ISSN 2527-3671 (*Online*)
ISSN 2087-8893 (*Print*)

Teknologi: Jurnal Ilmiah Sistem Informasi 14 (2) 2024 146-156

Tersedia *online* di www.journal.unipdu.ac.id
Unipdu

Halaman jurnal di www.journal.unipdu.ac.id/index.php/teknologi

PENERAPAN ALGORITMA K-MEANS UNTUK KLASTERISASI INDEKS STANDAR PENCEMARAN UDARA

Ihjal Mahendrasyah [a], Anita Diana [b], Rusdah [c], Deni Mahdiana [d]
[a,b,d] *Sistem Informasi, Universitas Budi Luhur, Jakarta, Indonesia.*
[c] *Magister Ilmu Komputer, Universitas Budi Luhur, Jakarta, Indonesia.*

email: [a]ihjalwota@gmail.com.

*Korespondensi

Dikirim 17 September 2023; **Direvisi** 11 September 2024; **Diterima** 1 November 2024; **Diterbitkan** 27 Desember 2024

Abstrak

Lingkungan merupakan satu ruang utuh yang berisi semua benda, daya, keadaan, serta makhluk hidup di dalamnya, termasuk juga manusia dan perilakunya yang mempengaruhi alam. Penting untuk menjaga lingkungan agar tidak timbul pencemaran yang dapat menciptakan kondisi yang tidak sehat. Pengelompokkan pencemaran lingkungan dapat mempermudah pemerintah dalam pertimbangan wilayah mana saja yang memerlukan atensi lebih dalam penegakkan perlindungan dan pengelolaan lingkungan hidup. Pengelompokan yang digunakan menggunakan algoritma K-Means Clustering yang dapat mengelompokkan data ke dalam kelompok yang sama dan data yang berbeda ke dalam kelompok yang berbeda. Klasterisasi dilakukan terhadap data indeks standar pencemaran udara provinsi DKI Jakarta berdasarkan parameter pencemaran udara. Sehingga dapat diketahui informasi mengenai kualitas udara terutama di wilayah provinsi DKI Jakarta. Data yang digunakan adalah data tahun 2021 dan diperoleh dari situs resmi Jakarta Open Data. Dataset berisi 365 hari pemantauan kualitas udara tahun 2021 serta parameter pencemaran udara seperti pm10, pm25, so2, co, 03, dan n02. Data yang telah diperoleh kemudian diolah dengan tools RapidMiner menggunakan Algoritma K-Means Clustering dalam 3 cluster, diketahui hasil klasterisasi yaitu kategori kualitas udara sehat pada cluster 0 terdiri dari 39 hari, kategori kualitas udara sedang pada cluster 1 yang terdiri dari 128 hari, dan kategori kualitas udara tidak sehat pada cluster 2 yang terdiri dari 198 hari. Sehingga dapat diketahui hasil klasterisasi dengan Algoritma K-Means terhadap kualitas udara di Provinsi DKI Jakarta tahun 2021 cenderung berada di kategori tidak sehat. Hasil klasterisasi ini diharapkan dapat menjadi masukan bagi pemerintah dalam upaya penanganan daerah yang mengalami pencemaran lingkungan.

Kata Kunci: pencemaran, lingkungan, udara, klasterisasi, k-means

Implementation of K-Means Algorithm for Air Pollution Standards Index Clustering

Abstract

The environment is whole space that contains all objects, power, conditions, and living things in it, including humans and their behavior that affects nature. It is important to protect the environment so that pollution does not arise which can create unhealthy conditions. Grouping environmental pollution can make it easier for the government to consider which areas require more attention in enforcing environmental protection and management. The grouping used uses the K-Means Clustering algorithm which can group data into the same group and different data into different groups. Clustering was carried out on the DKI Jakarta provincial air pollution standard index data based on air pollution parameters. So that information about air quality can be found, especially in the DKI Jakarta province. The data used is data for 2021 and is obtained from the official Jakarta Open Data website. The dataset contains 365 days of air quality monitoring in 2021 as well as air pollution parameters such as pm10, pm25, so2, co, 03, and n02. The data that has been obtained is then processed with the RapidMiner tool using the K-Means Clustering Algorithm in 3 clusters. unhealthy air quality in cluster 2 which consists of 198 days. So it can be seen that the results of clustering with the K-Means Algorithm on air quality in DKI Jakarta Province in 2021 tend to be in the unhealthy category. The results of this clustering are expected to be input for the government in efforts to deal with areas experiencing environmental pollution.

Keywords: : pollution, environment, air, clustering, k-means

Untuk mengutip artikel ini dengan *APA Style*:
Mahendrasyah. I, Diana. A, Rusdah, Mahdiana. D (2024). Penerapan Algoritma K-Means untuk Klasterisasi Indeks Standar Pencemaran Udara PENERAPAN. TEKNOLOGI: Jurnal Ilmiah Sistem Informasi, 14(2), 146-156: https://doi.org/10.26594/teknologi.v14i2.4088

1. Pendahuluan

Lingkungan ialah kesatuan ruang yang dimana memuat semua daya, benda, keadaan, dan juga makhluk hidup, termasuk manusia serta perilakunya, yang mempengaruhi kelangsungan kehidupan dan alam. Apa saja yang ada di sekeliling manusia yang mempengaruhi perkembangan kehidupan baik langsung maupun

ISSN 2502-3357 (*Online*)
Ihjal, M dkk./TEKNOLOGI 14 (2) 2024 146-156 ISSN 2503-0477 (*Print*)

tidak langsung juga adalah pengertian dari lingkungan lingkungan (Darmo Wihardjo & Rahmayanti, 2021). Penting akan menciptakan kesehatan, kedamaian serta kenyamanan agar tidak meninggalkan dampak pencemaran pada lingkungan. Dampak pencemaran lingkungan menciptakan kondisi yang tidak sehat, sehingga dapat mengganggu kenyamanan dan kesehatan makhluk hidup termasuk manusia (Khairunnisa et al., 2019).

Pertumbuhan kehidupan ekonomi serta urbanisasi yang cukup tinggi berpeluang tinggi dalam hal peningkatan pemakaian konsumsi energi, misalnya kebutuhan tungku-tungku industri, bahan bakar guna pembangkit tenaga listrik, dan juga transportasi. Pembakaran bahan bakar tersebut ialah sumber pencemar utama yang setiap harinya menimbulkan zat polutan menjadi pencemar udara. Akibatnya udara yang awalnya bersih menjadi tercemar dan bisa menimbulkan gangguan kesehatan serta dapat merusak lingkungan ekosistem. Kondisi ini memberikan dampak positif pada sektor perekonomian, namun juga memberikan dampak negatif berupa pencemaran udara (Abidin & Artauli Hasibuan, 2019).

Provinsi DKI Jakarta berperan sebagai kota yang paling padat penduduk, berdasarkan Badan Pusat Statistik DKI Jakarta pada tahun 2018 mengalami peningkatan pada produksi industri yang menciptakan emisi seperti pada industri tekstil, logam, kelistrikan dan kendaraan. Berdasarkan hal tersebut Dinas Lingkungan Hidup DKI Jakarta membangun stasiun pemantauan kualitas udara guna mengamati polusi udara seperti pada parameter NO2 (Nitrogen Dioksida), SO2 (Sulfur Dioksida), O3 (Ozon), CO (Karbon Monoksida) dan PM10 (Partikulat) (Imas Agista et al., 2020). Kemudian pada tahun 2019, Jakarta diketahui menduduki peringkat ketiga dari sepuluh kota yang berpolusi di Indonesia (Oktaviani & Hustinawati, 2021).

Berdasarkan kondisi polutan yang kian meningkat tersebut, analisa kualitas udara di DKI Jakarta dilakukan dengan membuat model klasterisasi terkait indeks standar pencemaran udara (ISPU) menerapkan penambangan data dengan metodologi penelitian *Cross-Industry Standard Process for Data Mining* (*CRISP-DM*) serta algoritma yang dipergunakan yaitu K-Means *Clustering* yang diharapkan dapat menjadi informasi kualitas udara di suatu wilayah. *Data mining* diartikan sebagai teknik untuk menemukan informasi bernilai tambah secara manual. *Data mining* telah digunakan untuk mengungkap hubungan antara data yang akan dikelompokkan pada satu atau lebih kelompok sehingga item dalam kelompok secara substansial mirip satu sama lain sebagai pendekatan praktis dan terarah untuk mengambil pola dan data (Mai Sarah Tarigan et al., 2022).

2. State of the Art

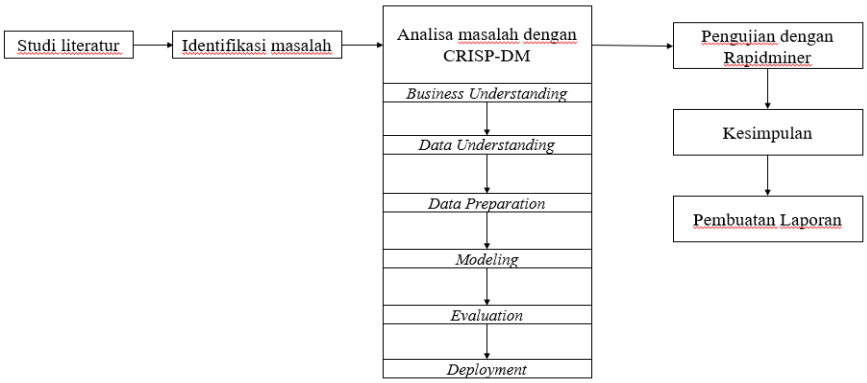
Pada penelitian serupa (Hermawan & Hasugian, 2022) membahas implementasi *data mining clustering* menggunakan K-Means dapat digunakan untuk mengetahui informasi permasalahan kesenjangan pembangunnan manusia di tiap Provinsi. Pada penelitian (Anjelita et al., 2020) dalam pengembangan data mining clustering membandingkan hasil terbaik dari dua algoritma yaitu K-Means dan K-Medoids, diketahui hasil terbaik pada algoritma K-Means Clustering. Penelitian (Sari et al., 2020) membahas penerapan K-Means dalam memberikan informasi tingkat kemiskinan di Provinsi Banten. Kemudian penelitian (Aritonang et al., 2020) menggunakan K-Means dalam mengelompokkan provinsi yang rawan dari kejahatan. Penelitian (Fadilah, 2022) membahas penerapan *clustering* menggunakan algoritma K-Means untuk mengetahui daerah yang rawan longsor di Provinsi Jawa Tengah.

Berdasarkan penelitian terdahulu, pada penelitian ini menerapkan metode *data mining clustering* dimana dalam penerapannya metode tersebut mengidentifikasi objek yang mempunyai kesamaan karakteristik khusus (Akbar Rismayadi et al., 2021). *Data mining* umumnya sebangun dengan proses mengekstraksi sejumlah besar data yang dikelompokkan menjadi data yang terorganisir dengan baik. Pengelompokan atau *Clustering* yang digunakan menggunakan algoritma K-Means. *Clustering* ialah metode dalam mengelompokkan data pada suatu *cluster* sehingga data yang mempunyai karakteristik sejenis dikelompokkan ke dalam satu *cluster* yang sama (Noor Permata Sari & Sukestiyarno, 2021).

Ihjal, M dkk./TEKNOLOGI 14 (2) 2024 146-156
ISSN 2502-3357 (*Online*)
ISSN 2503-0477 (*Print*)

3. Metode Penelitian

Tahapan pada penelitian ini mencakup dari sebagian tahapan serta menerapkan sebuah metodologi yaitu CRISP-DM (*Cross Industry Standart Process Model for Data Mining*) yang disajikan pada gambar 1, agar penelitian berjalan secara efektif dan terstruktur (Citra Mawani et al., 2023).



Gambar 1. Tahapan Penelitian

3.1. Studi Literatur

Tahapan pertama penelitian ini dimulai dengan pembelajaran yang dilakukan untuk memenuhi informasi dasar dan teori yang digunakan dalam penelitian. Studi literatur meliputi kegiatan pengumpulan data serta informasi yang dibutuhkan dan sumber data atau informasi diperoleh dari banyak media. Pada penelitian ini penulis mencari beberapa referensi dari perpustakaan Budi Luhur mengenai implementasi *data mining* dalam pengelompokkan wilayah. Informasi mengenai algoritma yang digunakan untuk pembagian atau pengelompokan seperti wilayah diketahui yaitu algoritma K-Means *Clustering*. Sumber yang sering dipakai untuk pencarian dataset pada penelitian sebelumnya adalah situs resmi Badan Pusat Statistik (BPS). Beberapa penelitian yang dibaca menggunakan pengujian *Davies Bouldin Index* (DBI) pada algoritma K-Means, namun pada penelitian ini ditambahkan pengujian dengan metode *Elbow*.

3.2. Identifikasi Masalah

Tahapan selanjutnya adalah mengidentifikasi masalah yang akan diteliti. Masalah yang berkaitan dengan topik penelitian diidentifikasi menggunakan informasi yang dikumpulkan dari studi literatur serta sumber lain yang berhubungan dengan topik penelitian ini. Identifikasi tersebut bertujuan dalam menegaskan batasan masalah supaya ruang lingkup penelitian tidak menyimpang dari tujuan.

3.3. Analisa Masalah dengan CRISP-DM

Tahapan berikutnya adalah analisa masalah dengan metodologi CRISP-DM. metodologi ini memberikan standar proses penambangan data yang akan digunakan untuk pemecahan masalah secara umum dari penelitian atau bisnis. Dalam tahapan CRISP-DM terdiri dari enam fase yang berurutan dan bersifat adaptif. Fase selanjutnya dalam urutan sangat bergantung pada keluaran dari fase sebelumnya (Ariwisanto Sianturi et al., 2019). Pada fase ini diterapkan metodologi CRISP-DM yang memiliki 6 tahap sebagai berikut:

a) *Business Understanding*

Tahap pertama yaitu memahami permasalahan yang menjadi bahan penelitian yaitu klasterisasi kualitas udara menurut indeks standar pencemaran udara. Dilakukannya analisa dikarenakan pencemaran udara semakin hari terus terjadi dan semakin parah terutama di daerah perkotaan. Maka hasil dari *clustering* ini akan menjadi informasi yang berguna bagi pemerintah dalam menangani masalah pencemaran udara di wilayah Provinsi DKI Jakarta.

b) *Data Understanding*

Dilakukan pemahaman data untuk bisa mengetahui data yang diperlukan untuk kebutuhan penelitian. Data diperoleh dari *website* resmi Jakarta *Open Data* berisi data indeks standar pencemaran udara Provinsi DKI Jakarta tahun 2021. Data yang diperoleh meliputi beberapa parameter pencemaran udara : pm10, pm25, so2, co, 03, dan n02. Data tersebut merupakan data pemantauan harian yang diakumulasi menjadi 365 hari selama 12 bulan.

c) *Data Preparation*

Tahap ini dijalankan pengolahan data supaya hasil penelitian yang diinginkan bisa tercapai meliputi seleksi variabel pada dataset, pembersihan data, hingga persiapan data awal. Data yang digunakan

ISSN 2502-3357 (*Online*)
Ihjal, M dkk./TEKNOLOGI 14 (2) 2024 146-156 ISSN 2503-0477 (*Print*)

berisi data pencemaran udara yang terdiri dari beberapa parameter, sehingga dilakukan seleksi variabel atau atribut. Selanjutnya dilakukan transformasi dengan metode normalisasi untuk menyamakan setiap rentang nilai atribut dengan skala tertentu. Normalisasi data pada penelitian ini menggunakan metode normalisasi *Min-Max* dengan rumus sebagai berikut (Azzahra Nasution et al., 2019).

$$x\_baru = (\frac{(x\_lama - minA)}{(maxA - minA)}) * (maxA\_baru - minA\_baru)) + minA\_baru$$

Diketahui pada rumus tersebut :
x_baru : nilai atribut baru hasil normalisasi
x_lama : nilai atribut sebelum normalisasi
minA : nilai atribut terkecil
maxA : nilai atribut terbesar
maxA_baru : rentang jarak terbesar
minA_baru : rentang jarak terkecil

d) Modelling

Pemilihan teknik pemodelan digunakan untuk mendapatkan nilai *cluster* yang optimal. Dalam *data mining*, terdapat beberapa model untuk menyelesaikan masalah yang sama, salah satunya Algoritma K-Means. K-Means *Clustering* bertujuan dalam membagi objek menjadi k *cluster* kemudian mengamati dimana tiap objek dalam *cluster* tersebut didapat melalui mean terdekat. Algoritma ini mengklasifikasikan observasi ke dalam kelompok k, dimana k adalah parameter *input*. Semua data tersebut kemudian ditempatkan pada setiap observasi dalam *cluster* berdasarkan seberapa dekat observasi tersebut dengan rata-rata *cluster*. Selanjutnya nilai rata-rata pada *cluster* dihitung secara berulang pada proses awal. Langkah-langkah dalam penerapan K-Means *Clustering* adalah sebagai berikut (Kamila et al., 2019):

1. Pilih jumlah *cluster* (k) yang diperlukan dalam dataset.
2. Menentukan nilai centroid pada tahap awal secara acak atau dapat diperoleh dari nilai maksimum pada *cluster* tingkat tinggi dan nilai minimum pada *cluster* tingkat rendah.
3. Tentukan jarak terdekat ke centroid untuk setiap data. Jarak Euclidean digunakan untuk menghitung jarak terpendek ke centroid.

$$D_e = \sqrt{(x_i - s_i)^2 + (y_i - t_i)^2}$$

Keterangan:
De : *euclidean distance*
I : banyaknya objek
(x,y) : koordinat objek
(s, t) : koordinat centroid

4. Hitung ulang *cluster* dengan anggota *cluster* baru. Rata-rata dari seluruh data digunakan sebagai pusat *cluster*. Proses hitung berhenti ketika nilai *cluster* tidak lagi berubah.

e) Evaluation

Tahap evaluasi akan dilakukannya pengujian dengan menggunakan metode *Elbow* pada *tools* RapidMiner untuk mengetahui jumlah *cluster* terbaik melalui pengamatan hasil perbandingan antara jumlah *cluster*, tujuan dari tahapan ini untuk menguji apakah perhitungan yang dilakukan sebelumnya telah sesuai dengan yang diharapkan.

f) Deployment

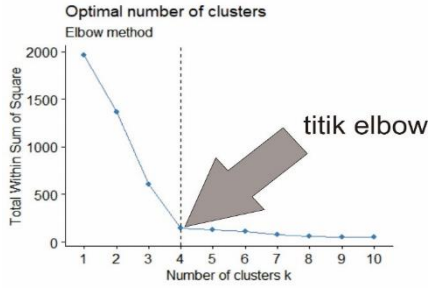
Langkah terakhir adalah memperluas pengetahuan atau informasi yang dihasilkan dari analisa yang telah dilakukan dalam bentuk laporan, sehingga hasil analisa dapat dengan mudah dipahami.

3.4. Pengujian dengan RapidMiner

Setelah dilakukannya tahap analisa masalah dengan metodologi CRISP-DM diketahui hasil *clustering* dari penerapan algoritma K-Means, setelah itu dilakukan tahapan pengujian dengan bantuan *tools* RapidMiner. Tahap pengujian dilakukan menggunakan metode *Elbow* dalam mengetahui jumlah *cluster* terbaik dengan mengetahui hasil perbandingan antara jumlah *cluster,* tujuan dari tahapan ini untuk menguji

Ihjal, M dkk./TEKNOLOGI 14 (2) 2024 146-156
ISSN 2502-3357 (*Online*)
ISSN 2503-0477 (*Print*)

apakah perhitungan yang dilakukan sebelumnya telah sesuai dengan yang diharapkan. Grafik Elbow dapat dilihat pada gambar 2.



Gambar 2. Grafik Metode *Elbow* (Ayu Indah Cahya Dewi & Ayu Kadek Pramita, 2019)

3.5. Kesimpulan

Pada tahap ini, mencakup kesimpulan dari keseluruhan proses yang telah dilakukan berdasarkan hasil analisis masalah hingga pengujian dengan RapidMiner.

3.6. Pembuatan Laporan

Tahap terakhir dari penelitian ini adalah membuat sebuah laporan dari hasil analisa dengan tujuan untuk memperluas pengetahuan atau informasi yang diperoleh, sehingga hasil penelitian dapat dengan mudah dipahami oleh pihak lain.

4. Hasil dan Pembahasan

Penelitian ini memakai data publik mengenai pencemaran udara berdasarkan nilai indeks standar pencemaran udara (ISPU) yang didapatkan dari *website* resmi Jakarta *Open Data* https://data.jakarta.go.id/dataset/indeks-standar-pencemaran-udara-isputahun-2021. Data yang diperoleh merupakan data tahun 2021 yang berekstensi .csv, data tersebut terdiri dari data harian selama 12 bulan pemantauan kualitas udara wilayah provinsi DKI Jakarta yang meliputi parameter pencemaran udara seperti pm10, pm25, so2, co, 03, dan n02.

4.1. Pra Pemrosesan Data

Tahapan yang dilakukan pada proses ini yaitu pengumpulan data ISPU, yang diperoleh berasal dari situs resmi Jakarta *Open Data* mengenai indeks standar pencemaran udara Provinsi DKI Jakarta selama 2021 menjadi satu dataset. Data ini meliputi beberapa parameter pencemaran udara seperti pm10, pm25, so2, co, 03, dan n02. Selanjutnya pada tahap ini mempersiapkan data mentah yang telah dikumpulkan untuk diubah ke dalam bentuk yang lebih mudah dipahami. Kemudian dilakukan *cleaning data* pada atribut yang tidak digunakan serta melakukan normalisasi untuk menjaga agar isi dalam dataset tetap konsisten.

4.2. Pengumpulan Data

Setelah pengumpulan data, kemudian proses akumulasi dilakukan pada data harian indeks standar pencemaran udara dari bulan Januari sampai dengan Desember menjadi satu dataset selama 2021.

Tabel 1. Data Indeks Standar Pencemaran Udara DKI Jakarta Tahun 2021

tanggal	pm10	pm25	so2	co	o3	no2	max	*critical*	*category*	*location*
01/01/2021	43	58	58	29	35	65	65	O3	SEDANG	DKI2
02/01/2021	58	86	86	38	64	80	86	PM25	SEDANG	DKI3
03/01/2021	64	93	93	25	62	86	93	PM25	SEDANG	DKI3
04/01/2021	50	49	67	24	31	77	77	O3	SEDANG	DKI2
05/01/2021	59	89	89	24	35	77	89	PM25	SEDANG	DKI3
06/01/2021	73	60	81	29	66	85	85	O3	SEDANG	DKI2
07/01/2021	36	50	52	22	55	72	72	O3	SEDANG	DKI2
08/01/2021	38	55	68	26	51	71	71	O3	SEDANG	DKI2
09/01/2021	60	67	77	34	42	80	80	O3	SEDANG	DKI2
10/01/2021	24	39	39	16	38	59	59	O3	SEDANG	DKI2
.....										
29/12/2021	61	98	54	15	37	29	98	PM25	SEDANG	DKI4

ISSN 2502-3357 (*Online*)
Ihjal, M dkk./TEKNOLOGI 14 (2) 2024 146-156 ISSN 2503-0477 (*Print*)

30/12/2021	60	102	53	17	38	44	102	PM25	TIDAK SEHAT	DKI4
31/12/2021	64	90	52	44	37	53	90	PM25	SEDANG	DKI4

Dengan rincian atribut data sebagai berikut:

a. Tanggal : Tanggal pengukuran kualitas udara
b. *location* : Lokasi pengukuran di stasiun
c. pm10 : Partikulat salah satu parameter yang diukur
d. pm25 : Partikulat salah satu parameter yang diukur
e. so2 : Sulfida salah satu parameter yang diukur
f. co : Carbon Monoksida salah satu parameter yang diukur
g. o3 : Ozon salah satu parameter yang diukur
h. no2 : NItrogen dioksida salah satu parameter yang diukur
i. *max* : Nilai ukur paling tinggi dari seluruh parameter
j. *critical* : Parameter yang hasil pengukurannya paling tinggi
k. *category* : Hasil perhitungan indeks standar pencemaran udara

4.3. Cleaning Data

Dilakukan pemilihan pada data yang tidak diinginkan untuk dihapus. Yaitu dengan menghapus atribut *max*, *critical*, *category*, dan *location*. Pada atribut *max* dan *critical* yang hanya merupakan informasi pengukuran paling tinggi dari tiap parameter. Atribut *category* yang merupakan label atau *class* juga tidak digunakan karena *clustering* merupakan *unsupervised learning* yang tidak membutuhkan atribut kelas, dan atribut *location* tidak digunakan karena tidak berdampak pada pengelompokan ISPU.

4.4. Transformasi Data

Transformasi dilakukan agar dapat mempermudah tahap analisa, maka dataset yang sebelumnya belum diolah akan dilakukan normalisasi. Proses normalisasi dilakukan dengan menggunakan rumus (1), melibatkan transformasi yaitu mengubah nilai jarak data menjadi lebih kecil misalnya ke dalam nilai 0 dan 1 (Farmana Putra et al., 2023). Sehingga menjadi format yang memungkinkan pemrosesan data yang efisien. Diketahui dengan rumus *Min-Max* pada data parameter pm10 tanggal 1 Januari 2021 sebagai berikut.

$$x\_baru = \left(\frac{(43-19)}{(179-19)}\right) * (1 - 0)) + 0 = 0{,}150$$

Normalisasi dilakukan seterusnya sampai 365 *record* dari setiap atribut parameter. Hasil pra pemrosesan data bisa dilihat pada tabel 2.

Tabel 2. Data ISPU Setelah Pra Pemrosesan

tanggal	pm10	pm25	so2	co	o3	no2
01/01/2021	0,150	0,200	0,236	0,550	0,115	0,448
02/01/2021	0,244	0,393	0,551	0,775	0,336	0,568
03/01/2021	0,281	0,441	0,629	0,450	0,321	0,616
04/01/2021	0,194	0,138	0,337	0,425	0,084	0,544
05/01/2021	0,250	0,414	0,584	0,425	0,115	0,544
06/01/2021	0,338	0,214	0,494	0,550	0,351	0,608
07/01/2021	0,106	0,145	0,169	0,375	0,267	0,504
08/01/2021	0,119	0,179	0,348	0,475	0,237	0,496
09/01/2021	0,256	0,262	0,449	0,675	0,168	0,568
10/01/2021	0,031	0,069	0,022	0,225	0,137	0,400
.....						
29/12/2021	0,263	0,476	0,191	0,200	0,130	0,160
30/12/2021	0,256	0,503	0,180	0,250	0,137	0,280
31/12/2021	0,281	0,421	0,169	0,925	0,130	0,352

4.5. Pemodelan

Setelah dilakukan tahap pra pemrosesan data pada dataset ISPU setelah itu dilakukan pemodelan *clustering* dengan menerapkan algoritma k-means. Berdasarkan tahapan *data mining* dengan algoritma k-means, akan dilakukan tahapan sebagai berikut:

Ihjal, M dkk./TEKNOLOGI 14 (2) 2024 146-156
ISSN 2502-3357 (*Online*)
ISSN 2503-0477 (*Print*)

a. Menentukan jumlah *cluster* yang digunakan berdasarkan atribut pada dataset yaitu 3 *cluster*, *cluster* pertama (C0) dengan kategori sehat, *cluster* kedua (C1) dengan kategori sedang dan *cluster* ketiga (C2) dengan kategori tidak sehat. Kategori yang digunakan sebagai landasan, dapat dilihat pada gambar 3.

Rentang	Kategori	Penjelasan
1-50	Baik	Tingat mutu udara yang sangat baik, tidak memberikan efek negatif terhadap manusia, hewan dan tumbuhan
51-100	Sedang	Tingkat mutu udara masih dapat diterima pada kesehatan manusia, hewan dan tumbuhan.
101-200	Tidak Sehat	Tingkat mutu udara yang bersifat merugikan pada manusia, hewan dan tumbuhan.
201-300	Sangat Tidak Sehat	Tingkat mutu udara yang dapat meningkatkan resiko kesehatan pada sejumlah segmen populasi yang terpapar.
301+	Berbahaya	Tingkat mutu udara yang dapat merugikan kesehatan serius pada populasi dan perlu penanganan cepat.

Gambar 3. Kategori Kualitas Udara Menurut Nilai ISPU Berdasarkan Lampiran Keputusan Kepala Bapedal No. 107 th 1997

b. Menentukan centroid awal, tahap ini akan dipilih titik centroid secara acak dari dataset.

Tabel 3. *Cluster* Awal

Cluster	pm10	pm25	so2	co	o3	no2
C0	0	0,028	0,045	0	0,206	0,008
C1	0,213	0,297	0,045	0,15	0,16	0,128
C2	1	0,2	0,079	0,125	0,176	0,096

c. Menghitung iterasi pertama jarak data antara nilai centroid menggunakan euclidian distance. Diketahui jarak data dari C0 sampai dengan C3 pada data berikut. Dengan menggunakan rumus (2), maka
Jarak data 1 dengan centroid 1:

$$= \sqrt{(0{,}150-0)^2+(0{,}200-0{,}028)^2+(0{,}236-0{,}045)^2+(0{,}550-0)^2+(0{,}115-0{,}206)^2+(0{,}448-0{,}008)^2}$$

$$= 0{,}826$$

Jarak data 1 dengan centroid 2:

$$= \sqrt{(0{,}150-0{,}213)^2+(0{,}200-0{,}297)^2+(0{,}236-0{,}045)^2+(0{,}550-0{,}150)^2+(0{,}115-0{,}160)^2+(0{,}448-0{,}128)^2}$$

$$= 0{,}563$$

Jarak data 1 dengan centroid 3:

$$= \sqrt{(0{,}150-1)^2+(0{,}200-0{,}200)^2+(0{,}236-0{,}079)^2+(0{,}550-0{,}125)^2+(0{,}115-0{,}176)^2+(0{,}448-0{,}096)^2}$$

$$= 1{,}089$$

Dan seterusnya sampai dengan record ke-365 dari tiap atribut. Sehingga diketahui hasil perhitungan data nilai centroid dari setiap cluster yang jaraknya terpendek dari pusat cluster pada Tabel 4.

Tabel 4. Hasil Perhitungan

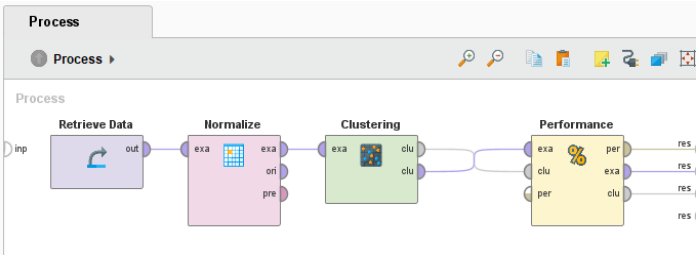
tanggal	C0	C1	C2	Jarak Terpendek
01/01/2021	0,826	0,563	1,089	0,563
02/01/2021	1,346	1,023	1,354	1,023
03/01/2021	1,268	0,933	1,272	0,933
04/01/2021	0,862	0,612	1,106	0,612
05/01/2021	1,119	0,792	1,179	0,792

ISSN 2502-3357 (*Online*)
ISSN 2503-0477 (*Print*)

d. Hasil perhitungan masing-masing data pada setiap cluster dapat ditentukan pada cluster mana data tersebut dikelompokkan. Data dikelompokkan berdasarkan nilai jarak terkecil. Data tersebut berhasil dipetakan ke dalam cluster sesuai dengan jarak yang diperoleh. Proses perhitungan dilanjutkan pada iterasi berikutnya sampai hasil cluster tidak berubah.

4.6. Pengujian

Pengujian dilakukan dengan melakukan pemodelan K-Means dengan memanfaatkan *tools* RapidMiner. RapidMiner adalah perangkat lunak yang dipakai dalam proses pengolahan data. RapidMiner bersifat *open source* yang dapat digunakan secara gratis untuk melakukan analisa penambangan data (Nindy Yuliarina & Hendry, 2022). Serta dilakukan uji validasi dengan mencari nilai *Davies Bouldin Index* dan grafik *Elbow*.



Gambar 4. Kategori Kualitas Udara

Berdasarkan hasil pengujian diketahui hasil dari proses klasterisasi. dataset yang telah diproses dibagi menjadi tiga kelompok yang terdiri dari *cluster* 0, *cluster* 1, dan *cluster* 2. Pembagian *cluster* didasarkan pada kedekatan setiap jarak cluster. Hasil Pemodelan *Clustering* dengan RapidMiner disajikan pada gambar 5, dan Nilai DBI pada gambar 6.

Cluster Model

```
Cluster 0: 39 items
Cluster 1: 128 items
Cluster 2: 198 items
Total number of items: 365
```

Gambar 5. Hasil Pemodelan Clustering dengan RapidMiner

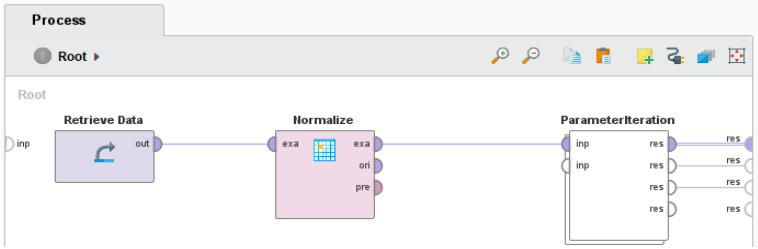
Davies Bouldin

```
Davies Bouldin: -1.172
```

Gambar 6. Nilai DBI dengan 3 *Cluster*

Pada gambar 4 dapat diketahui hasil pembagian kelompok data terhadap nilai *cluster*, yang dimana *cluster* 0 sebanyak 39 anggota, *cluster* 1 sebanyak 128 anggota, dan *cluster* 2 sebanyak 198 anggota. Hasil *clustering* yang telah diketahui selanjutnya dilakukan uji validasi.

Hasil dari jumlah *cluster* terbaik akan menjadi dasar dalam melakukan proses analisa *clustering*. Oleh karena itu perlu dilakukan uji validasi dengan *Davies Bouldin Index* (DBI) dan Metode *Elbow*. Metode DBI umumnya digunakan untuk mengukur performa model terbaik dari tiap *cluster*. Serta dilakukan pengujian dengan metode *Elbow* untuk memperoleh informasi tentang penentuan jumlah *cluster* terbaik dengan mengamati persentase hasil yang diperoleh saat membandingkan jumlah *cluster* yang membentuk siku (Virantika et al., 2022). Dilakukan percobaan pada RapidMiner dengan model *KmeansWithPlot* yang bisa dilihat pada gambar 7, serta hasil perbandingan tiap cluster pada gambar 8 berikut ini.



Gambar 7. Pengujian Menggunakan RapidMiner

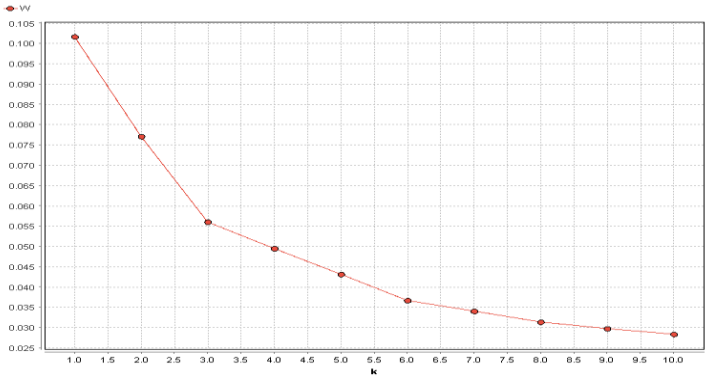
ISSN 2502-3357 (*Online*)
ISSN 2503-0477 (*Print*)

ProcessLog (10 rows, 3 columns)

k	DB	W
1	-∞	0.102
2	1.340	0.077
3	1.172	0.056
4	1.125	0.049
5	1.231	0.043
6	1.167	0.037
7	1.271	0.034
8	1.245	0.031
9	1.205	0.030
10	1.226	0.028

Gambar 8. Nilai DBI Tiap *Cluster*

Pada gambar 8 merupakan hasil proses algoritma K-Means dari 1 cluster sampai dengan 10 cluster. Dapat diketahui nilai Davies Bouldin Index yang paling kecil yaitu pada k=4 sebesar 1,125



Gambar 9. Grafik Hasil Metode *Elbow*

Dilanjutkan dengan pengujian Metode Elbow pada gambar 9 diketahui merupakan visualisasi grafik dari jumlah cluster optimal. Dapat diketahui bahwa cluster bernomor 2, 3, dan 4 menggambarkan sudut yang membentuk busur dan mengalami penurunan paling besar. Sehingga, dapat diketahui cluster yang optimal adalah 3 cluster.

5. Kesimpulan

Berdasarkan hasil analisa klasterisasi dengan menerapkan algoritma K-Means bertujuan dalam pengelompokkan data dengan mengoptimalkan kemiripan data dalam satu *cluster* serta meminimalkan kemiripan data antar *cluster*. Ukuran kesamaan atau kemiripan yang dipakai dalam *cluster* merupakan fungsi jarak. Sehingga dalam pengoptimalan pada kemiripan data diperoleh menurut jarak terpendek antara data terhadap titik centroid. Tujuan yang dicapai pada penelitian ini adalah mengetahui informasi kualitas udara di Provinsi DKI Jakarta berdasarkan parameter pencemaran udara, yang terdiri dari 5 atribut data yaitu pm10, so2, co, o3, dan no2. Pada proses klasterisasi dibantu dengan *tools* RapidMiner, Berdasarkan pengujian dengan Davies Bouldin Index (DBI) dan Metode *Elbow* dalam penentuan jumlah *cluster* terbaik diketahui berjumlah 3 *cluster*. *Cluster* ini dibagi menjadi kategori kualitas udara sehat pada *cluster* 0 yang terdiri dari 39 hari, kategori kualitas udara sedang pada *cluster* 1 yang terdiri dari 128 hari, dan kategori kualitas udara tidak sehat pada *cluster* 2 yang terdiri dari 198 hari. Sehingga dapat disimpulkan bahwa kualitas udara di Provinsi DKI Jakarta tahun 2021 cenderung berada di kategori tidak sehat.

6. Kontribusi Penulis

Mahendrasyah, I: *Data collection*, *Formal Analysis*, *Investigation*, *Methodology*, *Visualization*, dan *Writing – original draft*. **Diana, A:** *Supervision*, *Validation*, dan *Review*. **Rusdah:** *Supervision*, *Validation*, dan *Review*. **Mahdiana, D:** *Supervision*, *Validation*, dan *Review*.

7. Declaration of Competing Interest

Saya sebagai penulis menyatakan bahwa hasil penelitian ini tidak terpengaruh oleh konflik kepentingan.

ISSN 2502-3357 (*Online*)
Ihjal, M dkk./TEKNOLOGI 14 (2) 2024 146-156 ISSN 2503-0477 (*Print*)

8. Referensi

Abidin, J., & Artauli Hasibuan, F. (2019). PENGARUH DAMPAK PENCEMARAN UDARA TERHADAP KESEHATAN UNTUK MENAMBAH PEMAHAMAN MASYARAKAT AWAM TENTANG BAHAYA DARI POLUSI UDARA. *Prosiding SNFUR-4*, *7*, 1–3.

Akbar Rismayadi, A., Nur Fatonah, N., & Junianto, E. (2021). ALGORITMA K-MEANS CLUSTERING UNTUK MENENTUKAN STRATEGI PEMASARAN DI CV. INTEGREET KONSTRUKSI. *JURNAL RESPONSIF*, *3*(1), 30–36. http://ejurnal.ars.ac.id/index.php/jti

Anjelita, M., Windarto, A. P., Wanto, A., & Sudahri, I. (2020). Pengembangan Datamining Klastering Pada Kasus Pencemaran Lingkungan Hidup. *Seminar Nasional Teknologi Komputer & Sains (SAINTEKS)*, *3*(1), 309–313.

Aritonang, E. D., Satria Tambunan, H., Hardinata, J. T., Irawan, E., & Suhendro, D. (2020). Penerapan Data Mining Dalam Mengelompokkan Provinsi Rawan Kejahatan Menggunakan Algoritma K-Means. *KOMIK (Konferensi Nasional Teknologi Informasi Dan Komputer)*, *4*(1), 35–42. https://doi.org/10.30865/komik.v4i1.2576

Ariwisanto Sianturi, F., Marto Hasugian, P., Simangunsong, A., & Nadeak, B. (2019). *DATA MINING: Teori dan Aplikasi Weka* (H. Tamando Sihotang, Ed.; 1st ed., Vol. 1). CV.Rudang Mayang. https://iocscience.org

Ayu Indah Cahya Dewi, D., & Ayu Kadek Pramita, D. (2019). Analisis Perbandingan Metode Elbow dan Sillhouette pada Algoritma Clustering K-Medoids dalam Pengelompokan Produksi Kerajinan Bali. *JURNAL MATRIX*, *9*(3), 102.

Azzahra Nasution, D., Khotimah, H. H., & Chamidah, N. (2019). PERBANDINGAN NORMALISASI DATA UNTUK KLASIFIKASI WINE MENGGUNAKAN ALGORITMA K-NN. *CESS (Journal of Computer Engineering System and Science)*, *4*(1), 2502–7131.

Citra Mawani, A., Li Hin, L., & Anubhakti, D. (2023). DETEKSI DINI GEJALA AWAL PENYAKIT DIABETES MENGGUNAKAN ALGORITMA RANDOM FOREST. *Idealis: Indonesia Journal Information System*, *6*(2), 165–171. http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/indexAjengCitraMawani|http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/index|

Darmo Wihardjo, S., & Rahmayanti, H. (2021). *PENDIDIKAN LINGKUNGAN HIDUP* (S. Ramadhan, Ed.; 1st ed., Vol. 1). PT. Nasya Expanding Management.

Fadilah, N. (2022). PENERAPAN METODE ALGORITMA K-MEANS UNTUK CLUSTERING DAERAH RAWAN TANAH LONGSOR DI PROVINSI JAWA TENGAH. *Jurnal BATIRSI*, *6*(1), 1–5. https://bpbd.jatengprov.go.id/.

Farmana Putra, R., Sandra Yofa Zebua, R., Budiman, Wibawa Rahayu, P., Theo Ari Bangsa, M., Zulfadhilah, M., Choirina, P., Wahyudi, F., & Andiyan, A. (2023). *Data Mining : Algoritma dan Penerapannya* (Efitra & Sepriano, Eds.; 1st ed., Vol. 1). PT. Sonpedia Publishing Indonesia.

Hermawan, & Hasugian, H. (2022). Penerapan Data Mining Untuk Clustering Indeks Pembangunan Manusia Berdasarkan Provinsi Di Indonesia. *Seminar Nasional Mahasiswa Fakultas Teknologi Informasi (SENAFTI)*, *1*(1), 525–532.

Imas Agista, P., Gusdini, N., & Dewi Dyah Maharani, M. (2020). ANALISIS KUALITAS UDARA DENGAN INDEKS STANDAR PENCEMAR UDARA (ISPU) DAN SEBARAN KADAR POLUTANNYA DI PROVINSI DKI JAKARTA. *Jurnal SEOI – Fakultas Teknik Universitas Sahid Jakarta*, *2*(2), 39–57. https://doi.org/10.36441/seoi.v2i2.491

Kamila, I., Khairunnisa, U., & Mustakim. (2019). Perbandingan Algoritma K-Means dan K-Medoids untuk Pengelompokan. *Jurnal Ilmiah Rekayasa Dan Manajemen Sistem Informasi*, *5*(1), 119–125.

Ihjal, M dkk./TEKNOLOGI 14 (2) 2024 146-156
ISSN 2502-3357 (*Online*)
ISSN 2503-0477 (*Print*)

Khairunnisa, Syahrul Jiwandono, I., Nurhasanah, Kemala Dewi, N., Hadi Saputra, H., & Linggo Wati, T. (2019). Kampanye Kebersihan Lingkungan melalui Program Kerja Bakti Membangun Desa di Lombok Utara. *Jurnal Pendidikan Dan Pengabdian Masyarakat* , *2*(2), 230–234.

Mai Sarah Tarigan, P., Tata Hardinata, J., Qurniawan, H., Safii, M., & Winanjaya, R. (2022). IMPLEMENTASI DATA MINING MENGGUNAKAN ALGORITMA APRIORI DALAM MENENTUKAN PERSEDIAAN BARANG (STUDI KASUS : TOKO SINAR HARAHAP). *Just IT : Jurnal Sistem Informasi, Teknologi Informasi Dan Komputer*, *12*(2), 51–61. https://jurnal.umj.ac.id/index.php/just-it/index

Nindy Yuliarina, A., & Hendry. (2022). COMPARISON OF PREDICTION ANALYSIS OF GOFOOD SERVICE USERS USING THE KNN & NAIVE BAYES ALGORITHM WITH RAPIDMINER SOFTWARE. *Jurnal Teknik Informatika (Jutif)*, *3*(4), 847–856. https://doi.org/10.20884/1.jutif.2022.3.4.294

Noor Permata Sari, D., & Sukestiyarno, Y. L. (2021). Analisis Cluster dengan Metode K-Means pada Persebaran Kasus Covid-19 Berdasarkan Provinsi di Indonesia. *PRISMA, Prosiding Seminar Nasional Matematika*, *4*, 602–610. https://journal.unnes.ac.id/sju/index.php/prisma/

Oktaviani, A., & Hustinawati. (2021). PREDIKSI RATA-RATA ZAT BERBAHAYA DI DKI JAKARTA BERDASARKAN INDEKS STANDAR PENCEMAR UDARA MENGGUNAKAN METODE LONG SHORT-TERM MEMORY. *Jurnal Ilmiah Informatika Komputer*, *26*(1), 41–55. https://doi.org/10.35760/ik.2021.v26i1.3702

Sari, Y. R., Sudewa, A., Lestari, D. A., & Jaya, T. I. (2020). PENERAPAN ALGORITMA K-MEANS UNTUK CLUSTERING DATA KEMISKINAN PROVINSI BANTEN MENGGUNAKAN RAPIDMINER. *CESS (Journal of Computer Engineering System and Science)*, *5*(2), 192–198.

Virantika, E., Kusnawi, K., & Ipmawati, J. (2022). Evaluasi Hasil Pengujian Tingkat Clusterisasi Penerapan Metode K-Means Dalam Menentukan Tingkat Penyebaran Covid-19 di Indonesia. *JURNAL MEDIA INFORMATIKA BUDIDARMA*, *6*(3), 1657–1666. https://doi.org/10.30865/mib.v6i3.4325