Jurnal Komputer dan Informatika
Vol 15 No 2, Oktober 2020: hlm 258-264

ISSN 1410-7228 (Versi Cetak)
ISSN 2716-4233 (Versi Elektronik)

CLUSTERING MENGGUNAKAN METODE K-MEANS DENGAN BANTUAN METODE ELBOW

Brian Wijaya
Program Studi Teknik Informatika, Teknik Informasi, Universitas Tarumanagara
Jl. Letjen S.Parman No.1, Jakarta Barat, DKI Jakarta, Indonesia 11410
Email**:** *Brian.535200069@stu.untar.ac.id*

ABSTRAK

Banyak faktor seseorang untuk tetap menjaga kesehatan dirinya mulai dari menjaga asupan makanan yang masuk, hingga melakukan kegiatan olahraga rutin. Namun tidak hanya faktor dari dalam yang mempengaruhi kesehatan seseorang, terdapat juga beberapa faktor ekstenal seperti pencemaran udara. Senyawa-senyawa kimia yang menjadi penyebab pencemaran udara yaitu SO_2, CO, O_3 dan lain-lain. Penelitian ini dilakukan menggunakan metode *K-means* sebagai algoritma *clustering* dengan menggunakan data indeks pencemaran udara wilayah DKI Jakarta tahun 2020. Hasil penelitian menunjukkan bahwa metode *elbow* dapat menjadi salah satu metode dalam menentukan jumlah klaster yang akan diteliti selain dengan menggunakan melihat nilai *silhouette coefficient* ataupun dapat dikombinasikan bersama. PM10 dan CO pada visualisasi 3D menjadi faktor penentu dan memiliki pengaruh jika angka keduanya meningkat, maka dipastikan polutan lain akan ikut meningkat juga.

Kata Kunci: Klasterisasi, Pencemaran Udara, *K-Means, Elbow, Silhouette Score*

ABSTRACT

There are many factors that influence a person's ability to maintain their health, ranging from maintaining a healthy diet to engaging in regular exercise. However, it is not only internal factors that affect a person's health; there are also several external factors, such as air pollution. The chemical compounds that cause air pollution include SO2, CO, O3, and others. This study was conducted using the K-means method as a clustering algorithm with data on air pollution indices in the DKI Jakarta area in 2020. The results show that the elbow method can be used to determine the number of clusters to be studied, in addition to using the silhouette coefficient value or a combination of both. PM10 and CO in 3D visualization are determining factors and have an influence; if both numbers increase, it is certain that other pollutants will also increase.

Keywords: *Clustering, Air Polution, K-Means, Elbow, Silhouette Score*

1. PENDAHULUAN

Pencemaran udara menjadi salah satu penyebab kondisi seseorang menjadi kurang sehat. Dikarenakan jika seseorang menghirup udara yang tercermar oleh suatu senyawa tertentu dalam jumlah besar, maka orang tersebut akan mengalami ganguan kesehatan mulai dari penyakit ringan hingga penyakit berat seperti ganguan pernafasan. Masalah pencemaran udara menjadi sulit untuk di tangani terutama di kota-kota metropolitan dengan jumlah penduduk yang tinggi dan aktifitas yang padat. Banyak sekali hal yang menjadi kontribusi penyebab pencemaran udara seperti pembakaran kendaraan bermotor, asap pabrik, pengunaan pendingin ruangan dan masih banyak lagi.

Tujuan dari penerlitian ini adalah mendapatkan pembagian kelompok dari data indeks pencemaran udara yang diberikan untuk dibagi menjadi beberapa kolompok. Pengelompokan yang dimaksut adalah kadar tingginya senyawa kimia yang menyebar di udara di kawasan DKI Jakarta. Dengan demikian hasil dari pengelompokan tesebut dapat digunakan sebagai alat bantu dalam penganalisaan data.

2. METODE

2.1. K-means

Algoritma k-means merukan metode yang paling populer digunakan untuk melakukan pengelompokan[1]. K-means sendiri termasuk kedalam algoritma unsupervised learning, yang dapat diartikan bahwa data set yang digunakan tidak memiliki label penanda kelompok. Biasanya data set yang digunakan metode k-means berbentuk fitur angka atau fitur yang berbentuk simbol. Pada fitur angka perhitungan yang digunakan adalah euclidean distance, sedangkan dalam bentuk simbol menggunakan hamming distance[2]. Dibawah ini merupakan rumus dari euclidean distance.

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2} \quad (1)$$

$d(p,q)$ = dua titik pada dalam euclidean
$q_i,\ p_i$ = Euclidean vector

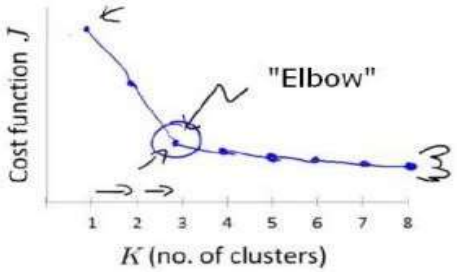
2.2. Metode Elbow

Metode Elbow merupakan metode tertua untuk menentukan jumlah klaster terbaik pada data set [3]. Pengerjaan metode ini menggunakan perhitungan sisa penambahan dari pangkat 2 (*Residual Sum of Squares*) pada percobaan dengan klaster berurut seperti dari 2-kalster sampai 7-kalster.

$$RSS = \sum_{i=1}^{n}(y_i - f(x_i))^2 \quad (2)$$

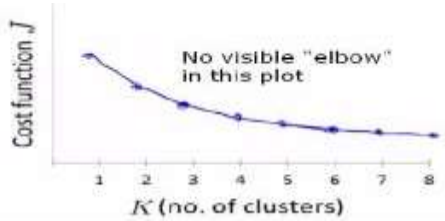
RSS = Residual Sum of Squares
y_i = jumlah dari variabel yang akan di prediksi
$f(x_i)$ = hasil prediksi dari y_i

Dengan bantuan grafik sebagai visualisasi dengan nilai x sebagai jumal n-kalster dan y sebagai hasil dari perhitungan dengan rumus diatas. Jika pada grafik tersebut terlihat bahwa perhitungan tidak lagi terjadi penurunan secara signifikan maka n-kalster tersebut merupakan jumlah kalster terbaik[4]. Di bawah ini merupakan visualisasi dari metode elbow.



Gambar 1. Grafik visualisasi elbow

Namun metode tertua ini memiliki sebuah kekurangan jika hasil grafik tidak memiliki tekukan. Maka metode ini menjadi ambigu atau tidak dapat digunakan. Seperti yang ditunjukan pada gambar 2.

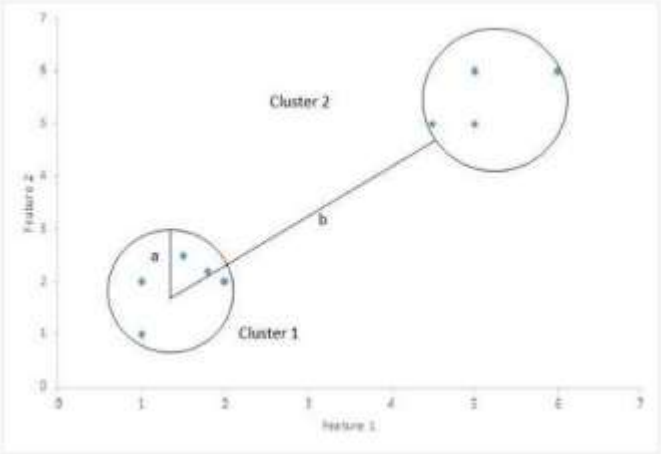


Gambar 2. Elbow yang ambigu

ISSN 1410-7228 (Versi Cetak)
ISSN 2716-4233 (Versi Elektronik)

2.3. *Silhouette Coefficient*

Tujuan dari silhouette coefficient bertujuan untuk memberikan evaluasi pada kalster tertentu yang akan digunakan sebagai perbandingan dengan kalster lain yang digunakan sebagai pengukur kepadatan dan pemisahan kalster [5].



Gambar 3. *Silhouette*

Indikasi (a) merupakn nilai rata-rata jarak setiap titik didalam kalster dan (b) merupakan nilai rata-rata jarak antara semua kalster [6].

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)} \quad (3)$$

s_i = nilai silhouette
a_i = nilai rata rata jarak setiap titik didalam kalster
b_i = nilai rata-rata jarak antar semua kalster

3. HASIL DAN PEMBAHASAN

3.1. Data

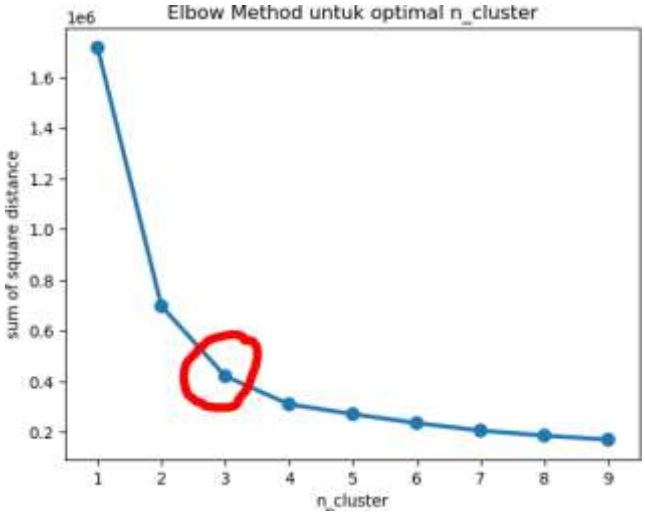
Data yang digunakan pada penelitian ini merupakan data yang didapat dari Satu Data Jakarta dengan nama Indeks Standar Pencemaran Udara (ISPU) tahun 2020 dari bulan September hingga November dengan jumlah data sebesar 455 data. Variable-variable yang terdapat pada data tersebut adalah tanggal, stasiun, PM10, SO_2, CO, O_3, NO_3, max, critical dan categori.

3.2. Pra-Pemrosesan Data

Sebelum data digunakan, perlu dilakukannya screening data yaitu melakukan pengecekan data satu persatu untuk menemukan apakah ada data kosong pada data set tersebut. Jika saat melakukan screening data ditemukan data kosong maka perlu melakukan pembenaran data dengan menggunakan nilai rata rata dari 5 data sebelum data kosong. Jika variabel data yang digunakan memiliki perbedaan satuan angka yang sangat jauh, perlu dilakukannya pengkonversi data dengan menggunakan MinMaxScaler. Pada data set ini variabel PM10 memiliki missing value sebanyak 9 data (1.97%), dan variabel SO_2, CO, O_3 dan NO_3 hanya memiliki data kosong masing-masing 5 data (1.1%). Setelah pengisian selesai, hasil dari pra-pemrosesan data tersebut pada setiap variabel tidak memiliki data kosong. Dan data yang digunakan memiliki variable angka yang saling berdekatan, sehingga tidak diperlukan konversi data menggunakan MinMaxScaler.

3.3. Eksperimen

Pada penelitian ini variabel data set yang digunakan sebagai eksperimen adalah PM10, SO_2, CO, O_3 dan NO_3. Variabel-variabel tersebut akan dipisahkan terlebih dahulu ke sebuah data frame baru. Kemudian dengan menggunakan metode K-means untuk mengelompokkan data variabel tersebut ke dalam pembagian sebanyak n_klaster. Proses pencarian n_klaster dilakukan dengan menggunakan metode *elbow*.

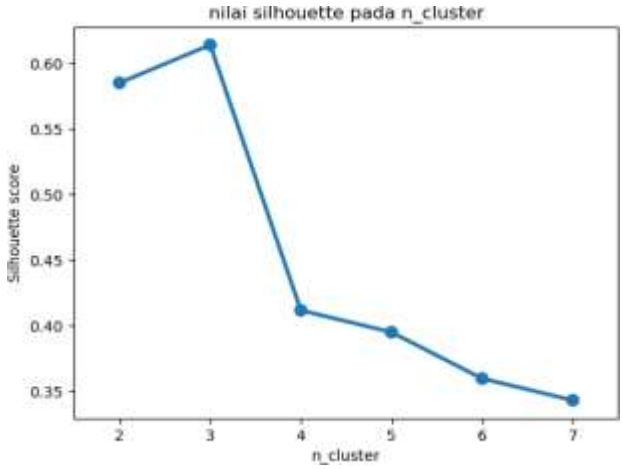


Gambar 4. Grafik elbow penelitian

Dilihat dari hasil *elbow* tersebut klaster dengan jumlah 3 merupakan pembagian terbaik pada data set tersebut. Setelah menumukan n_kalster, *clustering* dapat dilakukan dengan metode *k-means* dengan jumlah klaster sama dengan 3. Kemudian hasil dari klaster tersebut akan dievaluasi dengan menggunakan nilai *silhouette coefficient* untuk mengetahui seberapa baik pembagian klaster tersebut.

3.4. Evaluasi

Evaluasi dilakukan dengan menghitung nilai *silhouette coefficient*. Hasil dari uji coba yang dilakukan dari 2 klaster hingga 7 klaster dapat di lihat seperti pada Gambar 5 berikut ini.



Gambar 5. Grafik hasil perhitungan *silhouette* tiap klaster

Nilai tertinggi terdapat pada klaster berjumlah 3. Hal ini dapat dijadikan sebagai bukti bahwa metode *elbow* yang digunakan tervalidasi. Tabel 1 dibawah ini adalah ringkasan nilai dari perhitungan *silhouette coefficient* tersebut.

ISSN 1410-7228 (Versi Cetak)
ISSN 2716-4233 (Versi Elektronik)

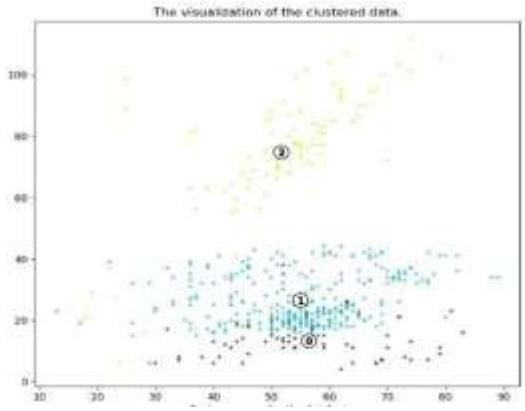
Tabel 1: Tabel nilai *silhouette*

2	0.585064
3	0.614003
4	0.411798
5	0.395119
6	0.359796

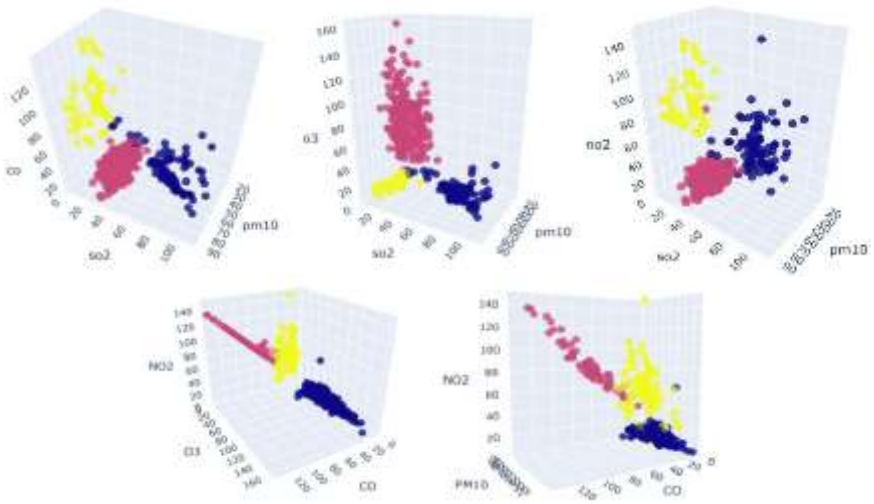
Pada pembagian 6-klaster percobaan diberhentikan dikarenakan nilai rata-rata tersebut mengalami penurunan secara terus-menerus, sehingga proses evaluasi dihentikan dan menggunakan n_klaster sama dengan 3 sebagai nilai pembagian *silhouette* terbaik yaitu sebesar 0.614. Hasil nilai tersebut menandakan pembagian klaster dapat masuk ke dalam kategori baik untuk digunakan pada tahap selanjutnya.

3.5. Analisis

Dari hasil evaluasi penelitian yang dilakukan dengan memvisualisasikan hasil klaster ke dalam 2 dimensi dan 3 dimensi untuk mempermudah dalam proses analisis pembagian klaster tersebut. Gambar 6 dan Gambar 7 menunjukkan hasil visualisasi klaster pada bentuk masing-masing 2D dan 3D.



Gambar 6. Visualisasi kalster 2 dimensi



Gambar 7. Visualisasi klaster 3 dimensi

Dapat di lihat dari hasil visualisasi pada gambar-gambar sebelumnya bahwa titik berwarna biru adalah Klaster 1, Klaster 2 berwarna merah, dan kuning merepresentasikan Klaster ke 3. Tabel 2, Tabel 3, dan Tabel 4 adalah detail pembagian klaster.

Tabel 2. Klaster 1

	pm10	so2	co	o3	no2
count	91.000000	91.000000	91.000000	91.000000	91.000000
mean	51.571429	74.824176	28.461538	12.956044	60.175824
std	12.448242	20.480665	11.012425	7.128834	21.091648
min	18.000000	6.000000	15.000000	3.000000	14.000000
25%	46.000000	68.000000	20.000000	8.000000	48.000000
50%	53.000000	77.000000	23.000000	12.000000	59.000000
75%	58.500000	87.000000	42.000000	16.000000	70.500000
max	79.000000	112.000000	53.000000	34.000000	148.000000

Tabel 3. Klaster 2

	pm10	so2	co	o3	no2
count	303.000000	303.000000	303.000000	303.000000	303.000000
mean	54.871287	26.584158	11.742574	76.336634	9.524752
std	12.027835	7.916285	5.866597	23.297789	4.706472
min	13.000000	6.000000	3.000000	23.000000	2.000000
25%	48.000000	20.000000	7.000000	61.000000	7.000000
50%	56.000000	25.000000	10.000000	75.000000	9.000000
75%	62.000000	33.000000	15.000000	89.000000	11.000000
max	89.000000	44.000000	41.000000	162.000000	63.000000

Tabel 4. Klaster 3

	pm10	so2	co	o3	no2
count	61.000000	61.000000	61.000000	61.000000	61.000000
mean	56.360656	13.262295	83.852459	7.672131	85.606557
std	12.989797	6.112560	22.638342	2.071403	20.939022
min	30.000000	4.000000	42.000000	3.000000	48.000000
25%	46.000000	8.000000	68.000000	6.000000	71.000000
50%	54.000000	13.000000	80.000000	8.000000	80.000000
75%	65.000000	15.000000	98.000000	9.000000	98.000000
max	83.000000	38.000000	135.000000	13.000000	135.000000

Hasil menunjukkan bahwa pembagian yang dilakukan pada klaster pertama nilai tertinggi terdapat pada SO2 dan NO2, PM10 dan O3 memperoleh nilai tertinggi pada klaster kedua, dan pada ketiga CO dan NO2 memiliki nilai tertinggi.

4. KESIMPULAN

K-Means merupakan metode Pengelompokan yang sangat populer dan cukup mudah untuk diimplementasi. Pencarian jumlah klaster terbaik dapat dicari dengan menggunakan metode *elbow* dan sebelum proses klasterisasi dilakukan, perhitungan nilai *silhouette coefficient* menjadi kunci untuk validasi dari hasil yang diperoleh metode *elbow*. Pada hasil klaster diketahui bahwa beberapa kelompok dengan 2 fitur dapat divisualisasikan dengan 2D dan penggunaan visualisasi 3D pada jumlah fitur lebih dari 2. Saran yang dapat diberikan untuk penelitian selajutnya sebaiknya dataset yang digunakan tidak memiliki data kosong untuk mendapatkan nilai pengelompokan terbaik

Dapat disimpulkan juga bahwa dari hasil perhitungan dengan metode *elbow*, n_klaster terbaik adalah sama dengan 3. Penentuan nilai n_klaster tersebut kemudian dievaluasi dan divalidasi lagi dengan menghitung nilai *silhouette coefficient*. Nilai n_klaster sama dengan 3 memiliki nilai *silhouette coefficient* tertinggi yaitu sebesar 0.614, sehingga digunakan dalam proses selanjutnya. Jumlah klaster yang sudah diperoleh kemudian divisualisasikan ke dalam 2D dan 3D. Ditemukan adanya kesamaan pola pergerakan atau peningkatan yang signifikan pada PM10 dan CO yang masuk ke seluruh klaster yang ada.

Hal ini mengindikasikan bahwa setiap kali PM10 atau CO mengalami peningkatan, maka dapat dipastikan polutan lain akan mengalami kenaikan juga. Oleh karena itu, pengaruh kerusakan lingkungan dapat diketahui seiring dengan meningkatnya jumlah polutan yang ada. Sebagai tambahan lagi untuk penelitian selanjutnya adalah menambahkan metode *clustering* yang berbeda dan lebih modern sebagai pembanding karena metode yang digunakan pada penelitian ini merupakan metode yang sudah cukup lama, sehingga diperlukan metode *clustering* yang lebih baru.

ISSN 1410-7228 (Versi Cetak)
ISSN 2716-4233 (Versi Elektronik)

DAFTAR PUSTAKA

[1] Sinaga, K.P. and Yang, M.S., "Unsupervised K-means Pengelompokan algorithm". *IEEE access*, *8*, pp.80716-80727. 2020. [Onlineserial]. Available: . https://web.cse.msu.edu/~cse802/notes/ConstrainedKmeans.pdf [Accessed Dec 4].

[2] Wagstaff, Kiri, Claire Cardie, Seth Rogers, and Stefan Schrödl. "Constrained k-means Pengelompokan with background knowledge." In *Icml*, vol. 1, pp. 577-584. 2001. [Online serial]. Available: https://web.cse.msu.edu/~cse802/notes/ConstrainedKmeans.pdf [Accessed Dec 4].

[3] Kodinariya, T.M. and Makwana, "P.R., Review on determining number of Kalster in K-Means Pengelompokan" *International Journal*, *1*(6), pp.90-95. 2013. [Online serial]. Available: https://www.researchgate.net/profile/Trupti-Kodinariya/publication/313554124 Review on Determining of Kalster in_K-means_Pengelompokan/links/5789fda408ae59aa667931d2/Review-on-Determining-of-Kalster-in-K-means-Pengelompokan.pdf [Accessed Dec 5].

[4] Syakur, M.A., Khotimah, B.K., Rochman, E.M.S. and Satoto, B.D., 2018, April. "Integration k-means Pengelompokan method and elbow method for identification of the best customer profile kalster". In *IOP conference series: materials science and engineering* (Vol. 336, No. 1, p. 012017). [Online serial]. Available: https://iopscience.iop.org/article/10.1088/1757-899X/336/1/012017/pdf [Accessed Dec 5].

[5] Layton, Robert, Paul Watters, and Richard Dazeley. "*Evaluating authorship distance methods using the positive Silhouette coefficient*." *Natural Language Engineering* 19, no. 4 (2013): 517-535. [Online serial]. Available: https://core.ac.uk/download/pdf/213010094.pdf [Accessed Dec 7].

[6] Dinh, Duy-Tai, Tsutomu Fujinami, and Van-Nam Huynh. "*Estimating the optimal number of kalsters in categorical data Pengelompokan by silhouette coefficient*." In *International Symposium on Knowledge and Systems Sciences*, pp. 1-17. Springer, Singapore, 2019. [Online serial]. Available: https://drive.google.com/file/d/1bCuA8b7I4wCNZkWYSOK6v6vsnRwk9C7I/view [Accessed Dec 7].