APLIKASI METODE *SILLHOUETTE COEFFICIENT*, METODE *ELBOW* DAN METODE *GAP STATISTIC* DALAM MENENTUKAN *K* OPTIMAL PADA ANALISIS *K-MEDOIDS*

Skripsi

Oleh

HILDA LAILATUL RAMADHANIA
NPM 1817031056

JURUSAN MATEMATIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMPUNG
BANDAR LAMPUNG
2022

ABSTRACT

APPLICATION OF SILLHOUETTE COEFFICIENT METHOD, ELBOW METHOD, AND STATISTIC GAP METHOD IN DETERMINING OPTIMAL K IN K-MEDOIDS ANALYSIS

By

HILDA LAILATUL RAMADHANIA

Cluster analysis is a technique used for grouping data. One method of grouping data often used in cluster analysis is the non-hierarchical method. However, this method is weak in determining the number of clusters before analysis. One of the non-hierarchical methods that are often used is the K-Means method. This method is distance-based which divides the data into some clusters with numeric attributes. For data containing outliers, using the K-Means method is not recommended. Therefore, this method is modified so that it becomes the K-Medoids method. The difference between the two methods lies in selecting the medoid or the median value as the cluster's center. The thesis describes the research results using three methods to determine the optimal number of clusters on some data. The methods are the Sillhouette coefficient, the Elbow, and the Gap Statistics. According to the average Dunn Index value, the Gap Statistics method gave the largest one. Thus, the Gap Statistics method is recommended for research involving outlier data.

Key Words: *Sillhouette coefficient*, *Elbow*, *Gap Statistic*, *K-Medoids*.

ABSTRAK

APLIKASI METODE *SILLHOUETTE COEFFICIENT*, METODE *ELBOW*, DAN METODE *GAP STATISTIC* DALAM MENENTUKAN *K* OPTIMAL PADA ANALISIS *K-MEDOIDS*

Oleh

HILDA LAILATUL RAMADHANIA

Analisis klaster adalah teknik yang digunakan untuk mengelompokkan data. Salah satu metode pengelompokan data yang sering digunakan dalam analisis klaster adalah metode non-hierarki. Namun metode ini lemah dalam menentukan jumlah cluster sebelum dilakukan analisis. Salah satu metode non-hierarki yang sering digunakan adalah metode *K-Means*. Metode ini berbasis jarak yang membagi data menjadi beberapa cluster dengan atribut numerik. Untuk data yang mengandung *outlier*, tidak disarankan menggunakan metode *K-Means*. Oleh karena itu, metode ini dimodifikasi sehingga menjadi metode *K-Medoids*. Perbedaan kedua metode tersebut terletak pada pemilihan nilai *medoid* atau median sebagai pusat cluster. Skripsi ini memaparkan hasil penelitian dengan menggunakan tiga metode untuk menentukan jumlah cluster yang optimal pada beberapa data. Metodenya adalah koefisien *Sillhouette*, *Elbow*, dan *Gap Statistics*. Menurut nilai rata-rata *Dunn Index*, metode *Gap Statistics* memberikan yang terbesar. Oleh karena itu, metode *Gap Statistics* direkomendasikan untuk penelitian yang melibatkan data *outlier*.

Kata kunci: *Sillhouette coefficient, Elbow, Gap Statistic, K-Medoids.*

APLIKASI METODE *SILLHOUETTE COEFFICIENT*, METODE *ELBOW* DAN METODE *GAP STATISTIC* DALAM MENENTUKAN *K* OPTIMAL PADA ANALISIS *K-MEDOIDS*

Oleh

HILDA LAILATUL RAMADHANIA

Skripsi

Sebagai Salah Satu Syarat untuk Mencapai Gelar
SARJANA MATEMATIKA

Pada

Jurusan Matematika
Fakultas Matematika dan Ilmu Pengetahuan Alam

JURUSAN MATEMATIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMPUNG
BANDAR LAMPUNG
2022

Judul Skripsi	:	**APLIKASI METODE *SILLHOUETTE COEFFICIENT*, METODE *ELBOW* DAN METODE *GAP STATISTIC* DALAM MENENTUKAN *K* OPTIMAL PADA ANALISIS *K-MEDOIDS***
Nama Mahasiswa	:	**Hilda Lailatul Ramadhania**
Nomor Pokok Mahasiswa	:	**1817031056**
Jurusan	:	**Matematika**
Fakultas	:	**Matematika dan Ilmu Pengetahuan Alam**

MENYETUJUI

1. Komisi Pembimbing

Widiarti, S.Si., M.Si.	**Prof. Dr. La Zakaria, S.Si., M.Sc.**
NIP 19800502 200501 2 003	NIP 19690213 199402 1 001

2. Ketua Jurusan Matematika

Dr. Aang Nuryaman, S.Si., M.Si.
NIP 19740316 200501 1 001

MENGESAHKAN

1. Tim Penguji

Ketua : **Widiarti, S.Si., M.Si.** .....

Sekretaris : **Prof. Dr. La Zakaria, S.Si., M.Sc.** .....

Penguji
Bukan Pembimbing : **Drs. Nusyirwan, M.Si.** .....

2. Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam

Dr. Eng. Suripto Dwi Yuwono, M.T.
NIP 19740705 200003 1 001

Tanggal Lulus Ujian Skripsi : **26 Juli 2022**

SURAT PERNYATAAN

Yang bertandatangan dibawah ini:

Nama	: **Hilda Lailatul Ramadhania**
Nomor Pokok Mahasiswa	: **1817031056**
Jurusan	: Matematika
Judul Skripsi	: **Aplikasi Metode *Sillhouette Coefficient*, Metode *Elbow* dan Metode *Gap Statisitic* dalam Menentukan *K* Optimal pada Analisis *K-Medoids***

Dengan ini menyatakan bahwa penelitian ini adalah hasil pekerjaan saya sendiri dan sepanjang pengetahuan saya tidak berisi materi yang telah dipublikasikan atau ditulis orang lain atau telah dipergunakan dan diterima sebagai persyaratan penyelesaian studi pada Universitas atau Institusi lain.

Bandar Lampung, 26 Juli 2022

Yang Menyatakan

Hilda Lailatul Ramadhania

NPM. 1817031056

RIWAYAT HIDUP

Penulis bernama lengkap Hilda Lailatul Ramadhania, dilahirkan di Simpang Sari, Lampung Barat 08 Desember 1999 sebagai putri kedua dari Bapak Hermansyah dan Ibu Fanda Yanti.

Penulis menempuh pendidikan Taman Kanak-kanak (TK) Yapsi Lampung Barat diselesaikan pada tahun 2006. Sekolah Dasar (SD) diselesaikan di SDN 03 Tugusari, Lampung Barat pada tahun 2012. Sekolah Menengah Pertama (SMP) di SMPN 01 Sumber Jaya, Lampung Barat dan lulus pada tahun 2015, lalu melanjutkan pendidikan Sekolah Menengah Atas (SMA) di SMAN 01 Sumber Jaya, Lampung Barat dan lulus pada tahun 2018.

Pada tahun 2018 penulis diterima sebagai mahasiswi di jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung melalui jalur PMPAP. Pada tahun 2021, penulis melakukan kerja praktik di Badan Pusat Statistik Kabupaten Way Kanan dan pada tahun yang sama sebagai salah satu bentuk pengabdian masyarakat penulis melaksanakan Kuliah Kerja Nyata di Desa Fajar Bulan, Kecamatan Way Tenong, Kabupaten Lampung Barat, Lampung.

PERSEMBAHAN

Alhamdulillahirobbil'alamin
Dengan mengucapkan rasa syukur yang begitu besar kepada Allah SWT, yang telah memberikan kesempatan, pengalaman, ketabahan, rasa semangat dan pantang menyerah dalam menulis karya ini
Maka
Penulis mempersembahkan karya ini kepada:

Kedua orang tua, Abang dan Adik Tersayang

Terimakasih atas Doa dan dukungan yang selalu diberikan, baik dalam bentuk material, moral, semangat, cinta dan kasih sayang.
Penulis ingin melakukan yang terbaik untuk setiap kepercayaan yang diberikan.
Penulis akan tumbuh, untuk menjadi yang terbaik yang penulis bisa.
Pencapaian ini adalah persembahan kecil penulis untuk kalian.

Dosen Pembimbing, dan seluruh Dosen Matematika FMIPA, Universitas Lampung.

Terima kasih telah berjasa memberikan arahan, semangat, pelajaran dan ilmu yang sangat berharga kepada penulis. Hanya Allah yang mampu membalas setiap kebaikan yang telah diberikan.

Teman-Teman penulis

Terimakasih atas setiap kebersamaan, kebaikan, keceriaan, semangat dan doa yang telah diberikan.

Almamater Universitas Lampung

KATA INSPIRASI

“ Angin tidak berhembus untuk menggoyangkan pepohonan, melainkan menguji kekuatan akarnya.”

(Ali bin Abi Thalib)

“ Sesungguhnya Allah tidak Akan mengubah keadaan suatu kaum, sebelum mereka mengubah keadaan diri mereka sendiri.”

(QS. Ar Rad: 11)

“ Allah tidak akan membebani seseorang melainkan sesuai dengan kesanggupannya.”

(QS. Al-Baqarah : 286)

“ Menuntut ilmu adalah taqwa. Menyampaikan ilmu adalah ibadah. Mengulang-ulang ilmu adalah zikir. Mencari ilmu adalah jihad.”

(Abu Hamid Al Ghazali).

SANWACANA

Alhamdulillahirobbil'alamin, puji syukur kehadirat Allah SWT yang telah melimpahkan rahmat dan karunia-Nya sehingga penulis dapat menyelesaikan skripsi yang berjudul "Aplikasi Metode *Sillhouette Coefficient*, Metode *Elbow* dan Metode *Gap Statisitic* dalam Menentukan *K* Optimal pada Analisis *K-Medoids".*

Penulis menyadari bahwa dalam penyelesaian skripsi ini tidak terlepas dari bantuan berbagai pihak. Oleh karena itu, penulis mengucapkan terima kasih kepada:

1. Ibu Widiarti, S.Si., M.Si., selaku Dosen Pembimbing 1 yang telah meluangkan waktu untuk membimbing, memberikan masukan, memotivasi, serta kritik dan saran yang membangun kepada penulis dalam penyusunan skripsi sehingga skripsi bisa menjadi lebih baik lagi.
2. Bapak Prof. Dr. La Zakaria, S.Si., M.Sc., selaku Dosen Pembimbing II, terima kasih atas bimbingan dan saran yang membangun kepada penulis dalam penyusunan skripsi ini.
3. Bapak Drs. Nusyirwan, M.Si., selaku Dosen Pembahas yang telah memberikan kritik dan saran yang positif dalam penyusunan skripsi ini.
4. Bapak Agus Sutrisno, S.Si.,M.Si.,selaku Dosen Pembimbing Akademik yang telah memberikan arahan, bimbingan, serta motivasi dan semangat bagi penulis selama masa studi.

5. Bapak Dr. Aang Nuryaman, S.Si., M.Si., selaku Ketua Jurusan Matematika Fakultas Matematika dan Ilmu Pengetahun Alam (FMIPA) Universitas Lampung.
6. Bapak Dr. Eng. Suripto Dwi Yuwono, M.T., selaku Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam (FMIPA) Universitas Lampung.
7. Kedua orang tua, abang, adik dan seluruh keluraga yang selalu memberikan semangat, motivasi, dukungan melalui doa terbaiknya maupun bentuk material kepada penulis agar dapat memberikan hasil yang terbaik.
8. Teman terbaik penulis Febi, Irma dan Yuni yang membantu selama proses perskripsian.
9. Teman-teman seperjuangan penuis angkatan 2018, khususnya teman-teman di kelas B dan juga teman seperbimbingan yang selalu memberikan semangat, dukungan dan doa.

Semoga seluruh kebaikan, dukungan, bantuan dan motivasi yang telah dicurahkan maupun diberikan mendapatkan balasan pahala yang setimpal dari Allah SWT.

Akhir kata, penulis menyadari bahwa skripsi ini masih jauh dari sempurna untuk itu penulis mengharapkan kritik dan saran yang bersifat membangun.

Bandar Lampung, 26 Juli 2022

Hilda Lailatul Ramadhania

DAFTAR ISI

Halaman

DAFTAR TABEL

Tabel..... Halaman

DAFTAR GAMBAR

Gambar..... Halaman

I. PENDAHULUAN

1.1 Latar Belakang dan Masalah

Analisis klaster merupakan teknik yang digunakan untuk mengelompokkan objek kedalam kelompok yang relatif homogen yang dinamakan klaster. Objek dalam setiap kelompok memiliki kemiripan satu sama lain dan berbeda jauh dengan objek klaster lainnya. Di dalam pengklasteran, objek hanya boleh masuk ke dalam satu klaster saja sehingga tidak terjadi tumpang tindih (Nugroho, 2008). Proses pengklasteran dilakukan dengan dua metode, yaitu metode hierarki dan metode non hierarki (Widyadhana, dkk., 2021). Metode analisis klaster non hierarki yang paling umum digunakan adalah algoritma *K-Means*. *K-Means* adalah metode pengklasteran berbasis jarak yang membagi data ke dalam sejumlah klaster dan berlaku pada atribut numerik. Algoritma *K-Means* memiliki kelemahan, yaitu tidak *robust* terhadap pencilan karena menggunakan nilai rata-rata sebagai pusat kelompoknya.

Metode *K-Medoids* atau biasa disebut PAM (*Partitioning Around Medoids*) diperkenalkan oleh Kaufman & Rousseeuw (1989). Metode *K-Medoids* merupakan varian umum dari metode *K-Means*. Perbedaan metode *K-Medoids* dari metode *K-Means*, yaitu pada pemilihan *medoid* atau nilai tengah sebagai pusatnya. Medoid adalah objek yang letaknya terpusat dalam suatu kelompok bertujuan untuk mengurangi sensitivitas dari partisi yang dihasilkan sehubungan dengan nilai-nilai ekstrim yang ada pada data set (Sindi, dkk., 2020). Akan tetapi, pada

metode *K-Medoids* memiliki permasalahan dalam penentuan jumlah kelompok sebelum dilakukan analisis (Utami & Saputro, 2018).

Terdapat beberapa penelitian mengenai perbandingan penentuan jumlah *k* optimal. Seperti yang dilakukan Utami & Saputro (2018) melakukan penelitian pengelompokan data yang memuat pencilan dengan kriteria *Elbow* dan koefisien *Sillhouette* (algoritma *K-Medoids*) yang menghasilkan 3 kelompok dengan menggunakan evaluasi klaster dari nilai koefisien *Sillhouette* didapatkan hasil sebesar 0,6409981. Kemudian penelitian Dewi & Paramita (2019) mengenai analisis perbandingan metode *Elbow* dan *Sillhouette* pada algoritma *clustering K-Medoids* dalam pengelompokan produksi kerajinan Bali menggunakan evaluasi klaster *Davies Bouldin Index* (DBI).

Ada beberapa metode yang biasanya digunakan dalam penentuan jumlah klaster yang tepat, diantaranya yang paling umum yaitu, metode *Elbow*, metode koefisien *Sillhouette,* dan *Gap Statistic*. Setiap metode mempunyai kelebihan dan kekurangan, sehingga perlu ketepatan dalam memadukan metode *clustering* yang digunakan, metode untuk menentukan jumlah klaster yang tepat dan struktur data serta ukuran data (Dewi & Paramita, 2019). Maka dari itu, peneliti tertarik untuk melakukan penelitian tentang membandingakan metode *Elbow*, metode koefisien *Sillhouette,* dan metode *Gap Statistic* dalam penentuan jumlah klaster dalam analisis *K-Medoids*. Evaluasi hasil jumlah klaster yang ditentukan dari setiap metode akan ditinjau berdasarkan nilai validasi *Dunn Index* yang dihasilkan.

1.2 Tujuan Penelitian

Penelitian ini bertujuan menentukan jumlah klaster optimal dalam analisis *K-Medoids* menggunakan metode *Sillhouette*, metode *Elbow* dan metode *Gap Statistic*, juga mengevaluasi hasil pengklasteran dari penentuan jumlah *k* optimal setiap metode menggunakan nilai validasi *Dunn Index.*

1.3 Manfaat Penelitian

Penelitian ini dapat memberi manfaat antara lain

1. Memberikan informasi mengenai pengklasteran terbaik dari hasil metode penentuan jumlah *k* optimal.
2. Memberikan wawasan tentang metode penentuan jumlah *k* optimal pada analisis klaster non hierarki.
3. Dapat dijadikan referensi bagi penelitian lanjutan tentang klasterisasi data.

II. TINJAUAN PUSTAKA

2.1 Analisis Klaster

Analisis klaster adalah teknik analisis untuk membagi kelompok utama individu atau objek menjadi beberapa bagian. Secara khusus, analisis klaster adalah mengelompokan sampel entitas (individu atau objek) ke dalam sejumlah kecil kelompok berdasarkan kesamaan diantara entitas (Hair, *et al*., 2010). Proses pengelompokan objek dari setiap partisi memiliki kemiripan berdasarkan matriks tertentu. Dalam proses inti *clustering* atau pengelompokan data, terdapat dua metode yang bisa digunakan, yaitu metode hierarki dan metode non hierarki.

2.1.1 Asumsi Analisis Klaster

Analisis klaster bukanlah teknik inferensi statistik dimana parameter dari sampel dinilai mewakili populasi. Sebaliknya, analisis klaster adalah metode untuk mengukur karakteristik struktural dari suatu pengamatan. Maka dari itu, persyaratan normalitas, linieritas dan homoskedastisitas yang sangat penting dalam teknik lain tidak banyak berpengaruh dalam analisis klaster. Ada dua asumsi yang harus dipenuhi analisis klaster, yaitu sampel representatif dan tidak adanya multikolinieritas antar variabel (Hair, *et al*., 2010).

a. Sampel Representatif

Sampel representatif adalah sampel yang diambil dapat mewakili populasi yang ada. Masalah yang perlu diperhatikan pada data yang digunakan yaitu pencilan. Jika pencilan dibuang menimbulkan bias dalam estimasi struktur yang harus diteliti. Karena keberadaan pencilan mengakibatkan terdapatnya sampel yang

tidak mewakili populasi. Maka, perlunya metode yang *robust* terhadap pencilan agar sampel representatif.

b. Tidak Terjadi Multikolinieritas

Multikolinieritas yaitu adanya hubungan linear yang jelas antar variabel. Adanya multikolinieritas dalam analisis klaster dapat mempersulit penentuan pengaruh dari masing-masing variabel dan dapat mempengaruhi hasil pengklasteran akhir. Pada analisis klaster sebaiknya variabel-variabel tidak terindikasi multikolinieritas. Nilai VIF (*Variance inflation factor*) dapat mengukur seberapa erat hubungan antar variabel. Jika nilai VIF < 10 artinya data tersebut tidak mengandung multikolinieritas (Hidayat & Hakim, 2021). Adapun rumus untuk mengetahui nilai VIF sebagai berikut:

$$VIF_j = \frac{1}{1- r_j^2} \quad (2.1)$$

Dengan, r_j^2 adalah koefisien determinasi variabel dependen X_j dengan variabel bebas lainnya selain variabel ke-*j*.

2.1.2 Ukuran Kemiripan Objek

Dalam analisis klaster pengelompokan didasarkan pada kemiripan antar objek yang diukur dengan menggunakan ukuran jarak (*distance*). Proses dalam pembentukan klaster, diperoleh dengan mencari dan mengelompokkan objek-objek yang memiliki kemiripan dan kedekatan antar suatu objek dengan objek lainnya. Ukuran jarak dengan nilai yang lebih besar menunjukan kesamaan yang lebih rendah (Hair, *et al*., 2010). Pada penelitian ini pengukuran kemiripan yang digunakan dalam proses pengelompokan adalah *Euclidean distance*.

Terdapat beberapa pengukuran jarak yang dapat digunakan untuk ukuran kemiripan, yang paling umum salah salah satunya adalah jarak *Euclidean.* Rumus jarak *Euclidean* dinyatakan sebagai berikut:

$$d_{ij} = \sqrt{\sum_{t=1}^{n}(x_{ik} - x_{jk})^2} \quad (2.2)$$

dengan,

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

n = jumlah variabel klaster

x_{ik} = data dari objek ke-*i* pada variabel ke-*k*

x_{jk} = data dari objek ke-*j* pada variabel ke-*k*.

2.2 Pencilan

Dalam melakukan analisis klaster, perlu diperhatikan perbedaan nilai pada datanya. Jika terdapat perbedaan nilai yang besar antar variabel akan menyebabkan perhitungan jarak menjadi tidak valid (Nugroho, 2008). Metode yang dapat digunakan pada kasus multivariat untuk mendeteksi pencilan adalah pengukuran jarak Mahalanobis (Filzmoser, 2005).

2.2.1 Jarak Mahalanobis

Jarak Mahalanobis pertama kali diperkenalkan oleh Mahalanobis pada tahun 1936. Jarak Mahalanobis diperoleh dengan menghitung jarak setiap observasi terhadap pusat datanya. Pengukuran jarak kuadrat Mahalanobis objek ke-*i* dapat dihitung dengan rumus sebagai berikut:

$$d_{MD}^2(i) = (x_i - \bar{x})' \Sigma^{-1}(x_i - \bar{x}) \quad (2.3)$$

dengan,

$d_{MD}^2(i)$ = jarak kuadrat Mahalanobis objek pada pengamatan ke-*i*

x_i = vektor data objek pengamatan ke-*i* berukuran *p* x *1*

$\bar{x}$ = vektor rata-rata dari tiap variabel berukuran *p* x *1*

$\boldsymbol{\Sigma}$ = matrix kovarian berukuran *p* x *p*, dimana *p* banyaknya variabel.

Pengamatan ke-*i* terindikasi pencilan jika,

$$d_{MD}^2(i) > x_{p,1-a}^2 \quad (2.4)$$

dimana $x_{p,1-a}^2$ merupakan batas pencilan dengan probabilitas $1 - \alpha$.

Langkah-langkah dalam mendeteksi pencilan dengan jarak Mahalanobis sebagai berikut (Johnson & Wichern, 2007):

1. Menentukan nilai vektor rata-rata (μ).
2. Menentukan nilai matriks varians kovarians **(Σ)**
3. Menentukan nilai jarak Mahalanobis pada setiap pengamatan dengan vektor rata-rata.
4. Mengurutkan nilai $d_{MD}^2(i)$ dari kecil ke besar $d_1^2 \leq d_2^2 \leq \cdots \leq d_n^2$.

Mengevaluasi jarak Mahalanobis dengan menggunakan x^2pada derajat kebebasan (*df*) sejumlah variabel yang digunakan dalam penelitian. Identifikasi data pencilan pada pengamatan ke-*i* didefinisikan sebagai pencilan apabila $d_{MD}^2(i) > x_{p,1-a}^2$.

2.3 Penentuan Jumlah *K* Optimal

Terdapat bebrapa metode yang digunakan untuk menentukan jumlah *k* optimal, yang paling umum digunakan yaitu metode *Sillhouette coefficient*, metode *Elbow* dan metode *Gap Statistic*.

a. Metode *Silhouette Coefficient*

Metode *Sillhouette* merupakan metode yang digunakan untuk menentukan jumlah klaster dengan melakukan pendekatan nilai rata-rata metode *Sillhouette* untuk menduga kualitas klaster yang terbentuk (Dewa & Jatipaningrum, 2019). Untuk menghitung nilai *Silhouette coefficient*, diperlukan perhitungan nilai *Sillhouette index* dari sebuah data ke-*i*.

Nilai *Silhouette coefficient* didapatkan dengan mencari nilai maksimal dari nilai *Sillhouette index* global dari jumlah klaster 2 sampai jumlah klaster *n-1* seperti pada persamaan berikut:

$$SC = maks_k SI\ (k) \tag{2.5}$$

dengan,

SC = Sillhouette Coefficient

SI = Sillhouette Index Global

k = jumlah klaster.

Tabel 1. Kategori nilai koefisien *Sillhouette* dan interpretasinya

Koefisien Sillhouette	**Interpretasi**
$0.7 < SC \leq 1$	Terdapat ikatan yang sangat baik (*strong structure*) antara objek dan kelompok yang terbantuk.
$0.5 < SC \leq 0.7$	Terdapat ikatan yang cukup baik (*medium structure*) antara objek dan kelompok yang terbentuk.
$0.25 < SC \leq 0.5$	Terdapat ikatan yang lemah (*weak structure*) antara kelompok objek dan kelompok yang terbentuk.
$SC \leq 0.25$	Tidak terdapat ikatan antara objek dan kelompok yang terbentuk.

Untuk menghitung nilai *SI* dari sebuah data ke-1, ada 2 komponen yaitu a_i dan b_i. Nilai $a_i$adalah rata-rata jarak ke-*i* terhadap semua data lainnya dalam satu klaster. Sedangkan b_i didapatkan dengan menghitung rata-rata jarak data ke-*i* terhadap semua data dari klaster lainnya yang tidak satu klaster dengan data ke-*i*, lalu diambil yang terkecil (Petrovic, 2006). Berikut persamaan untuk menghitung nilai a_i.

$$a(i) = \frac{1}{|C_i|-1} \sum_{j \in C, i \neq j} d(i,j) \tag{2.6}$$

dengan,

C_i = klaster ke-*i*

$d(i,j)$ = jarak objek ke-*i* dengan objek lainnya pada satu klaster *j*.

Kemudian menghitung nilai b_i dengan persamaan sebagai berikut :

$$b(i) = \min \frac{1}{|C_k|} \sum_{j \epsilon C_k} d(i,j) \quad (2.7)$$

dengan,

C_k = jumlah data pada klaster *k*

$d(i,j)$ = jarak objek ke-*i* dengan objek *j* pda klaster *k*.

Selanjutnya adalah rumus perhitungan mendapatkan nilai SI_i^j dapat dilihat pada persamaan

$$SI_i^j = \frac{b_i^j - a_i^j}{\max\{a_i^j, b_i^j\}} \quad (2.8)$$

dengan,

SI_i^j = *Sillhouette index* data ke-i dalam satu klaster

b_i^j = rata-rata jarak ke-*i* terhadap semua data yang tidak dalam satu klaster dengan data ke-*i*

a_i^j = rata-rata jarak data ke-*i* terhadap semua data dalam satu klaster.

Berikut ini adalah rumus perhitungan mendapatkan nilai SI_j dapat dilihat pada persamaan

$$SI_j = \frac{1}{m_j} \sum_{i=1}^{m_j} SI_i^j \quad (2.9)$$

dengan,

SI_j = rata-rata *Sillhouette index cluster j*

SI_i^j = *Sillhouette index* data ke-*i* dalam klaster *j*

m_j = jumlah data dalam klaster ke-*j*

i = *index* data (i = 1,2,... m_j)

Berikut ini adalah rumus perhitungan mendapatkan nilai *SI* global dengan persamaan

$$SI = \frac{1}{k} \sum_{j=1}^{k} SI_j \quad (2.10)$$

dengan,

SI = rata-rata *Sillhouette index* dari dataset

SI_j = rata-rata *Sillhouette index cluste*r *j*

k = jumlah klaster

b. Metode *Elbow*

Metode *Elbow* merupakan metode untuk menentukan jumlah klaster yang tepat melalui persentase hasil perbandingan antara jumlah klaster yang akan membentuk siku pada suatu titik. Jika nilai klaster pertama dengan nilai klaster kedua memberikan sudut dalam grafik atau nilainya mengalami penurunan paling besar dapat menjadi jumlah nilai klaster yang tepat untuk digunakan (Bholowalia & Kumar, 2014). Untuk mendapatkan perbandingannya adalah dengan menghitung *sum square error* (*SSE*) dari masing-masing nilai klaster *k*. Maka nilai *SSE* akan semakin kecil. Rumus *SSE* sebagai berikut:

$$SSE = \sum_{k=1}^{n} \sum_{x_i} |x_i - c_k|^2 \quad (2.11)$$

dengan,

k = klaster ke-*k*

x_i = objek data ke-*i*

c_k = pusat klaster ke-*i.*

c. Metode *Gap Statistic*

Metode ini pada mulanya diperkenalkan oleh Tibshirani, *et al*. (2001). Metode *Gap Statistic* dapat digunakan dalam menentukan jumlah kelompok optimum pada suatu himpunan data dengan membangkitkan data referensi *uniform*. Ide dari *Gap Statistic* adalah mengusulkan dispersi dalam kelompok versus jumlah kelompok.

Gap statistic merupakan metode untuk menduga kelompok optimum pada analisis klaster. Teknik ini berdasar pada perubahan dispersi dalam klaster dengan peningkatan jumlah kelompok dari data (Arima, *et al*., 2005). Secara detail *Gap Statistic* dapat dijelasakan sebagai berikut:

Misalkan $\{X_{ij}\}$ dengan $i = 1,2,...,n$ dan $j = 1,2,...,p$ merupakan himpunan data dengan *p* (peubah) pada observasi independen dengan *n* objek. Kemudian data dikelompokan menjadi *k* kelompok yaitu $C_1, C_2, ...,C_k$ dengan C_r menunjukkan pengamatan pada kelompok ke-*r*. Kemudian didefinisikan sebagai berikut:

$$Dr = \Sigma_{x_i x_j} d(x_i, x_j) \tag{2.12}$$

dengan,

Dr = jarak *Euclid* data observasi

$d(x_i, x_j)$ = jarak antara objek ke-*i* dan objek ke-*j* dimana $i \neq j$.

Kemudian menghitung W_k:

$$W_k = \sum_{r=1}^{k} \frac{1}{2n_r} Dr \tag{2.13}$$

dengan W_k adalah jumlah kuadrat dalam klaster, sedangkan n_r adalah banyaknya observasi (anggota) kelompok ke-*r*. Berikut adalah *Gap Statistic* untuk *k* tertentu:

$$Gap(k) = \left[\frac{1}{B}\right] \Sigma_b \{\log(W_{kb}^*) - \log(W_k) \tag{2.14}$$

Dimana *B* adalah *resampling* (dari data simulasi) dengan pengambilan sebanyak *B* kali dengan distribusi *uniform*. Tahapan penentuan jumlah klaster optimal menggunakan metode *Gap Statistic* sebagai berikut*:*

1) Mengelompokkan data dan mengubah-ubah banyaknya kelompok mulai dari k=1,2,...,*n*, dan hitung total variasi *intracluster* W_k, dengan k = 1,2,...,*n*.
2) Hasilkan kumpulan data referensi *B* dengan distribusi referensi *uniform*. Klasterkan masing-masing dari kumpulan data referensi ini dengan berbagai jumlah kelompok $k = 1,...,k_{max}$ dan menghitung total variasi *intracluster* $W_{kb.}$

3) Hitung estimasi *Gap Satistic* sebagai penyimpangan nilai W_k yang diamati dari W_{kb} dan juga hitung standar deviasinya.
4) Pilih jumlah klaster sebagai nilai terkecil dari *k* sehingga *Gap Statistic* berada dalam satu standar deviasi dari celah pada *k+1*.

2.4 Metode Pengklasteran

Ada dua metode pengklasteran yaitu metode hierarki dan metode non hierarki (pengklasteran *K-Means*). Pengklasteran yang ideal adalah pengklasteran yang tiap objek hanya masuk atau menjadi anggota dari salah satu klaster sehingga tidak terjadi tumpang tindih. Semua metode pada dasarnya menggunakan kesamaan atau ketidaksamaan objek (Nugroho, 2008).

a) Metode Hierarki

Metode ini memulai pengelompokan dengan dua atau lebih objek yang mempunyai kesamaan paling dekat. Kemudian proses diteruskan ke proses lain yang mempunyai kedekatan kedua. Demikian seterusnya sehingga klaster akan membentuk semacam pohon dimana ada hierarki (tingkatan yang jelas) antara objek. Untuk membantu memperjelas hierarki biasanya digunakan dendrogam. Menurut Hair, *et al*. (2010) terdapat lima metode hierarki yang paling sering digunakan yaitu *Single Linkage, Complate Linkage, Average Linkage, Ward Method,* dan *Centroid Method.*

b) Metode Non Hierarki

Berbeda dengan metode hierarki sebelumnya, metode ini justru dimulai dengan menentukan terlebih dahulu jumlah klaster yang diinginkan. Setelah jumlah klaster diketahui lalu proses klaster dilakukan tanpa mengikuti proses hierarki.

Metode ini juga disebut pengklasteran *K-Means*. Pengklasteran *K-Means* sangat cocok untuk data dengan ukuran yang besar karena memiliki kecepatan yang lebih tinggi dibandingkan metode hierarki. Namun, pemilihan banyaknya klaster dan *centroid* yang harus ditentukan lebih dahulu menjadi kelemahan metode ini. Hasil pengklasteran mungkin tergantung pada urutan observasi data (Nugroho, 2008). Terdapat beberapa jenis metode non hierarki diantaranya yaitu *sequential threshold, parallel threshold, dan optimizing partitioning* (metode partisi). Pada penelitian ini akan mengkaji mengenai metode non hierarki *optimizing partitioning* (metode partisi) dalam membandingan metode penentuan jumlah *k* optimal.

2.4.1 Metode Partisi

Dalam analisis klaster menggunakan metode partisi bertujuan untuk membagi data menjadi beberapa kelompok atau klaster. Dalam satu klaster memiliki tingkat kemiripan yang tinggi. Langkah kerja metode partisi yaitu apabila terdapat basis data sejumlah *n* objek. Selanjutnya data dipartisi menjadi *k* partisi dari data, dimana setiap partisi mewakili sebuah klaster $k \leq n$. Syarat yang harus terpenuhi untuk metode partisi sebagai berikut (Kaufman & Rousseeuw, 1989):

1. Setiap kelompok harus berisi setidaknya satu objek.
2. Setiap objek harus memiliki tepat satu kelompok.

Pada metode partisi terdapat beberapa metode yang sering digunakan antara lain metode *K-Means* dan metode *K-Medoids*. Metode partisi yang digunakan dalam penelitian ini yaitu metode *K-Medoids*.

2.4.1.1 Metode *K-Medoids*

Metode *K-Medoids* dikenal juga sebagai PAM (*Partitioning Around Medoids*) merupakan salah satu metode yang digunakan untuk proses klastering. Dalam metode ini data yang terdiri dari *n* objek di partisi menjadi *k* klaster dimana jumlah $k \leq n$. Maka dari itu tujuannya adalah untuk menemukan *k* objek tersebut (Kaufman & Rousseeuw, 1989). Perbedaan metode *K-Medoids* dengan metode *K-Means* terletak pada penentuan pusat klaster, dimana metode *K-Medoids* nilai tengah dari data sebagai medoids atau perwakilan untuk pusat klaster. Sedangkan algoritma *K-Means* menggunakan nilai rata-rata dari setiap klaster sebagai pusat klaster (Kaur, *et al*., 2014).

Berikut merupakan langkah-langkah metode *K-Medoids*:

1. Pilih jumlah klaster sebanyak jumlah klaster *k* (jumlah klaster).
   Alokasikan setiap data atau objek ke dalam klaster terdekat dengan menggunakan ukuran jarak *Euclidean* (persamaan 2.2).
2. Pilihlah objek secara acak pada setiap klaster menjadi kandidat *medoid* baru.
3. Hitunng jarak yang berada di setiap klaster melalui *medoid* baru.
4. Pilih jumlah klaster sebanyak jumlahl klaster *k* (jumlah klaster).
   Alokasikan setiap data atau objek ke dalam klaster terdekat dengan menggunakan ukuran jarak *Euclidean* (persamaan 2.2).
5. Pilihlah objek secara acak pada setiap klaster menjadi kandidat *medoid* baru.
6. Hitunng jarak yang berada di setiap klaster melalui *medoid* baru.
7. Hitung total simpangan (*S*) dengan menghitung nilai total *distance* baru dikurang total *distance* lama. Ganti objek menggunakan data klaster untuk memperoleh sekelompok *k* objek yang baru sebagai *medoid*, jika $S < 0$.
8. Ulangi tahap ke 3 sampai ke 5 hingga tidak terjadi perubahan *medoid*, sehingga didapatkan klaster beserta anggota klaster masing-masing.

2.5 Validasi Klaster

Untuk mengetahui seberapa baik hasil yang diperoleh dari proses *clustering* diperlukannya klaster validasi. Jika pada klasifikasi dapat diketahui dengan jelas mengenai pengukuran keberhasilan suatu proses klasifikasi dengan melihat nilai presisi, *recall*, dan akurasinya yang bersifat eksak atau pasti. Pengukuran hasil proses *clustering* memiliki banyak metode seperti pengukuran berdasarkan nilai korelasinya, nilai *cohessieon* dan *separation, Silhouette coefficient, Dunn Index,* dan *Davies-Bouldin Index* (Saikhu & Gita, 2013). Pada penelitian ini digunakan *Dunn Index* sebagai validitas klaster.

Dunn Index adalah salah satu pengukur validitas klaster yang diajukan oleh J. C. Dunn. Pada jurnal penelitial Luthfi & Wijayanto (2021) yang bersumber dari penelitian J. C Dunn pada tahun 1973 menjelaskan bahwa *index validitas Dunn* menghitung nilai minimum dari perbandingan antara nilai fungsi *dissimilaritas* antara dua klaster sebagai *separation* dan nilai maksimum dari diameter klaster sebagai *compacness*. Jumlah klaster terbaik ditunjukkan dengan semakin besarnya nilai *Dunn Index*. Rumus *Dunn Index* dapat ditulis sebagai berikut:

$$D = \left\{ \frac{min_{\substack{1\le i\le k \\ i+1\le j\le l}} \left(d(C_i, C_j) \right)}{max_{1\le l\le q}\, diam(C_k)} \right\} \tag{2.15}$$

dengan,

D = *Dunn Index*

q = jumlah klaster

$d(c_i, c_j)$ = jarak *Euclidean* antar pasangan objek pada klaster *i* dan klaster *j* (*intercluster*)

$diam(c_k)$ = jarak *Euclidean* antar objek dengan nilai rata-rata klaster (*intracluster*)

Dunn Index memiliki rentang nilai dari nol sampai tak hingga. Jika nilai *Dunn Index* semakin besar, maka hasil klaster akan semakin baik.

III. METODOLOGI PENELITIAN

3.1 Waktu dan Tempat Penelitian

Penelitian ini dilaksanakan pada semester genap tahun ajaran 2021/2022 yang bertempat di Jurusan Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Lampung.

3.2 Data Penelitian

Data penelitian yang digunakan adalah data penelitian Andini, dkk. (2019) yang diambil dari Badan Pusat Statistik Sumatera Utara, yaitu persentase rumah tangga Sumatra Utara pada tahun 2015 yang disimulasikan dengan menggunakan *software Rstudio* versi 4.0.3. Jumlah variabel yang digunakan sebanyak lima dengan keterangan setiap variabel sebagai berikut:

X1 = persentase rumah tangga menurut sumber air minum kemasan,
X2 = persentase rumah tangga menurut sumber air minum ledeng,
X3 = persentase rumah tangga menurut sumber air minum pompa,
X4 = persentase rumah tangga menurut sumber air minum sumur,
X5 = persentase rumah tangga menurut sumber air minum mata air.

Lalu membangkitkan pencilan sebesar 10% pada data yang dibangkitkan tanpa merubah jumlah objek pada data yang dibangkitkan. Banyaknya objek (n) yang dibangkitkan terdiri atas 30, 50, 100, 150, 210, 370, 500, 750, 885 dan 1000 data bangkitan.

3.3 Metode Penelitian

Penelitian ini dilakukan dengan mempelajari buku-buku teks, jurnal serta akses internet. Selain itu penelitian juga dilakukan secara simulasi yang bertujuan untuk mengetahui metode terbaik dalam menentukan jumlah *k* optimal pada data penelitian berdasarkan nilai validasi klaster dari jumlah *k* yang telah ditentukan. Adapun langkah-langkah penelitian yang dilakukan sebagai berikut:

1. Membangkitkan data penelitian.
   a) Membangkitkan data berdistribusi normal dengan lima variabel bebas. Masing-masing variabel bebas memiliki distribusi yaitu $X_1 \sim N(22.39, 18.31)$, $X_2 \sim N(12.70, 16.58)$, $X_3 \sim N(14.74, 16.24)$, $X_4 \sim N(22.12, 18.37)$, $X_5 \sim N(18.59, 18.38)$. Sehingga diperoleh:

      $\mu = (22.39, 12.70, 14.74, 22.12, 18.59)$

      dan

$$\boldsymbol{\Sigma} = \begin{bmatrix} 267.913 & -4{,}533 & -37.348 & 136.347 & -86.482 \\ -4.533 & 245.135 & -23.734 & 77.280 & 8.643 \\ -37.348 & -23.734 & 317.808 & -50.920 & -100.754 \\ 136.347 & 77.280 & -50.920 & 493.681 & 8.593 \\ -86.482 & 8.643 & -100.754 & 8.593 & 389.153 \end{bmatrix} \quad (3.1)$$

   b) Selanjutnya membangkitkan data berdistribusi normal multivariat dengan lima variabel bebas untuk setiap *n* sehingga $X \sim \boldsymbol{N}_5(n,\mu,\boldsymbol{\Sigma})$ dengan $\mu = (22.39, 12.70, 14.74, 22.12, 18.59)$ dan Σ pada persamaan (3.1).
   c) Kemudian membangkitkan data pencilan sebesar 10% dari data yang dibangkitkan dengan distribusi normal (10, 1).
   d) Memasukkan pencilan ke dalam data tanpa merubah jumlah data yang dibangkitkan. Ketentuan data ke-1 sampai ke-*p* sebagai pencilan dengan *p* = jumlah pencilan.

2. Melakukan uji asumsi klaster.
   a) Mendeteksi pencilan dengan metode jarak Mahalanobis dari persamaan (2.3)
   b) Uji asumsi multikolinieritas dengan melihat apakah terdapat hubungan linier atau tidak berdasarkan nilai VIF antar variabel.
3. Menentukan jumlah klaster optimal yang dapat dilihat berdasarkan plot dari setiap metode.
   a) Metode Koefisien *Sillhouette*

      Metode ini mengukur seberapa baik suatu objek diposisikan dalam satu klaster dengan melihat nilai rata-rata *Sillhouette coefficient* dari persamaan (2.5).
   b) Metode *Elbow*

      Metode ini menggunakan nilai persentase hasil perbandingan antar jumlah klaster yang akan membentuk siku pada suati titik untuk menentukan jumlah klaster yang tepat. Perbandingannya adalah dengan menghitung *sum square error* (*SSE*) pada persamaan (2.11).
   c) Metode *Gap Statistic*

      Metode ini digunakan untuk menentukan jumlah kelompok optimum dengan membangkitkan data referensi distribusi *unifrom*. Teknik ini berdasarkan peubah dispersi dalam klaster dengan peningkatan jumlah kelompok dari data. Jumlah kelompok (k) optimum diperoleh dengan membandingkan nilai $Gap(k)$ dari persamaan (2.14).
4. Melakukan pengklasteran dengan menggunakan metode *K-Medoids*. Dengan jumlah klaster (k) yang telah ditentukan sebelumnya berdasarkan masing-masing metode penentuan jumlah klaster optimal.
5. Menghitung nilai *Dunn Index* dalam penentuan jumlah klaster dari ketiga metode yang digunakan berdasarkan persamaan (2.15).
6. Melakukan evaluasi dari metode *Elbow*, metode koefisien *Sillhouette* dan metode *Gap Statistic* berdasarkan hasil nilai rata-rata *Dunn Index*.

V. KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan hasil dan pembahasan yang telah dilakukan pada Bab IV, dapat disimpulan bahwa penentuan jumlah klaster yang optimal pada masing-masing metode dipengaruhi oleh setiap objek dan jumlah data. Selain itu, berdasarkan nilai rata-rata *Dunn Index* dari keseluruhan data yang dibangkitkan didapatkan nilai rata-rata paling besar menggunakan metode *Gap Statistic* yaitu sebesar **0,125734**, sedangkan untuk nilai rata-rata *Dunn Index* metode koefisien *Sillhouette* sebesar **0,124483** dan metode *Elbow* sebesar. **0,106837** Nilai *Dunn Index* pada jumlah klaster yang sama akan menghasilkan nilai *Dunn Index* paling besar saat jumlah data lebih kecil.

5.2 Saran

Untuk pengembangan penelitian lebih lanjut disarankan untuk dilakukan beberapa persentase pencilan yang berbeda-beda. Hal ini untuk melihat pengaruh metode dalam penentuan jumlah klaster.

DAFTAR PUSTAKA

Andini, N., Ramadhani, T., Rahayu, S.P., & Lindasari, D. 2019. Pengujian Normal Multivariat dan Vektor Mean pada Data Prosentase Rumah Tangga Menurut Sumber Mata Air Minum Provinsi Aceh dan Sumatera Utara Tahun 2015.

Arima, C., Hakamada, K., Okamoto, M., & Hanai, T. 2005. Validity Index for Fuzzy K-means Clustering Using the Gap Statistic Method. *Sixteenth International Conference on Genome Informatics*.

Bholowalia, P. & Kumar, A. 2014. EBK-Means: A Clustering Technique based Elbow Method and K-Means in WSN. *International Journal of Computer Application*. 105(09): 17 – 24.

Dewa, F.A. & Jatipaningrum, M.T. 2019. Segmentasi E-comerace dengan Cluster K-Means dan Fuzzy C-Means. *Jurnal Statistika Industri dan Komputasi*. **04**(01): 53-67.

Dewi, D.A.I.C. & Paramita, D.A.K. 2019. Analisis Perbandingan Metode Elbow dan Sillhouette pada Algoritma Clustering K-medoids Dalam Pengelompokan Produksi Kerajinan Bali. *Jurnal Matrix*. **9**(03): 102-109.

Filzmoser, P. 2005. Identification of Multivariate Outliers: a Performance Study. *Austrian Journal of Statistics*. **34**(02): 127-138.

Hair, JR. J.F., Black, W.C., Babin, B.J., & Anderson, R.E. 2010. *Multivariate Data Analysis*. 7th Edition. Pearson Education Limited, England.

Hidayat, F.P., & Hakim, R.B.F. 2021. Implementasi Metode Clustering K-medoids Dalam Mengelompokan Jumlah Aduan di Kabupaten Sleman, hlm 106-114. Prosiding Seminar Matematika dan Pendidikan Matematika. Yogyakarta.

Johnson, R. A. & Wichern, D.W. 2007. *Applied Multivariate Statistical Analysis*. Pearson Education. Inc

Kaufman, L. & Rousseeuw, P.J. 1989. *Finding Groups in Data an Introduction to Cluster Analysis*. Willey, J. & Sons. New Jersey.

Kaur, N.K., Kaur, U., & Singh, D. 2014. K-medoid Clustering Algorithm- A Review. *International Journal of Computer Application and Technology*. **01**(01): 42-45.

Luthfi, E. & Wijayanto, A.W. 2021. Analisis Perbandingan Metode Hierarchical, K-Means, dan K-Medoids Clustering dalam Pengelompokan Index Pembangunan Manusia Indonesia. *Jurnal Inovasi*. **17**(4): 761-773.

Nugroho, S. 2008. *Statistika Multivariat Terapan*. Ed ke-1. UNIB Press. Bengkulu.

Petrovic, S. 2006. A Comparison Betweem the Sillhouette Index and the Davies-Bouldin Index in Labelling IDS Clusters. *In 11th Nordic Workshop on Secure IT-systems*.

Saikhu, A., & Gita, Y.B. 2013. Implementasi Deteksi Outlier Pada Algoritma Hierarchical Clustering. Prosiding Seminar Nasional Teknologi Informasi dan Multimedia. Yogyakarta.

Sindi, S., Ningse, W.R.O., Sihombing, I.A., Zer, F.I R.H., & Hartama, D. 2020. Analisis Algoritma K-medoids Clustering Dalam Pengelompokan Penyebaran Covid-19 di Indonesia. *Jurnal Teknologi Informasi*. **4**(01): 166-173.

Tibshirani, R., Walther, G., & Hastie, T. 2001. Estiming the Number of Cluster in a Data Set Via the Gap Statistic. *J. R. Statist*. **63**(02): 411-423.

Utami, D.S., & Saputro, D.R.S. 2018. Pengelompokan data yang Memuat Pencilan dengan Kriteria Elbow dan Koefisien Sillhouette (Algoritma K-medoids). Prosiding KNPMP III. Surakarta.

Widyadhana, D., Hastuti, R.B., Kharisudin, I., & Fauzi, F. 2021. Perbandingan Analisis Klaster K-means dan Average Linkage untuk Pengklasteran Kemiskinan di Provinsi Jawa Tengah. Prosiding Seminar Nasional Matematika. Semarang.