung memiliki tingkat pencemaran lingkungan hidup yang tinggi.

Pada gambar 10 menggambarkan hasil pengelompokan provinsi dengan metode K-Medoids dengan k=3. Warna merah yang semakin gelap pada gambar merepresentasikan tingkat pencemaran yang semakin tinggi, dan warna terang sebaliknya.



Gambar 10. Pemetaan *cluster* provinsi metode K-Medoids

Dilakukan analisis lebih lanjut untuk mengetahui sumber utama pencemaran lingkungan hidup menggunakan data tambahan yakni data jumlah desa tercemar menurut sumber utama pencemaran di tiap jenis pencemaran lingkungan hidup (tanah, air, dan udara) menurut provinsi di Indonesia tahun 2021. Dilakukan perhitungan rata-rata tiap *cluster*-nya untuk melihat sumber utama pencemaran. Hasil perhitungan dapat dilihat pada tabel 7.

Tabel 7. Analisis *clustering* K-Medoid dengan K=3

Kategori	Variabel	*Cluster* 1	*Cluster* 2	*Cluster* 3
Tanah	T.RT	**45.628**	**51.025**	27.449
	T.P	39.168	33.689	**56.084**
	T.L	15.204	15.286	16.467
Air	A.RT	**63.261**	**64.858**	49.170
	A.P	34.393	35.095	**50.728**
	A.L	2.346	0.047	0.102
Udara	U.RT	20.581	11.323	8.980
	U.P	38.863	**65.880**	**64.591**
	U.L	**40.556**	22.797	26.429

Pada Tabel 7 dapat dilihat bahwa pada *cluster* 1 sumber utama pencemaran tanah dan air adalah rumah tangga karena nilai rata-ratanya paling tinggi yakni 45.628 dan 63.261. Sedangkan sumber utama pencemaran udara adalah bersumber selain dari pabrik dan rumah tangga dengan nilai rata-rata tertinggi yakni 40.556. Pada *cluster* 2 sumber utama pencemaran tanah dan air berasal dari rumah tangga dilihat dari nilai rata-rata nya paling tinggi yakni bernilai 51.025 dan 64. 858. Sedangkan sumber utama pencemaran udara adalah dari pabrik dengan nilai rata-rata 65.880. Pada *cluster* 3 sumber utama pencemaran tanah, air, dan berasal dari pabrik dengan nilai rata-rata 56.084, 50.728, dan 64.591. Dengan demikian dapat dikatakan bahwa pada provinsi yang masuk dalam *cluster* 3 terdapat banyak pabrik yang regulasi pembuangan limbahnya kurang baik yang menyebabkan pencemaran di provinsi tersebut tinggi dengan ditunjukkannya jumlah desa yang tercemar tinggi.

KESIMPULAN DAN SARAN

Kesimpulan

Berdasarkan analisis *cluster* yang telah dilakukan didapatkan metode terbaik untuk *cluster* provinsi di Indonesia berdasarkan data pencemaran lingkungan hidup adalah metode K-Medoids dengan jumlah 3 *cluster*. Proporsi desa di provinsi yang masuk *cluster* pertama cenderung memiliki rata-rata pencemaran lingkungan hidup yang rendah, proporsi desa di provinsi yang masuk *cluster* kedua cenderung memiliki rata-rata pencemaran lingkungan hidup yang sedang, dan proporsi desa di provinsi yang masuk *cluster* terakhir cenderung memiliki rata-rata pencemaran lingkungan hidup yang tinggi. Dari ketiga *cluster* tersebut sumber utama pencemaran secara rata-rata bersumber dari pabrik.

Desa di provinsi yang masuk ke *cluster* 3 membutuhkan perhatian khusus karena cenderung memiliki rata-rata pencemaran lingkungan hidup yang tinggi. Pemerintah perlu memberikan regulasi dan pengawasan yang ketat mengenai cara pengolahan limbah dari pabrik karena sumber utama pencemaran pada *cluster* 3 adalah pabrik. Selain itu, perlu adanya edukasi terhadap rumah tangga tentang pentingnya kesadaran untuk menjaga lingkungan. Hal ini dapat dilakukan seperti tidak membakar atau membuang sampah sembarangan, mengurangi limbah plastik, menggunakan transportasi umum agar mengurangi asap kendaraan, dan Upaya lainnya yang dapat meningkatkan kualitas lingkungan dan kesehatan masyarakat.

Saran

Adapun saran untuk penelitian selanjutnya yakni dapat menggunakan algoritma *clustering* yang berbeda serta dapat menggunakan lebih banyak variabel yang dapat merepresentasikan pencemaran lingkungan.

DAFTAR PUSTAKA

Anton, H., & Rorres, C. (2018). Elementary Linear Algebra. In *Analytical Biochemistry* (Vol. 11). Retrieved from http://link.springer.com/10.1007/978-3-319-59379-1%0Ahttp://dx.doi.org/10.1016/B978-0-12-420070-8.00002-7%0Ahttp://dx.doi.org/10.1016/j.ab.2015.03.024%0Ahttps://doi.org/10.1080/07352689.2018.1441103%0Ahttp://www.chile.bmw-motorrad.cl/sync/showroom/lam/es/

Bell, J. N. B. (1997). Pollution: Causes, effects and control. *Environmental Pollution*, *96*(2), 276–277. https://doi.org/10.1016/s0269-7491(97)83364-7

Brock, G., Pihur, V., Datta, S., & Datta, S. (2008). ClValid: An R package for cluster validation. *Journal of Statistical Software*, *25*(4), 1–22. https://doi.org/10.18637/jss.v025.i04

Harris, P. T., Westerveld, L., Nyberg, B., Maes, T., Macmillan-Lawler, M., & Appelquist, L. R. (2021). Exposure of coastal environments to river-sourced plastic pollution. *Science of the Total Environment*, *769*. https://doi.org/10.1016/j.scitotenv.2021.145222

Hartono, D., Dachlan, A. N., Hastuti, S. H., Kartiasih, F., Saputri, N. K., Kurniawan, R., … Shirakawa, H. (2023). The Impacts of Households on Carbon Dioxide Emissions in Indonesia. *Environmental Processes*, *10*(4), 1–20. https://doi.org/10.1007/s40710-023-00666-3

Herman, E., Zsido, K. E., & Fenyves, V. (2022). Cluster Analysis with K-Mean versus K-Medoid in Financial Performance Evaluation. *Applied Sciences (Switzerland)*, *12*(16). https://doi.org/10.3390/app12167985

Ikhsanudin, M. R., & Wijayanto, A. W. (2024). Perbandingan Pengelompokkan Provinsi di Indonesia Menurut Kualitas Lingkungan Hidup Menggunakan Metode Hierarki dan Partisi Comparing Province Clustering in Indonesia Based on Environmental Quality Using Hierarchical and Partition Methods. *Jurnal Sistem Dan Teknologi Informasi*, *12*(1), 155–163. https://doi.org/10.26418/justin.v12i1.71495

Kementerian Lingkungan Hidup. (1988). Keputusan Menteri Negara Kependudukan Dan Lingkungan Hidup Nomor: Kep-02/Menklh/I/1988 Tentang Pedoman Penetapan Baku Mutu Lingkungan. *Menteri Negara Kependudukan Dan Lingkungan Hidup*.

KILIÇ, Z. (2021). Water Pollution: Causes, Negative Effects and Prevention Methods. *İstanbul Sabahattin Zaim Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, *3*(2), 129–132. https://doi.org/10.47769/izufbed.862679

Lin, L., Yang, H., & Xu, X. (2022). Effects of Water Pollution on Human Health and Disease Heterogeneity: A Review. *Frontiers in Environmental Science*, *10*(June). https://doi.org/10.3389/fenvs.2022.880246

Matlis, G., Dimokas, N., & Karvelis, P. (2024). Unveiling University Groupings: A Clustering Analysis for Academic Rankings. *Data*, 1–41. https://doi.org/https://doi.org/10.3390/data9050067

Mittal, R., & Mittal, C. G. (2013). Impact of Population Explosion on Environment. *WeSchool "Knowledge Builder" - The National Journal*, *1*(1), 1–5. Retrieved from http://weschool.rtmonline.in

Petrisor, I. (n.d.). About Environmental Pollution | Environmental Pollution Centers. Retrieved May 8, 2024, from https://www.environmentalpollutioncenters.org/

Remilekun Adeuti, B. (2020). Analysis of Environmental Pollution in Developing Countries. *American Scientific Research Journal for Engineering*, *65*(1), 39–48. Retrieved from http://asrjetsjournal.org/

Rice, M., Balmes, J., Malhotra, A., Castner, J., Garcia, E., Hicks, A.,Sockrider, M. (2021). Outdoor air pollution and your health. *American Journal of Respiratory and Critical Care Medicine*, *204*(7), P13–P14. https://doi.org/10.1164/rccm.2046P13

Septianingsih, A. (2022). Pemetaan Kabupaten Kota Di Provinsi Jawa Timur Berdasarkan Tingkat Kasus Penyakit Menggunakan Pendekatan Agglomeratif Hierarchical Clustering. *Jurnal Lebesgue : Jurnal Ilmiah Pendidikan Matematika, Matematika Dan Statistika*, *3*(2), 367–386. https://doi.org/10.46306/lb.v3i2.139

Shang, Q., Yu, Y., & Xie, T. (2022). A Hybrid Method for Traffic State Classification Using K-Medoids Clustering and Self-Tuning Spectral Clustering. *Sustainability (Switzerland)*, *14*(17). https://doi.org/10.3390/su141711068

Siringoringo, R., & Jamaludin. (2019). Peningkatan Performa Cluster Fuzzy C-Means Pada Pengklasteran Sentimen Menggunakan Particle An Improved Fuzzy C-Means For Sentiment Clustering Based On Particle Swarm Optimization. *JTIIK - Jurnal Teknlogi Informasi Dan Ilmu Komputer*, *6*(4), 349–354. https://doi.org/10.25126/jtiik.2018561090

Sompotan, D. D., & Sinaga, J. (2022). Pencegahan Pencemaran Lingkungan. *SAINTEKES: Jurnal Sains, Teknologi Dan Kesehatan*, *1*(1), 6–13. https://doi.org/10.55681/saintekes.v1i1.2

Syafiyah, U., Puspitasari, D. P., Asrafi, I., Wicaksono, B., & Sirait, F. M. (2022). Analisis Perbandingan Hierarchical dan Non-Hierarchical Clustering Pada Data Indikator Ketenagakerjaan di Jawa Barat Tahun 2020. *Seminar Nasional Official Statistics*, *2022*(1), 803–812. https://doi.org/10.34123/semnasoffstat.v2022i1.1221

Ukaogo, P. O., Ewuzie, U., & Onwuka, C. V. (2020). Environmental pollution: Causes, effects, and the remedies. In *Microorganisms for Sustainable Environment and Health*. INC. https://doi.org/10.1016/B978-0-12-819001-2.00021-8

Vijaya, V., Sharma, S., & Batra, N. (2019). Comparative Study of Single Linkage, Complete Linkage, and Ward Method of Agglomerative Clustering. *Proceedings of the International Conference on Machine Learning, Big Data, Cloud and Parallel Computing: Trends, Prespectives and Prospects, COMITCon 2019*, 568–573. https://doi.org/10.1109/COMITCon.2019.8862232

WHO date boks. (2019). Polusi Udara Sebabkan 7 Juta Kematian per Tahun di Dunia. Retrieved May 8, 2024, from Data Boks website: https://databoks.katadata.co.id/datapublish/2019/06/07/polusi-udara-sebabkan-7-juta-kematian-per-tahun-di-dunia