ISSN: 2339-2541
JURNAL GAUSSIAN, Volume 11, Nomor 2, Tahun 2022, Halaman 228 - 236
Online di: https://ejournal3.undip.ac.id/index.php/gaussian/

KERNEL K-MEANS *CLUSTERING* UNTUK PENGELOMPOKAN SUNGAI DI KOTA SEMARANG BERDASARKAN FAKTOR PENCEMARAN AIR

Anestasya Nur Azizah[1*], Tatik Widiharih[2], Arief Rachman Hakim[3]
[1,2,3] Departemen Statistika, Fakultas Sains dan Matematika, Universitas Diponegoro
***e-mail**: anestasyanurazizah@gmail.com

ABSTRACT

K-Means Clustering is one of the types of non-hierarchical cluster analysis which is frequently used, but has a weakness in processing data with non-linearly separable (do not have clear boundaries) characteristic and overlapping cluster, that is when visually the results of a cluster are between other clusters. The Gaussian Kernel Function in Kernel K-Means Clustering can be used to solve data with non-linearly separable characteristic and overlapping cluster. The difference between Kernel K-Means Clustering and K-Means lies on the input data that have to be plotted in a new dimension using kernel function. The real data used are the data of 47 rivers and 18 indicators of river water pollution from Dinas Lingkungan Hidup (DLH) of Semarang City in the first semester of 2019. The cluster results evaluation is used the Calinski-Harabasz, Silhouette, and Xie-Beni indexes. The goals of this study are to know the step concepts and analysis results of Kernel K-Means Clustering for the grouping of rivers in Semarang City based on water pollution factors. Based on the results of the study, the cluster results evaluation show that the best number of clusters K=4.

***Keywords**: Gaussian Kernel, Kernel K-Means Clustering, Cluster Results Evaluation*

1. PENDAHULUAN

Analisis kelompok atau yang dikenal dengan *cluster* merupakan salah satu analisis statistik yang dapat memisahkan data atau observasi ke dalam sejumlah *cluster* menurut karakteristiknya masing-masing. Salah satu metode analisis *cluster* yang sering digunakan adalah K-Means, dimana memiliki kelemahan dalam memproses data yang bersifat *non-linearly separable*. Data yang bersifat *non-linearly separable* adalah data yang tidak dapat terpisah secara linear dan tidak dapat dibentuk garis pemisah atau pembatas yang jelas. Cara untuk mengetahui *non-linearly separable* pada data adalah dengan melihat pada plot hasil akhir *cluster* ada tidaknya objek yang tumpang tindih atau *overlap*. Contoh data *overlap* pada *cluster* adalah saat objek yang dikategorikan sebagai anggota suatu *cluster* terletak diantara *cluster* lainnya, sehingga tidak dapat dibuat garis pembatas linear.

Konsep Kernel K-Means *Clustering* sebagai penanganan permasalahan data yang *non-linearly separable* dan *overlap* dengan memetakan data ke *feature space* atau ruang fitur menggunakan kernel. Pemetaan data dengan kernel dijadikan sebagai variabel yang dicari kelasnya menggunakan K-Means (Aprianto, 2018). Perluasan dari metode K-Means ke dalam Kernel K-Means *Clustering* juga direalisasikan melalui pernyataan jarak dalam bentuk fungsi kernel (Girolami, 2002). Jarak dalam *cluster* yang biasa digunakan antara lain Euclidean, namun dengan perhitungan *kernel trick* mengakibatkan adanya pengkuadratan pada jarak Euclidean yang akhirnya disebut dengan jarak *Square* Euclidean.

Metode Kernel K-Means *Clustering* (KKC) yang digunakan adalah Kernel Gaussian, atau dikenal dengan Kernel Radial Basis Function (RBF) Gaussian karena memiliki sifat yang fleksibel dalam penggunaannya. Menurut Indraswari *et al.* (2017), Kernel RBF Gaussian adalah kernel yang secara umum dapat digunakan untuk semua jenis data. Evaluasi hasil *cluster* yang digunakan adalah indeks Calinski-Harabasz, Silhouette, dan Xie-Beni.

Penelitian yang berhubungan dengan Kernel K-Means *Clustering* dilakukan oleh Maysaroh (2015) tentang Kernel K-Means *Clustering* pada faktor-faktor risiko penyebab

penyakit hipertensi di Indonesia. Penelitian ini dilakukan analisis *cluster* pada 47 sungai di Kota Semarang berdasarkan 18 faktor pencemaran air sungai dan data penelitian diperoleh dari Dinas Lingkungan Hidup Kota Semarang.

2. TINJAUAN PUSTAKA

Cluster adalah pengelompokan data berdasarkan kesamaan karakteristik antar objek. Saat ini, menemukan data dengan karakteristik serupa dan memberi label yang sesuai menjadi salah satu tantangan terbesar dalam aplikasi analisis data.

Indeks *Kaiser-Mayer-Olkin* (KMO) digunakan untuk mengetahui ketepatan penggunaan analisis faktor. H_0 diterima apabila nilai KMO antara 0,5 sampai 1, maka dapat disimpulkan analisis faktor tepat digunakan (Bilson, 2005).

Hipotesis:

H_0 : Jumlah data cukup untuk difaktorkan

H_1 : Jumlah data tidak cukup untuk difaktorkan

Statistik uji:

$$KMO = \frac{\sum_{i=1}^{P}\sum_{j=1}^{P} r_{X_i X_j}^2}{\sum_{i=1}^{P}\sum_{j=1}^{P} r_{X_i X_j}^2 + \sum_{i=1}^{P}\sum_{j=1}^{P} a_{X_i X_j}^2} \quad (1)$$

i $= 1,2,3,\ldots,P$ dan $j = 1,2,3,\ldots,P$

P = jumlah variabel

$r_{X_i X_j}^2$ = koefisien korelasi antara variabel X_i dan X_j

$a_{X_i X_j}^2$ = koefisien korelasi parsial antara variabel X_i dan X_j

Multikolinearitas digunaan untuk mengetahui ada atau tidaknya hubungan antara satu variabel prediktor dengan variabel prediktor yang lain. Cara untuk mendeteksi adanya kasus kolinearitas menurut Gujarati (2004) dilihat melalui *Variance Inflation Factors* (VIF) yang dinyatakan sebagai berikut:

$$VIF_p = \frac{1}{1-R_p^2} \quad (2)$$

R_p^2 merupakan koefisien determinasi antara X_p dengan variabel prediktor lainnya. VIF_p yang lebih besar dari 10 menunjukkan adanya kolinieritas antar variabel prediktor. Jika terjadi multikolinearitas, maka ditangani dengan melakukan *Principal Component Analysis* (PCA).

Jarak antar objek merupakan faktor yang sangat berpengaruh terhadap hasil *cluster* yang dibentuk. Cara untuk menghitung jarak antar objek pada Kernel K-Means *Clustering* adalah dengan metode jarak *Square* Euclidean.

Jarak *Square* Euclidean merupakan jarak yang dikembangkan dari jarak Euclidean. Jarak *Square* Euclidean adalah jumlah kuadrat perbedaan deviasi di dalam nilai untuk setiap variabel (Hair *et al*., 2010). Rumus untuk jarak *Square* Euclidean adalah sebagai berikut:

$$d_{ij} = \sum_{p=1}^{P}\left(x_{ip} - x_{jp}\right)^2 \quad (3)$$

i, j $= 1,2,3,\ldots,N$

p (variabel) $= 1,2,3,\ldots,P$

N = banyaknya data pada variabel ke-p

d_{ij} = jarak antara objek ke-i dan objek ke-j

x_{ip} = data dari objek ke-i pada variabel ke-p

x_{jp} = data dari objek ke-j pada variabel ke-p

Standarisasi data biasanya dilakukan ketika dalam suatu penelitian terdapat variabel yang memiliki satuan berbeda dengan variabel lainnya dan mengubahnya hingga seluruh

data menjadi standar. Dalam kasus analisis *cluster* perlu dilakukan standarisasi dengan mengubah ke bentuk *z-score* (nilai standar).

Z-score dapat dihitung dengan rumus berikut:

$$Z_{np} = \frac{X_{np} - \overline{X_p}}{S_p} \quad (4)$$

P = jumlah variabel
N = banyak data dalam variabel ke-p
Z_{np} = nilai *Z-score* pada data ke-n variabel ke-p
S_p = simpangan baku atau standar deviasi variabel ke-p
X_{np} = data ke-n variabel ke-p
$\overline{X_p}$ = rata-rata variabel ke-p

Metode K-Means merupakan salah satu metode *cluster* non hierarki. Pada K-Means, misal didefinisikan $X = \{x_1, x_2, x_3, \dots, x_n\}$ adalah sebuah himpunan data dalam ruang berdimensi D, yang dinotasikan R^D, sedangkan k adalah sebuah bilangan integer positif lebih dari satu. Apabila $X_n \in R^D$, maka algoritma K-Means *Clustering* akan mempartisi ke dalam K *cluster* serta dapat dinyatakan dengan himpunan $X_1, X_2, X_3, \dots, X_k$ yang saling lepas, sehingga $X_1 \cup X_2 \cup X_3 \cup \dots, \cup X_k = X$, dimana setiap *cluster* memiliki nilai tengah (*centroid*) dari data-data dalam *cluster* tersebut (Maysaroh,2015).

Algoritma K-Means secara acak menentukan K buah data sebagai titik tengah (*centroid*), kemudian dihitung jarak antara data dengan *centroid*, untuk selanjutnya data akan ditempatkan ke dalam *cluster* yang terdekat dihitung dari titik tengah *cluster*. Titik tengah (*centroid*) *cluster* didefinisikan sebagai:

$$m_k = \frac{1}{N_k} \sum_{X_n \in C_k} X_n \quad (5)$$

Penggunaan metode kernel dikenal dengan "*kernel trick*", yang memungkinkan penggabungan fungsi pemetaan dengan beberapa fungsi hasil kali dalam (*inner product*) (Cristianini dan Taylor, 2000).

Fungsi kernel dapat didefinisikan suatu fungsi K dimana untuk semua vektor input $\boldsymbol{x_i}, \boldsymbol{x_j}$ akan memenuhi kondisi:

$$K(\boldsymbol{x_i}, \boldsymbol{x_j}) = \emptyset(\boldsymbol{x_i})^T \cdot \emptyset(\boldsymbol{x_j}) \quad (6)$$

Dengan $\emptyset(.)$ adalah fungsi pemetaan dari ruang input (*input space*) ke ruang fitur (*feature space F*), atau secara matematis $\emptyset: \boldsymbol{x} \rightarrow \emptyset(\boldsymbol{x}) \in F$. Fungsi kernel memungkinkan untuk mengimplementasikan suatu model pada ruang berdimensi lebih tinggi (ruang fitur) tanpa harus mendefinisikan fungsi pemetaan dari ruang input ke ruang fitur, sehingga untuk kasus yang *non-linearly separable* pada ruang input diharapkan akan menjadi *linearly separable* pada ruang fitur dengan menggunakan *hyperplane* sebagai *decision boundary* secara efisien (Maysaroh, 2015).

Menurut Cristiani *et al*. (2004) dalam Murfi (2009), fungsi kernel ada 4 macam :

a. Kernel Linier

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{x}_i{}^T \boldsymbol{x}_j \quad (7)$$

b. *Polynomial Kernel*

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\gamma \boldsymbol{x}_i{}^T \boldsymbol{x}_j + r)^d \ , \gamma > 0 \quad (8)$$

c. Fungsi Kernel Gaussian (*Radial Basis Function*)

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = exp\left\{-\gamma \left\| \boldsymbol{x}_i - \boldsymbol{x}_j \right\|^2\right\} \quad (9)$$

d. Eksponensial (sigmoid) Kernel

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = tanh(\gamma \boldsymbol{x}_i{}^T \boldsymbol{x}_j + r) \quad (10)$$

dimana γ, r, d = parameter kernel

i, j = 1,2,3, … , n

Istilah Gaussian sering dibahas di dalam statistika, contohnya Distribusi Gaussian. Distribusi Gaussian memiliki nama lain distribusi normal dan digambarkan dengan kurva distribusi yang berbentuk lonceng. Distribusi Gaussian memiliki dua parameter, yaitu μ (rata-rata) data, dan σ (standard deviasi atau simpangan baku) (Han dan Kamber, 2006).

Kernel Gaussian merupakan fungsi kernel yang biasa digunakan karena sifatnya yang fleksibel dan mampu dapat digunakan untuk semua jenis data. Kernel Gaussian tidak memiliki syarat-syarat khusus dalam proses pengerjaannya, karena itu dinilai lebih mudah dalam perhitungan dan lebih sederhana untuk komputasinya.

Menurut Aprianto (2018), kernel digunakan untuk mengatasi permasalahan dimensi, dimana kita dapat mendefinisikan kernel seperti pada persamaan di bawah ini:

$$K(x_i, x_j) = exp\left(-\frac{(x_i, x_j)^2}{2h^2}\right) \quad (11)$$

dengan masing-masing x_i dan x_j adalah vektor dan $(x_i, x_j)^2$ dikenal dengan istilah jarak *Square* Euclidean antara kedua vektor. Sedangkan h merupakan parameter bebas yang ditentukan di awal.

Jika $\gamma = \frac{1}{2h^2}$, maka persamaan yang lebih sederhana dari kernel ditunjukkan sebagai persamaan berikut:

$$K(x_i, x_j) = exp\left(-\gamma(x_i, x_j)^2\right) \quad (12)$$

Kernel K-Means *Clustering*, pada prinsipnya mirip dengan K-Means tradisional, letak perbedaan yang mendasar ada pada perubahan masukannya. Data poin pada Kernel K-Means akan dipetakan pada dimensi baru yang lebih tinggi menggunakan fungsi non linear sebelum dilakukan analisis *cluster* (Cristianini & Taylor, 2000).

Kernel K-Means *Clustering* yang digunakan berdasarkan metode Kernel Gaussian. Sebelum dilakukan analisis *cluster*, dibentuk matriks *feature space* dengan ukuran ($N \times N$) yang disimbolkan $K(\boldsymbol{X^T}, \boldsymbol{X})$. Setiap elemen dari matriks *feature space* dinotasikan $K(x_i, x_j)$, dimana i adalah baris dan j adalah kolom, dihitung menggunakan Kernel Gaussian seperti persamaan (11). Setelah itu memilih sejumlah K titik dari N data untuk menghitung jarak masing-masing *cluster* sebagai pusat awal. Jarak pada *kernel space* dapat dihitung menggunakan trik jarak kernel seperti persamaan berikut:

$$\begin{aligned}\left\|\emptyset(x_i) - \emptyset(x_j)\right\|^2 &= \left(\emptyset(x_i) - \emptyset(x_j)\right).\left(\emptyset(x_i) - \emptyset(x_j)\right) \\ &= \emptyset(x_i).\emptyset(x_i) + \emptyset(x_j).\emptyset(x_j) - 2\emptyset(x_i).\emptyset(x_j) \\ \left\|\emptyset(x_i) - \emptyset(x_j)\right\|^2 &= K(x_i, x_i) + K(x_j, x_j) - 2K(x_i, x_j) \quad (13)\end{aligned}$$

Data ke-n akan menjadi anggota *cluster* ke-k (C_k) apabila hasil perhitungan jaraknya adalah yang paling minimum setelah dihitung dengan seluruh pusat kelompok. *Cluster* yang mendapat anggota baru akan dihitung kembali nilai tengahnya. Jarak baru antara data ke- n dengan pusat *cluster* ke-k pada *kernel space* dijelaskan pada rumus di bawah ini:

$$\|\emptyset(x_n) - \emptyset(m_k)\|^2 = K(x_n, x_n) + K(m_k, m_k) - 2K(x_n, m_k) \quad (14)$$

$C^*(X_n)$ adalah simbol nilai minimum jarak pada *kernel space* serta untuk memutuskan data ke-n masuk pada *cluster* (C_k) dengan jarak terdekat. $C^*(X_n)$ dijabarkan sebagai berikut:

$$C^*(X_n) = \arg min_i(\|\emptyset(x_n) - \emptyset(m_k)\|^2) \quad (15)$$

Pusat Kernel K-Means *CLustering* didefinisikan sebagai:

$$m_k = \frac{1}{N_k}\sum_{X_n \in C_k} \emptyset(X_n) \quad (16)$$

Pendekatan yang umum dipakai untuk mengevaluasi kebaikan dari hasil analisis *cluster* dikenal dengan *cluster validation* (Maulik dan Bandyopadhyay, 2002). *Cluster Validation* yang akan dibahas dibatasi pada validitas *cluster* dengan pendekatan *internal clustering validation*, yaitu evaluasi hasil *cluster* tanpa informasi dari luar, dalam hal ini berdasarkan seberapa dekat jarak di dalam *cluster* dan jarak antar *cluster*.

Beberapa jenis *internal clustering validation* yang digunakan pada penelitian Liu *et al*. (2010) adalah:

1. *Calinski-Harabasz Index*

Indeks ini mengukur jumlah kuadrat antar *cluster* dan jumlah kuadrat di dalam *cluster*. Penentuan banyaknya jumlah *cluster* (K) adalah dengan menggunakan nilai maksimum dari nilai *Calinski-Harabasz* CH_K dapat dilihat dari rumus berikut:

$$CH_K = \frac{SSB}{SSW} \times \frac{N-K}{K-1} \tag{17}$$

dimana, K = banyaknya *cluster*, SSB= jumlah kuadrat jarak antar *cluster*, dan SSW= jumlah kuadrat jarak di dalam *cluster*.

2. *Silhouette Index*

Silhouette Index biasa disimbolkan dengan S_{index} digunakan dengan menghitung rata-rata ketidakmiripan antar titik dalam *cluster* yang sama dan *cluster* yang berbeda.

Silhouette Index dihitung dengan:

$$S_{index} = \frac{b_{(i)} - a_{(i)}}{max\{a_{(i)}, b_{(i)}\}} \tag{18}$$

dimana,

$a_{(i)}$ adalah rata-rata jarak data ke-i terhadap semua data lainnya dalam satu *cluster*, dan $b_{(i)}$ adalah hasil rata-rata jarak data ke-i terhadap semua data dari *cluster* lain yang kemudian diambil data yang paling kecil.

3. *Xie-Beni Index*

Indeks *Xie-Beni* juga bertujuan untuk menghitung rasio total variasi di dalam *cluster* dan pemisahan *cluster* (Widiyanto, 2019). Nilai *Xie-Beni* yang rendah mengindikasikan partisi *cluster*yang lebih baik. Indeks *Xie-Beni* dituliskan sebagai berikut:

$$XB_{(K)} = \frac{S}{d_{min}} \tag{19}$$

Matriks $\boldsymbol{V}$ berukuran $K \times P$

$v_{kp} = \frac{\sum_{n=1}^{N} (u_{nk})^m x_{ip}}{\sum_{n=1}^{N} (u_{nk})^m}$ $\quad ; i = 1,2, \dots, N$

Membangkitkan data random $\boldsymbol{U}$ dengan ukuran $N \times K$ dengan syarat $\sum_{k=1}^{K} u_{nk} = 1$, sehingga $u_{nk} = \left[\frac{[\sum_{p=1}^{P} d_{Pk}]^{\frac{1}{m-1}}}{\sum_{k=1}^{K} [\sum_{p=1}^{P} d_{pk}]^{\frac{1}{m-1}}} \right]^{-1}$ dan $d_{pk} = \left(x_{ip} - v_{kp}\right)^2$

$S = \frac{\sum_{n=1}^{N} \sum_{k=1}^{K} (u_{nk})^m d_{nk}(x_n, m_k)}{N}$ $\quad ; \; d_{min} = \underset{i \neq p}{min} d_{ip}(m_i, m_p)$ $\quad ; i, p = k = 1,2, \dots, K$

dimana, N = banyak objek penelitian, K= banyak *cluster*, P= banyak variabel, x_{ip}= data ke-i variabel ke-p, u_{nk}= nilai keanggotaan objek ke-n dengan pusat *cluster* ke-k, m= pangkat pembobot $m > 1$, $\|x_n - m_k\|$ = jarak euclidean titik data (x_n) dengan pusat *cluster* (m_k).

Daya tampung beban pencemaran adalah kemampuan air pada suatu sumber air,untuk menerima masukan beban pencemaran tanpa mengakibatkan air tersebut menjadi tercemar. Beberapa sumber pencemaran air yang sering dijumpai pada masyarakat berupa limbah rumah tangga, limbah industri, limbah pertanian, pertambangan minyak lepas pantai, kebocoran minyak tanker,dan lain sebagainya.

Aspek kimia-fisika dari air yang umum diuji oleh Dinas Lingkungan Hidup (DLH) Kota Semarang untuk menentukan tingkat pencemaran air adalah *Chemical Oxygen Demand* (COD), suhu, residu terlarut, residu tersuspensi, oksigen terlarut, pH, kadmium, tembaga, seng, timbal, besi, sulfida, mangan, nitrat, nitrit, amoniak bebas, sulfat dan klorida.

3. METODE PENELITIAN

Data yang digunakan dalam penelitian ini adalah data sekunder tentang karakteristik lokasi sungai dan indikator pencemaran air sungai secara kimia yang diperoleh dari Dinas Lingkungan Hidup (DLH) Kota Semarang pada semester pertama tahun 2019. Data yang terkumpul berupa sampel dari 47 sungai dan terdapat 18 variabel diambil oleh Unit Pelaksana Teknis Daerah (UPTD) Laboratorium Lingkungan DLH Kota Semarang.

Variabel yang digunakan pada penelitian ini adalah 18 indikator pencemaran air sungai yang diperoleh dari Dinas Lingkungan Hidup (DLH) Kota Semarang. X_1 = *Chemical Oxygen Demand* (COD), X_2 = Suhu, X_3 = Residu Terlarut, X_4 = Residu Tersuspensi, X_5 = Oksigen Terlarut, X_6 = pH, X_7 = Kadmium, X_8 = Tembaga, X_9 = Seng, X_{10} = Timbal, X_{11} = Besi, X_{12} = Sulfida, X_{13} = Mangan, X_{14} = Nitrat, X_{15} = Nitrit, X_{16} = Amoniak Bebas, X_{17} = Sulfat, dan X_{18} = Klorida.

Software yang digunakan untuk pengolahan data ini adalah RStudio versi 1.1.463. Setelah data diperoleh maka dilakukan analisis data menggunakan metode Kernel K-Means dengan tahapan sebagai berikut:

1. Memasukkan data untuk pengolahan statistik deskriptif.
2. Melakukan standarisasi data.
3. Melakukan uji KMO.
4. Melakukan pemeriksaan multikolinearitas.
5. Memasukkan jumlah banyaknya *cluster* atau $K = 2,3,4,5$ dengan inisial *cluster* adalah C_1, C_2, C_3, C_4
6. Menginput Kernel matriks K berukuran $N \times N$.
7. Menghitung pusat awal (m_k).
8. Menghitung jarak baru (δ_{kn}) , untuk semua nilai n pada semua pusat *cluster* (m_k).
9. Menentukan nilai $C^*(X_n)$ untuk memutuskan data ke-n masuk pada *cluster* (C_k) dengan jarak terdekat.
10. Memperbarui $C_k = \{\{X_n\} | C^*(X_n) = k\}$ sampai semua nilai C_k konvergen
11. Menghitung nilai indeks validitas.
12. Menentukan jumlah *cluster* terbaik dari nilai indeks validitas *cluster*.
13. Menginterpretasikan profil *cluster* yang terbentuk.

4. HASIL DAN PEMBAHASAN

Hasil analisis deskriptif menunjukkan dari 18 variabel, yang memiliki rata-rata tertinggi adalah variabel residu terlarut (X3) sebanyak 502 mg/l dan nilai tertingginya adalah 996 mg/l. Parameter yang diutamakan dalam pencemaran air yaitu COD (X1) dengan rata-rata sebesar 99,78 mg/l, nilai paling tinggi sebesar 991 mg/l dan nilai paling rendah 20,42 mg/l.

Parameter pencemaran air sungai di Kota Semarang menunjukkan 2 dari 18 menunjukkan satuan yang berbeda, yaitu suhu dengan derajat celcius, pH tidak mempunyai notasi satuan serta untuk satuan untuk keenambelas parameter lainnya adalah mg/L. Karena terdapat perbedaan satuan di dalam analisis, maka dilakukan standarisasi data dengan mengubah ke bentuk *z-score* (nilai standar).

Hipotesis uji KMO adalah sebagai berikut:

H_0 : Jumlah data cukup untuk difaktorkan

H_1 : Jumlah data tidak cukup untuk difaktorkan

Hasil olah data menujukkan nilai KMO keseluruhan sebesar 0,7 , karena nilai tersebut berkisar diantara 0,5 sampai 1 sehingga H_0 diterima. Jadi, dapat disimpulkan bahwa sampel yang ada mewakili populasi atau *sample representatif* terpenuhi.

Ringkasan hasil pemeriksaan multikolinearitas atau nilai VIF berdasarkan *output* dapat dilihat pada Tabel 1.

Tabel 1. Ringkasan Nilai VIF Variabel Penelitian

Variabel	Nilai VIF	Variabel	Nilai VIF	Variabel	Nilai VIF
X1	7,592026	**X7**	1,342442	**X13**	5,720127
X2	1,528445	**X8**	5,667627	**X14**	3,654007
X3	4,818741	**X9**	1,858269	**X15**	3,363445
X4	4,840455	**X10**	1,576857	**X16**	1,837905
X5	2,646161	**X11**	6,271290	**X17**	6,848046
X6	2,072980	**X12**	1,293460	**X18**	2,901177

Hasil uji multikolinieritas menunjukkan bahwa nilai VIF ada pada rentang 1,293460 hingga 7,592026 atau tidak ada yang melebihi 10, sehingga tidak terjadi multikolinearitas antara variabel.

Ringkasan hasil *final cluster* adalah, untuk 2 *cluster* menghasilkan jumlah *cluster* adalah 21 dan 26 sungai. Pembagian sungai untuk 3 *cluster* adalah 16, 15, dan 16 sungai. Pembagian sungai untuk 4 *cluster* adalah 12, 14, 11, dan 10 sungai. Pembagian sungai untuk 5 *cluster* adalah 12, 8, 9, 10, dan 8 sungai.

Indeks validitas kelompok yang akan digunakan untuk menentukan jumlah kelompok dalam penelitian ini menggunakan indeks validitas Calinski-Harabasz, Silhouette, dan Xie-Beni.

Tabel 3. Hasil Evaluasi *Cluster*

Jumlah Kelompok	Calinski-Harabasz	Silhouette	Xie-Beni
2	**6,92751**	**0,1265328**	4,790146
3	6,107342	0,0847007	2,585736
4	5,203373	0,08423381	**2,031633**
5	4,978457	0,05959548	2,421528

Dari Tabel 3 memperlihatkan 3 indeks validitas yang mengevaluasi hasil *cluster* menunjukkan jumlah hasil *cluster* terbaik berbeda-beda, maka hasil akhir diputuskan dari indeks yang memiliki kinerja lebih baik. Diputuskan bahwa jumlah *cluster* yang baik untuk pencemaran air sungai di Kota Semarang adalah K=4 dari indeks validitas Xie-Beni, karena memiliki penanganan yang lebih lengkap dari indeks validitas lainnya.

Setelah mendapatkan hasil *cluster*, dilakukan pengurutan anggota setiap *cluster* dan menghitung rata-rata setiap variabel di setiap *cluster* menggunakan data asli atau sebelum dilakukan standarisasi.

Hasil untuk jumlah *cluster* sebanyak 4 beranggotakan 12 sungai untuk *cluster* 1 (25,53%), total 14 sungai untuk *cluster* 2 (29,79%), 11 sungai untuk *cluster* 3 (23,40%), dan sebanyak 10 sungai (21,28%) untuk *cluster* 4. Pada 4 *cluster* ini, *cluster* 1 dan *cluster* 2 merupakan sungai di Kota Semarang yang memiliki faktor pencemaran air yang lebih

tinggi dari *cluster* lainnya. Terdapat 6 variabel pada *cluster* 1 yang memiliki nilai lebih tinggi dari *cluster* lainnya adalah residu terlarut (X_3), pH (X_6), timbal (X_{10}), sulfida (X_{12}), mangan (X_{13}), dan klorida (X_{18}). *Cluster* 2 yang memiliki nilai lebih tinggi dari *cluster* lainnya adalah residu tersuspensi (X_4), tembaga (X_8), seng (X_9), besi (X_{11}), nitrat (X_{14}), dan nitrit (X_{15}).

Cluster 3 memiliki 3 variabel dimana rata-rata kandungan faktor pencemaran air pada *cluster* ketiga yang lebih tinggi dibandingkan pada *cluster* lainnya yaitu, oksigen terlarut (X_5), kadmium (X_7), dan amoniak bebas (X_{16}). *Cluster* 4 juga memiliki 3 variabel dimana rata-rata kandungan faktor pencemaran air pada *cluster* keempat yang lebih tinggi dibandingkan pada *cluster* lainnya yaitu *Chemical Oxygen Demand* atau COD (X_1), suhu (X_2) dan sulfat (X_{17}).

Berdasarkan *output* hasil *cluster* terbaik $K = 4$ terhadap 47 sungai di Kota Semarang dan jumlah anggota masing-masing *cluster* adalah 12, 14, 11 dan 10.

Tabel 4. Anggota per *Cluster* dengan K=4

No	***Cluster* 1**	***Cluster* 2**	***Cluster* 3**	***Cluster* 4**
1	Asin	Babon V	Beringin Hulu	Babon I
2	Bajak Hulu	Banjir Kanal Barat Hilir	Kaligarang Hulu	Babon II
3	Banger Hilir	Beringin Hilir	Karangayu Hulu	Banjir Kanal Timur Hilir
4	Banger Hulu	Candi Hulu	Kreo	Candi Hilir
5	Kaligawe Hilir	Mangkang Hilir	Kripik Hilir	Kedungmundu Hulu
6	Karanganyar Hilir	Pedurungan Hilir	Kripik Hulu	Majapahit Hulu
7	Karanganyar Hulu	Plumbon Hilir	Mangkang Hulu	Pedurungan Hulu
8	Siangker Hilir	Plumbon Hulu	Segoro Hilir	Semarang Hilir
9	Silandak Hulu	Ronggolawe Hulu	Segoro Hulu	Tandang Hilir
10	Tambakharjo Hilir	Semarang Hulu	Siangker Hulu	Tugurejo Hulu
11	Tenggang Hilir	Silandak Hilir	Tapak Hulu	
12	Tenggang Hulu	Sringin Hilir		
13		Sringin Hulu		
14		Tugurejo Hilir		

5. KESIMPULAN

1. Variabel faktor pencemaran yang memiliki rata-rata paling tinggi adalah variabel residu terlarut (X3) dengan kandungan sebanyak 502 mg/l dan nilai tertinggi adalah 996 mg/l.
2. Jumlah *cluster* terbaik untuk metode Kernel K-Means *Clustering* adalah menggunakan ukuran kebaikan Calinski-Harabasz, Silhouette dan Xie-Ben adalah sebanyak 2 dan 4 *cluster*.
3. Hasil untuk jumlah *cluster* sebanyak 4 beranggotakan 12 sungai untuk *cluster* 1, total 14 sungai untuk *cluster* 2, 11 sungai untuk *cluster* 3, dan sebanyak 10 sungai untuk *cluster* 4. Pada 4 *cluster* ini, *cluster* 1 dan *cluster* 2 merupakan sungai di Kota

Semarang yang memiliki faktor pencemaran air yang lebih tinggi karena terdapat 6 variabel yang memiliki nilai lebih tinggi dari *cluster* lainnya.

DAFTAR PUSTAKA

Aprianto, K. 2018. Optimasi Kernel K-Means dalam Pengelompokan Kabupaten/Kota Berdasarkan Indeks Pembangunan Manusia di Indonesia. *Journal of Mathematics and Its Applications* Vol. 15, No. 1, Hal: 1-15.

Bilson, S. 2005. *Analisis Multivariat Pemasaran*. Jakarta: Gramedia Pustaka Utama.

Cristianini, N., dan Taylor, J.S. 2000. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. United Kingdom: Cambridge University Press.

Girolami, M. 2002. Mercer Kernel-Based Clustering in Feature Space. *Journal Transaction on Neural Networks* Vol.13, No 3, Hal: 780-784.

Gujarati, D. N. 2004. *Basic Econometrics* (4th ed.). New York: The McGraw-Hill.

Hair, J. F. Jr., Black, W. C., Barry, J. B., dan Anderson, R. E. 2010. *Multivariate Data Analysis Seventh Edition*. New Jersey: Pearson Education.

Han, J., dan Kamber, M. 2006. *Data Mining: Concepts and Techniques Second Edition*. San Fransisco: Morgan Kauffmann.

Indraswari, R., Arifin, A. Z., dan Herumurti, D. 2017. RBF Kernel Optimization Method With Particle Swarm Optimization On SVM Using The Analysis Of Input Data's Movement. *Journal of Computer Science and Information* Vol. 10, No. 1, Hal: 36-42.

Liu, Y., Li, Z., Xiong, H., Gao, X., dan Wu, J. 2010. Understanding of Internal Clustering Validation Measures. *Proceeding of IEEE International Conference on Data Mining*. IEEE New York: 13-17 Desember 2010.

Maulik, U., dan Bandyopadhyay, S. 2002. Performance Evaluation of Some Clustering Algorithms and Validity Indices. *Journal: IEEE Transactions On Pattern Analysis And Machine Intelligence* Vol. 24, No. 12, Hal: 1650-1654.

Maysaroh, S. 2015. Analisis Kelompok Dengan Metode Kernel K-Means (Studi Kasus Pengelompokan Kabupaten/Kota di Indonesia Berdasarkan Penduduk Dengan Faktor-Faktor Risiko Penyebab Penyakit Hipertensi). *Tesis*. Program Studi Statistika Fakultas Matematika dan Ilmu Pengetahuan Alam Institut Teknologi Sepuluh Nopember Surabaya.

Murfi, H. 2009. *Metode Kernel*. Bahan Kuliah: Machine Learning. Program Studi Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Indonesia Depok.

Widiyanto, M. T. A. C. 2019. Perbandingan Validitas Fuzzy Clustering pada Fuzzy C-Means Dan Particle Swarms Optimazation (PSO) pada Pengelompokan Kelas. *Jurnal Informatika Sunan Kalijaga* Vol. 4, No. 1, Hal: 22-37.