E-ISSN 2527-9378

Jurnal Statistika Industri dan Komputasi
Volume 09, No. 02, Juli 2024, pp. 60-71

PENERAPAN METODE K-MEANS PADA DATA ORDINAL UNTUK PENGELOMPOKAN DAERAH BERDASARKAN KUALITAS UDARA DI DAERAH ISTIMEWA YOGYAKARTA

Sofyan Sosnegher Ndelawa[1], Rokhana Dwi Bekti[2]*, Maria Titah Jatipaningrum[3], Febriani Astuti[4]
[1,2,3,4] Program Studi Statistika, Fakultas Sains dan Teknologi Informasi, Universitas AKPRIND Indonesia
Email: rokhana@akprind.ac.id
**corresponding author*

Abstract. *Quality is an important factor for human health and is a long-term concern, especially in urban areas. So it must be maintained with efforts to control air pollution both for the central government and local governments, especially in the DIY Province. KPI is a value that shows the quality of goodness or air quality based on its provisions. It is important to group the regions in each province to make it easier for the public to know the development of air quality. This research uses the K-Means method with the Hamming Distance approach and the BOS Cluster distribution to group regions based on air quality in the Special Region of Yogyakarta. Based on K-Means with Hamming Distance approach, it is obtained that Cluster 1 consists of 2 data (regions) and Cluster 2 consists of 2 data (regions), Cluster 3 is 40 data (regions) and Cluster 4 = 3 data (regions). For K-Means with BOS Cluster distribution, it is obtained that Cluster 1 consists of 42 data (regions) and Cluster 2 consists of 2 data (regions), Cluster 3 as much as 4 data (regions) and Cluster 4 as much as 2 data (regions). Thus, based on the results of the validity of the K-Means Hamming Distnace approach is the best method in grouping air quality in the Special Region of Yogyakarta Province based on indicators.*

Keywords*: Air Quality, ISPU, Clustering, K-Means, Hamming Distance, BOS Cluster*

Abstrak. Udara merupakan kebutuhan terpenting dalam kehidupan ini, sehingga kualitas dan cara menjaganya harus perlu diperhatikan. Kualitas udara merupakan faktor penting bagi kesehatan manusia dan merupakan perhatian jangka panjang, terutama di daerah perkotaan. Sehingga harus dijaga dengan upaya-upaya pengendalian pencemaran udara baik bagi pemerintah pusat maupun pemerintah daerah khsusnya wilayah di Daerah Provinsi DIY. IKU merupakan nilai yang menunjukkan mutu kebaikan atau kualitas udara berdasarkan ketentuannya. Penting dilakukan pengelompokan terhadap wilayah-wilayah di setiap Provinsi agar mempermudah Masyarakat mengetahui perkembangan kualitas udara. Penelitian kali ini menggunakan metode K-Means dengan pendekatan *Hamming Distance* dan distribusi *BOS Cluster* untuk pengelompokan daerah berdasarkan kualitas udara di Daerah Istimewa Yogyakarta. Berdasarkan *K-Means* dengan pendekatan *Hamming Distance* diperoleh untuk *Cluster* 1 terdiri atas 2 data (wilayah) dan pada *Cluster* 2 terdiri 2 data (wilayah), *Cluster* 3 sebanyak 40 data (wilayah) dan *Cluster* 4 = 3 data (wilayah). Untuk *K-Means* dengan distribusi *BOS Cluster* diperoleh untuk *Cluster* 1 terdiri atas 42 data (wilayah) dan pada *Cluster* 2 terdiri 2 data (wilayah), *Cluster* 3 sebanyak 4 data (wilayah) dan *Cluster* 4 sebanyak 2 data (wilayah). Sehingga, berdasarkan hasil validitas *K-Means* pendekatan *Hamming Distnace* merupakan metode terbaik dalam pengelompokan kualitas udara di Provinsi Daerah Istimewa Yogyakarta berdasarkan indikator kualitas udara.

Kata kunci: Kualitas Udara, ISPU, *Clustering*, *K-Means* , *Hamming Distance*, *BOS Cluster*

1. Pendahuluan

Kehidupan manusia sangat dipengaruhi oleh keadaan lingkungan sekitar, salah satunya kualitas udara. Semakin baik kualitas udara, semakin baik pula kualitas kesehatan masyarakat. Dengan demikian permasalahan tentang kualitas udara sangat penting udara terus diatasi. Pencemaran udara (*air pollution*) telah menjadi perhatian manusia beberapa dekade yang lalu.

Secara umum ada dua kelompok standar kualitas udara, yakni kualitas udara *ambien* (lingkungan) dan kualitas udara emisi industri. Udara *ambien* (lingkungan) adalah udara bebas yang berada di atmosfer. Jenis udara ini biasanya bersumber dari geografi dan iklim suatu daerah. Sedangkan udara emisi adalah udara yang berasal dari sumber emisi, seperti kendaraan bermotor, industri, konstruksi dan lain-lain. Dalam penggunaannya, suatu batasan emisi bahan pencemar tertentu adalah sama dengan 30 kali dari pada standar udara *ambien* (lingkungan).

Agar informasi tentang mutu udara mudah dipahami oleh masyarakat, hasil pemantauan mutu udara dari stasiun pemantauan dalam hal ini dibawahi langsung Kementerian Lingkungan Hidup dan Kehutanan (KLHK) yang disampaikan dalam bentuk Indeks Standar Pencemar Udara (ISPU). Indeks Standar Pencemaran Udara merupakan angka tanpa satuan, digunakan untuk menggambarkan kondisi mutu udara ambien di lokasi tertentu dan didasarkan kepada dampak terhadap kesehatan manusia. Standar ini digunakan oleh pemerintahan Indonesia dengan menyesuaikan dengan standar yang telah digunakan oleh organisasi kesehatan dunia dalam hal ini WHO (World Health Organization). ISPU bertujuan agar memberikan kemudahan dari keseragaman informasi mutu udara ambien kepada masyarakat di lokasi dan waktu tertentu serta sebagai bahan pertimbangan dalam melakukan upaya-upaya pengendalian pencemaran udara baik bagi pemerintah pusat maupun pemerintah daerah. Menurut Peraturan Menteri Lingkungan Hidup dan Kehutanan Nomor P.14/Menlhk/Setjen/Kum.1/7/2020 tentang Indeks Standar Pencemar Udara, ISPU merupakan angka tanpa satuan yang menggambarkan kondisi kualitas udara ambien di lokasi tertentu, yang didasarkan pada dampak terhadap kesehatan manusia, nilai estetik, dan makhluk hidup lainnya. Pada peraturan tersebut mencantumkan parameter ISPU yakni Hidrokarbon (HC), Karbon monoksida (CO), Sulfur dioksida (SO2), Nitrogen dioksida (NO2), Ozon (O3), dan Partikulat (PM10 and PM2,5).

Untuk melakukan pengelompokan data dalam skala besar dapat digunakan metode *Data mining*. *Data mining* bertujuan untuk mengidentifikasi pola, hubungan, dan tren yang tersembunyi dalam data dalam mendapatkan informasi yang lebih baik untuk berbagai bidang. *Clustering* merupakan salah satu teknik atau metode yang digunakan dalam proses *data mining*. Pada dasarnya *Clustering* merupakan suatu metode untuk mencari dan mengelompokkan data yang memiliki kemiripan karakteristik (*similarity*) antara satu data dengan data yang lain (Oscar, 2013). Banyak metode pengelompokan *Cluster*ing yang dapat digunakan diantaranya K-Means, Fuzzy C-means, Mixture Modelling K-Medoids, Single Linkage, Complete Linkage, Average Linkage dan Average Group Linkage. Dalam Penelitian ini, peneliti menggunakan algoritma K-Means sebagai metode untuk mengatasi permasalah tersebut.

K-Means merupakan salah satu algoritma *Cluster*ing yang masuk dalam kelompok Unsupervised learning yang digunakan untuk mengelompokkan data kedalam beberapa kelompok dengan sistem partisi. Metode ini mempartisi data ke dalam *Cluster*/kelompok sehingga data yang memiliki karakteristik yang sama dikelompokkan ke dalam satu *Cluster* yang sama dan data yang mempunyai karakteristik yang berbeda dikelompokkan ke dalam kelompok yang lain. Adapun tujuan dari data *Cluster*ing ini adalah untuk meminimalisasikan *Objective Function* yang diset dalam proses *Cluster*ing, yang pada umumnya berusaha meminimalisasikan variasi di dalam suatu *Cluster* dan memaksimalisasikan variasi antar *Cluster* (Y. Agusta, 2007). Secara umum dalam pengolahannya, metode k-means menggunakan data yang atributnya adalah berupa numerik (angka). Data selain angka (non-numerik) juga bisa diterapkan tetapi terlebih dahulu harus dilakukan transformasi (pengkodean) untuk mempermudah perhitungan jarak/kesamaan karakteristik yang dimiliki dari setiap objek.

Selanjutnya, dalam melakukan pengukuran jarak antara suatu objek dengan obyek lainnya pada metode K-Means. Berbagai macam bentuk jarak (distance), antara lain adalah *Euclidean Distance, City Block (Manhattan) Distance, Chebyshev Distance, Minkowski Distance, Canberra Distance,* dan *Hamming Distance* (Murti dkk., 2005). Pada kasus lain Hamming distance adalah metrik jarak yang digunakan untuk mengukur perbedaan atau kesamaan antara dua buah string atau vektor dengan panjang yang sama dan dapat digunakan untuk membandingkan elemen-

elemen ordinal pada posisi yang sama dalam vektor-vektor yang diukur. Penggunaan data kategori (ordinal) dapat memberikan wawasan yang lebih tentang tingkatan, preferensi dalam pengambilan kesimpulan, khususnya dalam permasalahan lingkungan sosial.

Pendekatan lain bagi K-Means untuk data ordinal yaitu dengan model BOS (Binary Ordinal Search). Biernacki dan Jacques (2016) mengusulkan apa yang disebut model Pencarian Ordinal Biner, yang disebut sebagai model BOS (Bounded Self-Selection). Ini adalah distribusi probabilitas khusus untuk data ordinal yang di parameterisasi dengan parameter yang bermakna (μ,π).

Metode pengelompokan menggunakan K-Means banyak digunakan oleh berbagai peneliti, seperti yang dilakukan oleh Adrian dkk., (2016) menyatakan bahwa hasilnya, dari profile data sebelum dilakukan pengelompokan diperoleh kadar CO, SO2, NO2, dan O3 terendah berada pada titik-titik wilayah permukiman dan tertinggi berada pada titik perempatan jalan, training camp, kampus fakultas teknik, dan industri dan hasil dari perbandingan *Cluster*ing validity index terbentuk sebanyak 2 kelompok. Kelompok 1 memiliki titik tengah kadar pencemar gas NO2, SO2, CO, dan O3 yang lebih tinggi dibandingkan kelompok 2. Kelompok 1 terdiri atas 45 anggota, dimana sebagian besar kelompok ini merupakan titik industri, persimpangan jalan, serta pusat keramaian. Sedangkan kelompok 2 terdiri atas 30 anggota, dimana sebagian besar kelompok ini merupakan titik permukiman.

Penelitian yang dilakukan oleh Murti dkk., (2010), dimana melakukan mengembangkan suatu aplikasi *Cluster*ing pada data-data non-numerik pada kasus biro jodoh dengan menggunakan algoritma *K-means* dan *Hamming Distance*. Uji coba dan evaluasi dilakukan dengan menggunakan dataset nyata yaitu data biro jodoh Grasco, Sakinah Surabaya, Libe, dan O'Dea. Dari uji coba tersebut didapatkan bahwa *Cluster*ing dapat dilakukan pada atribut atribut categorical yang ditransformasikan terlebih dahulu ke dalam bentuk numerik. Selain itu, kesamaan (similarity) dan karakteristik dari masing-masing keanggotaan biro jodoh bisa diketahui.

Berdasarkan penelitian-penelitian sebelumnya, penelitian kali ini menggunakan metode K-Means untuk pengelompokan daerah berdasarkan kualitas udara di Daerah Istimewa Yogyakarta dengan menggunakan data berbentuk ordinal, diharapkan dapat menghasilkan nilai jumlah *Cluster*/kelompok wilayah-wilayah di Provinsi DIY berdasarkan data kualitas udara. Dimana menggunakan 3 (tiga) parameter ISPU (Indeks Standar Kualitas Udara) yakni, CO, NO_2 dan SO_2. Diharapkan dengan analisis ini dapat membantu pemerintah dalam pengambilan kebijakan agar pengoptimalan pencegahan pencemaran terhadap kualitas udara yang ada di provinsi Daerah Istimewa Yogyakarta

2. Metode Penelitain

2.1. Desain Penelitian

Desain yang digunakan adalah pendekatan kuantitatif deskriptif yaitu penelitian yang menggunakan data berupa angka atau numerik. Untuk interpretasi hasil dilakukan dengan analisis dan susunan data yang sudah ada sesuai dengan kebutuhan peneliti. Dengan maksud peneliti ingin menerapkan penggunaan metode *K-Means Clustering* untuk data ordinal dengan menggunakan perhitungan *Hamming Distance* dan pendekatan *BOSS Cluster*.

2.2. Objek Penelitian

Lokasi penelitian dilakukan di wilayah berdasarkan jumlah titik pengawasan kualitas udara Provinsi di Daerah Istimewa Yogyakarta. Titik-titik pengawasan tersebut sebagai berikut.

1. Sekitar Jalan Raya;

2. Sekitar Industri;
3. Sekitar Pemukiman;
4. Sekitar Yogyakarta

2.3. Metode Pengumpulan Data

Data yang digunakan pada penelitian ini adalah data sekunder yang diperoleh dari Dinas Lingkungan Hidup dan Kehutanan (DLHK) Provinsi Daerah Istimewa Yogyakarta Tahun 2019.

2.4. Variabel Penelitian

Adapun variabel yang digunakan dalam penelitian merupakan parameter-parameter Indeks Standar Kualitas Udara (ISPU), yakni Karbon Dioksida (CO), Sulfur Dioksida (SO_2) dan Nitrogen Dioksida (NO_2). Adapun data dari setiap variabel tersebut terdiri dari 3 kategori yakni, baik, sedang, dan tidak sehat/buruk.

2.5. Metode Analisis

Metode yang digunakan dalam penelitian ini adalah *K-Means Clustering* pada data ordinal dengan menggunakan perhitungan *Hamming Distance* dan pendekatan *BOSS Cluster*. Dengan tahapan-tahapan sebagai berikut:

1. Memilih parameter yang diduga memiliki tingkat yang cukup signifikan terhadap pencemaran udara;
2. Menghitung nilai Indeks Standar Pencemaran Udara (ISPU). Selanjutnya dilakukan pengkategorian ISPU yang nantinya digunakan untuk pengelompokan;
3. Melakukan analisis deskriptif untuk melihat karakteristik data;
4. Melakukan pengelompokan menggunakan metode analisis *K-Means* dengan *Hamming Distance*
5. Melakukan pengelompokan menggunakan metode pendekatan *Binary Ordinal Search (BOS) Cluster*
6. Perbandingan *Cluster* terbaik yang terbentuk Melakukan profiling *Cluster* untuk menentukan metode terbaik dalam pengelompokan

3. Hasil dan Pembahasan

3.1. Analisis Deskriptif Kualitas Udara di Daerah Istimewa Yogyakarta

Berikut ini merupakan tabel hasil analisis deskriptif pada wilayah Daerah Istimewa Yogyakarta berdasarkan Indeks Standar Pencemaran Udara (ISPU) tahun 2019:

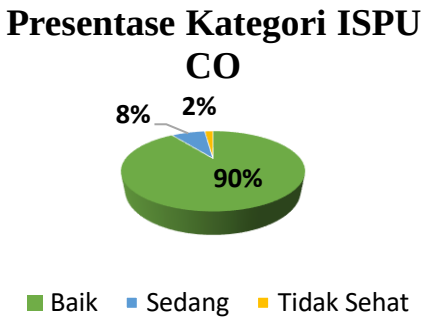
Tabel 1. Karakteristik Data Kualitas Udara DIY 2019

Variabel/Parameter	Mean	Standar Deviasi	Max	Min	Total (n)
CO	78,10	51,98	203,29	11,20	50
SO_2	33,92	17,47	93,31	10,19	50
NO_2	167,75	37,42	167,75	7,39	50

Penelitian ini menggunakan 3 variabel yang merupakan parameter pada Indeks Standar Pencemaran Udara (ISPU) yang digunakan pada wilayah Provinsi Daerah Istimewa Yogyakarta.

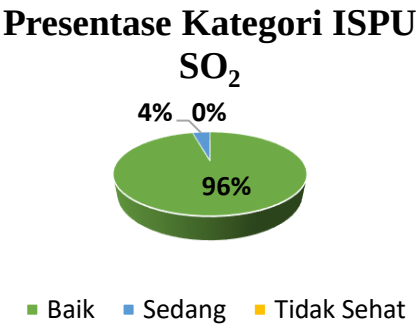
Variabel tersebut meliputi; Karbon Monoksida (CO), Sulfur Dioksida (SO_2), dan Nitrogen Dioksida (NO_2).

3.2. Karakteristik Kategori Indeks Pencemaran Udara (ISPU)



Gambar 1. Visualisasi Peresentase Kategori ISPU CO

Berdasarkan Gambar 1, terdapat 90% lokasi dengan ISPU CO berada dalam kategori Baik. Hal ini berarti bahwa sebagian besar lokasi masuk kategori baik, yang menunjukan kualitas udara di DIY relatif baik.



Gambar 2. Visualisasi Persentase ISPU SO_2

Berdasarkan Gambar 2, terdapat 88% lokasi dengan ISPU NO_2 berada dalam kategori baik. Hal ini berarti bahwa sebagian besar lokasi masuk kategori baik, yang menunjukan kualitas udara di DIY relatif baik.

Gambar 3. Visualisasi Persentase Kategori ISPU NO_2

Berdasarkan Gambar 3, terdapat 88% lokasi dengan ISPU NO_2 berada dalam kategori baik. Hal ini berarti bahwa sebagian besar lokasi masuk kategori baik, yang menunjukan kualitas udara di DIY relatif baik.

3.3. Transformasi Data

Untuk menggunakan metode *BOSCluster*, data yang telah dikategorikan dilakukan pengkodean pada setiap kategori dalam data. Berikut pengkodean kategori tersebut:

Tabel 2. Tranformasi Setiap Kategori ISPU ke Data Ordinal

Kategori	Kode
Baik	1
Sedang	2

Berikut merupakan hasil *binning* dengan data kategori yang akan digunakan.

Tabel 3. Hasil *Binning* Kategori

Kategori	Hasil *Binning*
Baik	[**1 0 0**]
Sedang	[**0 1 0**]
Tidak Sehat	[**0 0 1**]

Tabel 4. Hasil *Binning* Kategori ISPU

No.Loc	**Hasil *Binning* Kategori ISPU**					
	CO	*Binning* CO	SO_2	*Binning* SO_2	NO_2	*Binning* NO_2
1	Baik	[1 0 0]	Baik	[1 0 0]	Baik	[1 0 0]
5	Sedang	[0 1 0]	Baik	[1 0 0]	Baik	[1 0 0]
7	Tidak Sehat	[0 0 1]	Baik	[1 0 0]	Sedang	[0 1 0]
41	Baik	[1 0 0]	Sedang	[0 1 0]	Baik	[1 0 0]
42	Baik	[1 0 0]	Baik	[1 0 0]	Sedang	[0 1 0]
…						
50	Baik	[1 0 0]	Sedang	[0 1 0]	Baik	[1 0 0]

3.3.1.Pengelompokan menggunakan Hamming Distance

1. Menentukan jarak anatar amatan

   Wilayah ke-1 : |**1 0 0**|

   Wilayah ke-2 : |**1 0 0**|

   maka,

$$d_{1,2} = |1-1| + |0-0| + |0-0| = 0$$

Perhitungan jarak antar objek berlanjut hingga data atau objek wilayah ke-50, sehingga didapatkan matriks jarak hamming antar objek sebagai berikut.

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ \vdots \\ 50 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & \ldots & 50 \end{array} \\ \begin{bmatrix} 0 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & 0 \end{bmatrix} \end{array}$$

Jumlah *Cluster* yang akan dibentuk yaitu 4 *Cluster* atau *k* = 4. Pada jarak hamming bisa mengambil data acak dan melihat pembentukan k dari sebuah data. Pada penelitian kali ini misal akan dikelompokan data sebanyak 4 *Cluster*.Diambil secara acak pusat *Cluster* pada data pertama dengan *k = 2* sebagai berikut;

Centroid 1 (C1) = [**0 0 1**]
Centroid 2 (C2) = [**1 1 1**]
Centroid 3 (C3) = [**1 0 0**]
Centroid 4 (C4) = [**1 1 0**]

Jika angka pada jarak tersebut mendekati nilai pusat *Cluster* 1 maka data pada objek tersebut masuk pada *Cluster* 1. Berikut hasil *Cluster* dari data dengan menggunakan *k* = 2 tersebut.

Tabel 5. Pembagian Hasil *Cluster* Berdasarkan Metode *K-Means*

	Cluster 1	*Cluster* 2	*Cluster* 3	*Cluster* 4
Data Ke-	5, 6, 20	8, 9, 29, 30, 42	7, 19	1, 2, 3, 4, 10, 11, 12, 13, 14, 15, 16, 17, 18, 21, 22, 23, 24, 25, 26, 27, 28, 31, 32, 33, 34, 35, 36, 37, 38, 39,40, 41, 43, 44, 45, 46, 47, 48, 49, 50

Berikut penjelasannya menggunakan Tabel berikut.

Tabel 6. Lokasi disetiap *Cluster* Berdasarkan Metode *K-Means Hamming Distance*

No. *Loc*	**Nama Wilayah**	***Cluster***
7	Depan TVRI Jl. Magelang…	3
19	Depan Hotel Shapir Jl. Lak…	3

No. *Loc*	**Nama Wilayah**	***Cluster***
1	Simpang empat Ngepl….	4
2	Simpang tiga Toya….	4

No. *Loc*	**Nama Wilayah**	***Cluster***
8	Depan UPN Seturan…	2
9	Depan Ruiko Janti Jl…	2
29	Terminal Wonosari…	2
30	Perempatan RSUD…	2
42	Tapak/ PT. Madu Baru…	2

3	Terminal wates Kulon Progo	4
⋮		⋮
⋮		⋮
50	Sebelah Selatan PC. GKBI…	4

No. *Loc*	**Nama Wilayah**	***Cluster***
5	Depan GKBI Medari Sleman	1
6	Perempatan Denggung Sleman	1
20	Depan Kantor Kecamatan Jetis….	1

3.3.2.Profiling *Cluster*

Karakteristik *Cluster* dilihat berdasarkan nilai rata-rata setiap variabel pada masing-masing *Cluster*. Karena data berupa oridinal, maka profiling dibuat 2 tabel.

Tabel 7. Frekuensi Setiap Kategori pada *Cluster*

Cluster	ISPU CO			ISPU SO_2			ISPU NO_2		
	Baik	Sedang	Tidak Sehat	Baik	Sedang	Tidak Sehat	Baik	Sedang	Tidak Sehat
1	0	3	0	3	0	0	2	1	0
2	1	0	1	2	0	0	1	1	0
3	39	1	0	38	2	0	37	3	0
4	5	0	0	5	0	0	3	1	1

Tabel 8. Profiling Data Sebelum dikategorikan

Cluster	ISPU CO	ISPU SO2	ISPU NO2
1	66,47	67,61	65,9
2	92,39	15,44	16.2
3	28,5	36,44	27,8
4	30,8	21,2	28,7

Berdasarkan Tabel 8, dapat diketahui bahwa *Cluster* 3 dan 4 cenderung memiliki kualitas udara yang relatif rendah dibandingkan *Cluster* 1 dan 2, karena memiliki rata-rata ISPU CO, SO_2 dan NO_2 ayng relatif rendah.

3.3.3.Pengelompokan menggunakan BOS*Cluster*

Dalam metode pengelompokan menggunakan model BOS, data yang digunakan akan ditranformasi menjadi data ordinal, dimana sebelumnya data kualitas udara berdasarkan wilayah Daerah Istimewa Yogyakarta yang berupa numerik akan dikategorikan dan selanjutnya ditranformasi ke bentuk data ordinal. Berikut hasil transformasi data tersebut.

Tabel 9. Hasil Tranformasi

Variabel	Data ke-	Kategori	Hasil Tranformasi
CO	1	Baik	1
NO_2	41	Sedang	2
SO_2	9	Tidak Sehat	3

Pada data yang disediakan n = 50 dan jumlah variabel J = 3, sehingga dapat diketahui bahwa matriks $\boldsymbol{x}$ berukuran 50×3, dimana setiap barisnya adalah variabel ordinal univariat yang terdiri dari J respons ordinal x_{ij}, dengan $1 \leq i \leq 50$ dan $1 \leq j \leq 3$.

Mengestimasi parameter campuran dengan dengan menggunakan algoritma Expectation-Maximization (EM) dengan variabel laten tambahan. Sehingga didapatkan hasil sebagai berikut.

Tabel 10. Hasil Pembentukan Parameter

	$\boldsymbol{\mu}$			$\boldsymbol{\pi}$		
Cluster	V_1	V_2	V_3	V_1	V_2	V_3
1	1	1	1	0.9001887	1	1
2	3	1	2	0.3646419	1	1
3	1	1	2	1	1	0.7321329
4	1	2	1	1	1	1

Tabel 11. Hasil Pembentukan *Cluster*

	Cluster 1	*Cluster* 2	*Cluster* 3	*Cluster* 4
Data Ke-	1, 2, 3, 4, 5, 6, 10, 11, 12, 13, 14, 15, 16, 17, 18, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29 32, 33, 34, 35, 36, 37, 38, 39, 40, 42, 43, 44, 45, 46, 47, 48, 49.	7, 19	8, 9, 30, 31	41, 50

Berdasarakan hasil tersebut, terlihat bahwa untuk *Cluster* 1 terdiri atas 42 data (wilayah) dan pada *Cluster* 2 terdiri 2 data (wilayah), *Cluster* 3 = 4 data (wilayah) dan *Cluster* 4 = 2 data (wilayah). Dimana, anggota *Cluster* 1 terdiri dari data ke- 1, 2, 3, 4, 5, 6, 10, 11, 12, 13, 14, 15, 16, 17, 18, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29 32, 33, 34, 35, 36, 37, 38, 39, 40, 42, 43, 44, 45, 46, 47, 48, 49, pada *Cluster* 2 terdiri data ke- 7, 19. Sedangkan untuk anggota *Cluster* 3 hanya terdapat 4 data yakni data ke-8, 9, 30, 31, dan pada *Cluster* 4 terdiri dari data ke- 41, 50.

3.3.4.Profiling *Cluster*

Karakteristik *Cluster* dilihat berdasarkan nilai rata-rata setiap variabel pada masing-masing *Cluster*. Karena data berupa oridinal, maka profiling dibuat 2 tabel, berikut hasil profiling.

Tabel 12. Frekuensi Setiap Kategori pada *Cluster*

Cluster	ISPU CO			ISPU SO2			ISPU NO2		
	1	2	3	1	2	3	1	2	3
1	0	3	0	3	0	0	2	1	0
2	1	0	1	2	0	0	1	1	0
3	39	1	0	38	2	2	37	3	0
4	5	0	0	5	0	0	3	1	1

Tabel 13. Profiling Data Sebelum dikategorikan

Cluster	ISPU CO	ISPU SO2	ISPU NO2
1	66,7	17,91	36,32
2	92,39	10,45	51,37
3	20,92	23,32	22,65
4	36,42	16,24	58,02

3.4. Perbandingan validasi metode *K-Means Hamming Distance* dan BOSS*Cluster*

Perbandingan hasil yang dimaksud adalah menemukan metode terbaik antara K-Means Hamming Distance dan BOSS *Cluster* berdasarkan kriteri nilai validitas. Berikut hasil perbandingan kedua metode tersebut dengan menggunakan *Silhoutte Index* dan *Within-Sum of Square* (WSS).

Tabel 14. Hasil Validasi dari dua metode dengan K=4 dan K=2

Algoritma *K-Means*	*K* Optimal	*Silhoute Index*	*Sum of Square*
Hamming Distance	4	0.91	98.4%
BOS Cluster	4	0.88	85.1%
Hamming Distance	2	0.89	85.5%
BOS Cluster	2	0.78	84.8%

Dari hal tersebut, didapatkan bahwa algoritma terbaik adalah algoritma yang memiliki hasil nilai *Silhouette Index* mendekati satu, oleh karena itu dapat disimpulkan bahwa metode terbaik dalam pengelompokan kualitas udara di Provinsi Daerah Istimewa Yogyakarta berdasarkan indikator kualitas udara adalah *Hamming Distance* pada k=4 dengan nilai *Silhouette Index* (SI) sebesar 0,91 dan *Within- Sum of Square* (WSS) sebesar 98.4%. Hal ini berarti bahwa gambaran dalam proses klasterisasi hampir sepenuhnya bisa dijelaskan atau masuk dalam indikasi sangat baik antar klasternya.

4. Kesimpulan

Dalam penelitian ini berdasarkan hasil analisis dan pembahasan, hasil pengelompokan menggunakan K-Means dengan pendekatan Hamming Distance diperoleh untuk *Cluster* 1 terdiri atas 2 data (wilayah) dan pada *Cluster* 2 terdiri 2 data (wilayah), *Cluster* 3 sebanyak 40 data (wilayah) dan *Cluster* 4 = 3 data (wilayah). Berdasarkan perhitungan validasi *Cluster* diperoleh nilai silhouette index sebesar 0,91 mendekati 1, hal ini berarti bahwa kualitas *Cluster* terhadap analisis kualitas udara di Provinsi DIY memiliki struktur kuat. Hasil pengelompokan menggunakan K-Means dengan distribusi BOS *Cluster* diperoleh untuk *Cluster* 1 terdiri atas 42 data (wilayah) dan pada *Cluster* 2 terdiri 2 data (wilayah), *Cluster* 3 sebanyak 4 data (wilayah) dan *Cluster* 4 sebanyak 2 data (wilayah). Berdasarkan perhitungan validasi *Cluster* diperoleh nilai silhouette index sebesar 0,88 mendekati 1, hal ini berarti bahwa kualitas *Cluster* terhadap analisis kualitas udara di Provinsi DIY memiliki struktur cukup kuat. Metode terbaik dalam pengelompokan wilayah berdasarkan kualitas udara Provinsi di Daerah Istimewa Yogyakarta

adalah K-Means dengan pendekatan Hamming Distance pada k=4 dengan nilai *Silhouette Index* sebesar 0,91 dan nilai *Within-Sum Square* sebesar 98.4%

Daftar Pustaka

Agusta, Y. (2007). K-Means - Penerapan, Permasalahan dan Metode Terkait. *Jurnal Sistem dan Informatika*, 47-60.

Aulasari, K., & Kertaningtyas, M. (2021). Analisis Kualitas Udara Menggunakan Algoritma K-Means. *Jurnal Informatika & Rekayasa Elektronika (JIRE)*.

Aulisari, K., & Kertanityas, M. (2021). Analisis Kualitas Udara Menggunakan Algoritma K-Means. *JIRE*, 95-101.

Biernacki, C., & Jasques, J. (2015). Model-based Clustering of Multivariate Ordinal Data Relying On a Stochastic Binary Search Algorithm. *International Journal of Applied Mathematics and Computer Science*.

Damayanti, T. V., & Handriyono, R. E. (2022). Monitoring Kualitas Udara Ambien Melalui Stasiun Pemantau Kualitas Udara Wonorejo, Kebonsari Dan Tandes Kota Surabaya. *Environmental Engineering Journal ITATS*.

DLHK. (2019). *Laporan Analisa Hasil Pemantauan Kualitas Udata Kota Yogyakarta*. Daerah Istimewa Yogyakarta: Dinas Lingkungan Hidup Kota Yogyakarta.

Firmansyah, A., Dewi, A. L., Hirma, E. S., Briliyanto, M. A., Nurjannah, M., & Nooraeni, R. (2020). Pengelompokan Titik Wilayah di Provinsi Daerah Istimewa Yogyakarta Berdasarkan Kualitas Udara Menggunakan Algoritma Fuzzy C-Means. *Jurnal Matematika dan Statistika serta Aplikasinya*, 99-108.

Giordan, M., & Diana, G. (2011). A Clustering Methods For Categorical Ordinal Data. *Communications in Statistics - Theory and Methods*, 1315-1334.

Hasrul, M., & Alwi, W. (2018). Analisis Klaster Untuk Pengelompokan Kabupaten/Kota di Provinsi Sulawesi Selatan Berdasarkan Indikator Kesejateraan Rakyat. *Jurnal Matematika dan Statistika Serta Aplikasinya*, 35-42.

Junaedi, H., Budianto, H., Maryati, I., & Melani, Y. (2011). Data Transformation Pada Data Mining. *Prosiding Konferensi Naional "Inovasi dalam Desain dan Teknologi"*, 93-99.

Khomarudin, A. N. (2016). Teknik Data Mining : Algoritma K-Means Clustering. *IlmuKomputer.com*.

Kurniawan, A. (2017). Pengukuran Parameter Kualitas Udara (CO, NO2, SO2, O3 dan PM10) Di Bukit Kototabang Berbasis ISPU. *Jurnal Teknosains*, 1-82.

Larose, C. D., & Larose, D. T. (2014). *Discovering Knowledge in Data: An Introduction to Data Mining*. New York: WILEY.

Lee, P., & Yu, P. L. (2012). Mixtures of Weighted Distance-based Models for Ranking Data With Applications in Political Studies. *Jurnal Computational Statistics & Data Analysis*, 2486-2500.

Murti, D. H., Suciati, N., & Nanjaya, D. J. (2005). Clustering Data Non-Numerik Dengan Pendekatan Algoritma K-Means dan Hamming Distance Studi Kasus Biro Jodoh. *Teknik ITS*, 1-47.

Nasir, J. (2020). Penerapan Data Mining Clustering dalam Mengelompokan Buku dengan Metode K-MEANS. *Jurnal SIMETRIS*.

Praktikno, A. S., Prastiwi, A. A., & Rahmawati, S. (2020). Pemetaan Ukuran Pemusatan Data. *OSF Preprint*.

Pratikno, A. S., Prastiwi, A. A., & Rahmawati, S. (2022). Ukuran Pemusatan Rata-rata. *OSF Preprints*.

Pratiwi, B. P., Handayani, A. S., & Sarjana. (2020). Pengukuran Kinerja Sistem Kualitas Udara dengan Teknologi WSN Menggunakan *Confusion Matrix*. *Jurnal Informatika Upgirls*.

San, O. M., Hynh, V.-N., & Nakamori, Y. (2004). An Alternative Extension Of The K-Means Algorithm For Clustering Categorical Data. *International Journal of Applied Mathematics And Computer Science*, 242-247.

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

Santoso, S. (2010). *Menguasai Statistik Multivariat.* Jakarta: Media Komputindo.

Sartika, E. (2010). Pengolahan Berskala Ordinal. *Jurnal Teknik Politeknik Negeri Bandung*, 60-69.

Selosse, M., Jacques, J., & Biernacki, C. (2021). ordinalClust: An R Package to Analyze Ordinal Data. *The R Journal*.

Setyawan, Y., Noeryanti, & Suryowati, K. (2018). *STATISTIKA DASAR Dilengkapi dengan Software R.* Yogyakarta: AKPRIND PRESS.

Ulinnuh, N., & Veriani, R. (2020). Analisis Cluster dalam Pengelompokan Provinsi di Indonesia Berdasarkan Variabel Penyakit Menular Menggunakan Metode Complete Linkage, Average LInkage dan Ward. *Jurnal Nsional Informatika dan Teknologi Jaringan*, 101-108.

Yunita, F. (2018). Penerapan Data Mining Menggunakan Algoritma K-Means Clustering Pada Penerimaan Mahasiswa (Studi Kasus : Universitas Islam Indragiri). *Jurnal SISTEMASI*, 238-249.

Yunita, R. D., & Kiswando, A. A. (2017). Kajian Indeks Standar Pencemaran Udara (ISPU) Sulfur Dioksida (SO2) Sebagai Polutan Udara Pada Tiga Lokasi Di Kota Bandar Lampung. *Analit : Analytical and Enviromental*.

Ziberna, A., Kejzar, N., & Golob, P. (2004). A Comparison of Different Approaches to Hierarchical Clustering of Ordinal Data. *Metodoloski Zvezki*, 57-73.