IMPLEMENTASI METODE *ENSEMBLE ROBUST CLUSTERING USING LINKS* (ROCK) UNTUK KLASTERISASI SEKOLAH MENENGAH ATAS (SMA) DI BANDAR LAMPUNG

(Skripsi)

Oleh

DINI DESITA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMPUNG
BANDAR LAMPUNG
2023

ABSTRACT

IMPLEMENTATION OF THE ENSEMBLE ROBUST CLUSTERING USING LINKS (ROCK) METHOD FOR SENIOR HIGH SCHOOL (SMA) CLUSTERIZATION IN BANDAR LAMPUNG

By

Dini Desita

Cluster analysis is a method in interdependent multivariate analysis to group a set of objects based on their similar characteristics. One of the cluster analysis methods that can be used to group mixed data types (numeric and categorical) is Ensemble ROCK which has the advantage that there is a better accuracy value compared to the agglomerative hierarchical method because it has good scalability and has better quality in handle categorical data. In this analysis, the K-Medoids method is used to form clusters on numerical data and the ROCK method is used on categorical data. The purpose of this study is to apply the Ensemble ROCK cluster analysis to classify high schools in Bandar Lampung and obtain the best number of clusters. The data used is data from 60 high schools in Bandar Lampung based on high school facilities and infrastructure as well as human resources. The best grouping results are selected based on the ratio value of Sw and Sb. From the results and analysis it was found that the number of clusters-2 with a value of θ=0.05 was the best cluster result in mixed data with the smallest Sw and Sb ratio value of 0.0310. The characteristics of cluster-1 are clusters consisting of 25 schools which have better quality than cluster-2 which consists of 35 schools.

Keyword: Cluster Analysis, *Ensemble* ROCK, *K-Medoids,* Senior High School, Ratio *Sw* dan *Sb*

ABSTRAK

IMPLEMENTASI METODE *ENSEMBLE ROBUST CLUSTERING USING LINKS* (ROCK) UNTUK KLASTERISASI SEKOLAH MENENGAH ATAS (SMA) DI BANDAR LAMPUNG

Oleh

Dini Desita

Analisis *cluster* merupakan salah satu metode dalam analisis multivariat yang digunakan untuk mengelompokkan sekumpulan objek berdasarkan kemiripan karakteristiknya. Metode *ensemble* ROCK adalah salah satu teknik dalam analisis *cluster* yang dapat digunakan untuk mengelompokkan data dengan tipe campuran (numerik dan kategorik). Metode ini memiliki ketepatan dan sifat skalabilitas yang baik untuk data berskala campuran. Dalam penelitian ini, metode *K-Medoids* digunakan untuk membentuk *cluster* pada data numerik dan metode ROCK digunakan pada data kategorik. Tujuan dari penelitian ini adalah menerapkan analisis *cluster ensemble* ROCK untuk mengelompokkan SMA di Bandar Lampung berdasarkan sarana dan prasarana serta sumber daya manusia SMA dengan menggunakan metode *ensemble* ROCK. Data yang digunakan adalah data 60 SMA di Bandar Lampung. Hasil pengelompokkan terbaik dipilih berdasarkan nilai *ratio* $Sw\ dan\ Sb$. Dari hasil dan analisis data diperoleh bahwa pada nilai $\theta = 0.05$ dihasilkan nilai *ratio* $Sw\ dan\ Sb$ terkecil sebesar 0.0310 sehingga dapat disimpulkan bahwa jumlah *cluster* optimum adalah 2. Dengan karakteristik *Cluster*-1 merupakan *cluster* yang terdiri dari 25 sekolah yang memiliki kualitas yang lebih baik dibandingkan pada *Cluster*-2 yang terdiri dari 35 sekolah.

Kata Kunci: Analisis *Cluster*, *Ensemble* ROCK, *K-Medoids*, Sekolah Menengah Atas, Ratio Sw dan Sb

IMPLEMENTASI METODE *ENSEMBLE ROBUST CLUSTERING USING LINKS* (ROCK) UNTUK KLASTERISASI SEKOLAH MENENGAH ATAS (SMA) DI BANDAR LAMPUNG

Oleh
DINI DESITA
1917031042

Skripsi

Sebagai Salah Satu Syarat untuk Mencapai Gelar
SARJANA MATEMATIKA

Pada

Jurusan Matematika
Fakultas Matematika dan Ilmu Pengetahuan Alam
Universitas Lampung

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMPUNG
BANDAR LAMPUNG
2023

Judul Skripsi	:	**IMPLEMENTASI METODE *ENSEMBLE ROBUST CLUSTERING USING LINKS* (ROCK) UNTUK KLASTERISASI SEKOLAH MENENGAH ATAS (SMA) DI BANDAR LAMPUNG**
Nama Mahasiswa	:	Dini Desita
Nomor Pokok Mahasiswa	:	**1917031042**
Jurusan	:	**Matematika**
Fakultas	:	**Matematika dan Ilmu Pengetahuan Alam**

MENYETUJUI

1. Komisi Pembimbing

Drs. Eri Setiawan, M.Si.	**Agus Sutrisno, S.Si., M.Si.**
NIP. 19581101 198803 1 002	NIP. 19700831 199983 1 002

2. Mengetahui
Ketua Jurusan Matematika

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

MENGESAHKAN

1. Tim Penguji

   Ketua : **Drs. Eri Setiawan, M.Si.** .....

   Sekretaris : **Agus Sutrisno, S.Si., M.Si.** .....

   Penguji
   Bukan Pembimbing : **Dr. Khoirin Nisa, S.Si., M.Si** .....

2. Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam

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

Tanggal Lulus Ujian Skripsi : **24 Januari 2023**

PERNYATAAN SKRIPSI MAHASISWA

Saya yang bertanda tangan di bawah ini:

Nama	:	**Dini Desita**
Nomor Pokok Mahasiswa	:	**1917031042**
Jurusan	:	**Matematika**
Judul Skripsi	:	**IMPLEMENTASI METODE *ENSEMBLE ROBUST CLUSTERING USING LINKS* (ROCK) UNTUK KLASTERISASI SEKOLAH MENENGAH ATAS (SMA) DI BANDAR LAMPUNG**

Dengan ini menyatakan bahwa skripsi ini adalah hasil pekerjaan saya sendiri dan semua tulisan yang tertuang dalam skripsi ini telah mengikuti kaidah karya penulisan ilmiah Universitas Lampung.

Bandar Lampung, 24 Januari 2023

Penulis

Dini Desita

NPM. 1917031042

RIWAYAT HIDUP

Penulis memiliki nama lengkap Dini Desita yang lahir di Metro pada tanggal 08 Juni 2001. Penulis merupakan anak tunggal dari pasangan Bapak Sumadi dan Ibu Warsini.

Penulis menyelesaikan Pendidikan Sekolah Dasar di SD Negeri 5 Metro pada tahun 2013, Pendidikan Sekolah Menengah Pertama di SMP Negeri 1 Metro yang diselesaikan pada tahun 2016, dan Pendidikan Sekolah Menengah Atas di SMA Negeri 6 Metro pada tahun 2019.

Penulis melanjutkan Pendidikan Strata Satu (S1) di Jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam (FMIPA) Universitas Lampung pada tahun 2019 melalui jalur SBMPTN. Sebagai bentuk penerapan ilmu perkuliahan pada tahun 2022, penulis melaksanakan Kuliah Kerja Nyata (KKN) di Kelurahan Sumberejo Kecamatan Simpang Kanan Kabupaten Tanggamus dan Kerja Praktik (KP) di Badan Pusat Statistik (BPS) Kota Metro.

Selama menjadi mahasiswa penulis aktif mengikuti organisasi sebagai Bidang 1 Keanggotaan pada periode 2020-2021 di Koperasi Mahasiswa Universitas Lampung (KOPMA UNILA). Penulis juga mengikuti kepanitiaan Coop Education Festival (COUNFEST) sebagai anggota Kestari pada tahun 2019. Penulis juga mengikuti salah satu Program Merdeka Belajar-Kampus Merdeka (MBKM) yaitu Kampus Mengajar Angkatan 2 di SMP N 5 Metro Periode Agustus 2021 - Desember 2021.

KATA INSPIRASI

"Dan barang siapa yang bertakwa kepada Allah,
niscaya Allah menjadikan baginya
kemudahan dalam urusannya"
(Q.S. At-Talaq: 4)

"Bisa dibilang rahasia sukses itu tidak ada dan
kalaupun ada itu hanya ada dua, yang pertama
bertahan sampai akhir dan jangan menyerah, kedua
kalau kamu nyerah kamu harus balik lagi ke rahasia
pertama yaitu bertahan sampai akhir"
(Zhong Chenle)

"Semua ada waktunya. Jangan membandingkan hidupmu
Dengan orang lain. Tidak ada perbandingan antara
Matahari dan bulan, mereka bersinar saat
Waktunya tiba"
(B.J. Habibie)

PERSEMBAHAN

Bismillahirrahmanirrahim

Alhamdullilahirobbil'alamin,

Puji dan syukur saya haturkan kepada Allah AWT atas ridho-nya sehingga penulis dapat menyelesaikan skripsi ini. Saya persembahkan karya sederhana ini untuk:

Diri Sendiri

Terima kasih karena tetap kuat dan tetap bertahan di setiap situasi dan berusaha dalam mempersiapkan semua nya

Kedua Orang Tua

Terima kasih atas segala kasih sayang, pengorbanan doa, dan nasihat. Terima kasih atas segala dukungan dan motivasinya untuk penulis. Terimakasih telah mengajarkan banyak hal dan pelajaran yang berharga sehingga menjadikan penulis menjadi seseorang yang kuat. Terima kasih selalu ada untuk penulis disaat sedih dan senang

Dosen Pembimbing dan Pembahas

Terima kasih kepada dosen pembimbing dan pembahas yang sudah sangat membantu, memberikan motivasi, serta memberikan arahan dan ilmu yang berharga

Sahabat-sahabatku

Terimakasih atas semua keceriaan dan semangat yang telah diberikan

Almamater tercinta, Universitas Lampung

SANWACANA

Puji syukur kehadirat Allah SWT, atas segala rahmat dan karunianya sehingga penulis dapat menyelesaikan skripsi dengan judul **"Implementasi Metode *Ensemble Robust Clustering Using Links* (ROCK) untuk Klasterisasi Sekolah Menengah Atas (SMA) di Bandar Lampung"**. Skripsi ini disusun sebagai salah satu syarat memperoleh gelar Sarjana Matematika (S.Mat) pada Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung.

Dalam kesempatan ini, penulis ingin mengucapkan terimakasih kepada :

1. Bapak Drs. Eri Setiawan, M.Si., selaku pembimbing utama atas kesediaan waktu, pemikiran dalam memberikan evaluasi, arahan, dan saran yang membangun dalam proses penyusunan skripsi ini.
2. Bapak Agus Sutrisno, S.Si., M.Si., selaku pembimbing kedua atas kesediaan waktu, arahan dan saran yang membangun dalam proses penulisan skripsi ini.
3. Ibu Dr. Khoirin Nisa, S.Si., M.Si., selaku dosen pembahas atas kesediaan waktu, saran dan masukan yang membangun selama proses penyusunan skripsi ini.
4. Ibu Dra. Dorrah Aziz, M.Si., selaku dosen pembimbing akademik yang telah memberikan bimbingan, motivasi dan nasihat selama penulis menjalankan perkuliahan.
5. Bapak Dr. Aang Nuryaman, S.Si., M.Si. selaku Ketua Jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung.
6. Bapak Dr. Eng. Suripto Dwi Yuwono, M.T., selaku dekan Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung.

7. Seluruh Dosen, Staf dan Civitas Akademik Jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung.
8. Bapak dan Ibu tersayang beserta keluarga besar yang selalu memberikan semangat dukungan, dan do'a kepada penulis.
9. Untuk keluarga besar teravv, Maji, Sinta, Puja, Aldi, Wiranto dan Fazri khamsahamida untuk segala motivasi, dukungan, kebersamaan dan saling support dalam menjalani perkuliahan serta selama proses penyusunan skripsi ini.
10. Untuk dia yang tidak bisa disebut namanya, terimakasih atas segala doa, dukungan, motivasi yang sudah menemani penulis dari semester 4 hingga saat ini.
11. Untuk sahabatku Icha, Ujul dari SMA dan sahabatku Dira, Farah, Lutvia dari SMP, terimakasih atas segala motivasi, dukungan dan doa selama ini.
12. Keluargaku KKN Simpang Kanan untuk semangat, motivasi dan kebersamaannya selama ini.
13. Teman-teman seperjuangan Matematika 2019 dan Abang Yunda yang telah membantu selama perkuliahan selama ini.
14. Terimakasih kepada NCT, NCT Dream dan Chenle sebagai penyemangat yang selalu memberikan keceriaan penulis selama menyusun skripsi ini
15. Seluruh pihak terkait yang membantu dalam menyelesaikan skripsi ini yang tidak dapat penulis sebutkan satu per satu.

Penulis menyadari skripsi ini masih jauh dari kata sempurna dan masih terdapat banyak kekurangan. Oleh sebab itu, saran dan kritikan yang membangun senantiasa penulis harapkan demi menyempurnakan skripsi ini. Dan semoga dapat bermanfaat bagi pihak yang membutuhkan.

Bandar Lampung, 24 Januari 2023

Penulis

Dini Desita

NPM. 1917031042

DAFTAR ISI

Halaman

DAFTAR TABEL

Tabel Halaman

DAFTAR GAMBAR

Gambar Halaman

I. PENDAHULUAN

1.1 Latar Belakang dan Masalah

Analisis multivariat merupakan salah satu teknik analisis statistik untuk menganalisis data yang terdiri dari banyak variabel. Analisis multivariat dapat menganalisis pengaruh dari beberapa variabel terhadap variabel lainnya secara bersamaan. Pada kenyataannya terdapat masalah yang tidak dapat diselesaikan hanya dengan menghubungkan dua variabel atau melihat pengaruh satu variabel dengan variabel lainnya sehingga untuk mengatasinya dapat menggunakan analisis multivariat. Dari banyaknya metode statistika yang dapat digunakan dalam menganalisis data, analisis multivariat adalah metode yang cocok untuk meringkas data dengan jumlah variabel yang banyak.

Salah satu teknik analisis multivariat yang dapat digunakan untuk mengelompokkan data observasi dalam jumlah besar dan variabel yang banyak adalah analisis *cluster*. Analisis *cluster* adalah salah satu analisis statistik multivariat untuk mengelompokkan sekumpulan objek berdasarkan kemiripan karakteristiknya. Objek tersebut akan dikelompokkan kedalam satu atau lebih kelompok yang berbeda sehingga objek yang terdapat dalam satu kelompok tersebut akan memiliki kemiripan karakteristik antara satu sama lain. Analisis *cluster* berbeda dengan teknik analisis multivariat lainnya karena analisis ini berfokus untuk membandingkan objek berdasarkan karakteristik yang dipakai disetiap objeknya.

Banyak para ahli yang telah mengembangkan analisis *cluster* ini untuk pengetahuan sehingga analisis ini dapat diterapkan di berbagai bidang serta memiliki manfaat untuk mendapatkan karakteristik kelompok yang mudah diperoleh. Analisis *cluster* terdapat 2 metode yaitu metode hierarki (*hierarchical methods*) dan metode non-hierarki (*nonhierarchical methods*).

Secara umum, pengelompokkan dalam analisis *cluster* hanya untuk salah satu tipe data kategorik ataupun numerik saja, namun pada kenyataannya dalam analisis *cluster* hanya fokus pada tipe data numerik. Menurut Hair, *et al.* (2009), analisis *cluster* pada tipe data ketegorik lebih rumit dibandingkan pada tipe data numerik. Sehingga untuk mengatasi tipe data kategorik dikembangan beberapa algoritma seperti metode *K-modes*, *squeezer*, ROCK, *K-prototype*, dan lain-lain. Selain itu, permasalahan yang terjadi dari analisis *cluster* adalah jika jenis data berskala campuran yaitu berisi variabel dengan tipe data kategorik dan data numerik. Dalam perkembangannya, metode analisis yang sering dilakukan untuk mengklasterisasi data yang berskala campuran adalah dengan transformasi data yaitu mengubah tipe data kategorik menjadi tipe data numerik ataupun mengubah tipe data numerik menjadi kategorik. Namun, metode transformasi data memiliki kelemahan yang menyebabkan sulitnya menentukan transformasi data yang tepat agar tidak kehilangan banyak informasi dari data aslinya (Alvionita, dkk., 2017). Berdasarkan kelemahan metode transformasi tersebut maka dikembangkan metode *cluster ensemble* untuk mengelompokkan data campuran.

Metode *cluster ensemble* merupakan teknik *cluster* yang menggabungkan beberapa algoritma pengelompokkan yang berbeda untuk mendapatkan pengelompokkan hasil *clustering* individu yang lebih baik. Ide yang muncul untuk menggabungkan hasil *clustering* yang berbeda (*cluster ensemble*) adalah sebagai pendekatan alternatif untuk meningkatkan kualitas hasil dari beberapa algoritma *clustering*. Dalam penelitian ini akan digunakan metode *cluster ensemble Robust Clustering Using Links* (ROCK). Metode ROCK adalah pengembangan algoritma pengelompokkan yang hierarki untuk menganalisis dengan konsep tautan pada tipe data kategorik. Metode ini memiliki kelebihan

yaitu terdapat nilai akurasi yang lebih baik karena memiliki sifat skalabilitas yang baik serta memiliki kualitas yang lebih baik dalam menangani data kategorik (Reddy & Kavitha, 2012).

Beberapa penelitian yang telah dilakukan oleh peneliti terdahulu mengenai metode *cluster ensemble* adalah penelitian Sharma dan Yadav (2013) melakukan perbandingan antara metode ROCK dan *K-Means* yang membuktikan bahwa metode ROCK lebih optimal dalam analisis cluster. Penelitian Alvionita, dkk. (2017) membandingkan metode *Similarity Weight and Filter Method* (SWFM) dan ROCK untuk mengelompokkan aksesi jeruk, diperoleh bahwa metode ROCK memiliki kinerja yang lebih baik dibandingkan metode SWFM. Penelitian Wulandari, dkk (2020) melakukan pengelompokkan evaluasi daerah tertinggal di Jawa Timur menggunakan metode *ensemble* ROCK dengan metode *K-Means* untuk analisis data numerik dan metode *K-Modes* untuk menganalisis data kategorik menghasilkan pengelompokkan penetapan daerah tertinggal tahun 2020 terdiri dari 4 *cluster* dan nilai *threshold* optimum sebesar 0,04. Kemudian penelitian Ambiani (2021) melakukan pengelompokkan pada data Indeks Kesejahteraan Masyarakat di Indonesia menggunakan metode *cluster ensemble* menghasilkan 2 cluster dengan nilai *threshold* optimum sebesar 0,15.

SMA di Bandar Lampung memiliki sarana dan prasarana yang berbeda-beda dari segi kualitas antara satu dengan lainnya. Kualitas SMA dapat dikategorikan menjadi SMA unggul dan SMA tidak unggul. Kualitas SMA dapat ditentukan dari sarana dan prasarana SMA seperti jumlah siswa, jumlah guru, jumlah ruang kelas, luas lahan, besar daya listrik, akreditasi SMA, status SMA dan waktu penyelenggaraan SMA. Berdasarkan data dari *website* resmi KEMENDIKBUD terdapat 67 SMA dengan status SMA Negeri dan SMA Swasta yang tersebar diberbagai wilayah yang ada di Bandar Lampung. Tetapi pada penelitian ini penulis membatasi hanya 60 data SMA yang digunakan, setelah dilakukannya pembersihan data dari *missing value*.

Oleh karena itu, dalam penelitian ini akan dilakukan implementasi metode *cluster ensemble* ROCK untuk mengelompokkan Sekolah Menengah Atas

(SMA) di Bandar Lampung dengan menggunakan algoritma *K-Medoids* untuk analisis data numerik dan metode ROCK untuk analisis data kategorik. Penelitian ini diharapkan mampu memberikan hasil *cluster* yang terbaik dalam pengelompokkan SMA di Bandar Lampung.

1.2 Tujuan Penelitian

Tujuan dari penelitian ini adalah :

1. Menerapkan metode *ensemble* ROCK pada data campuran untuk mengelompokkan Sekolah Menengah Atas (SMA) di Bandar Lampung.
2. Memperoleh jumlah *cluster* yang terbaik berdasarkan ratio s_w dan s_b dari pengelompokkan *ensemble* metode ROCK.

1.3 Manfaat Penelitian

Adapun manfaat dari penelitian ini adalah :

1. Menambah pengetahuan dalam penerapan *ensemble clustering* pendekatan metode ROCK.
2. Memberikan tambahan informasi mengenai metode *clustering* jenis data campuran untuk dijadikan referensi pada penelitian selanjutnya.
3. Mendapatkan karakteristik yang terbaik terhadap pengelompokkan Sekolah Menengah Atas (SMA) berdasarkan *cluster* yang terbentuk.
4. Membantu memberikan saran bagi Pemerintah Kota Bandar Lampung serta Dinas Pendidikan dan Kebudayaan Kota Bandar Lampung dalam menentukan kebijakan sebagai pertimbangan dalam mengatasi Sekolah Menengah Atas (SMA) yang masih minim sarana prasarana,kurang meratanya distribusi guru tiap sekolah serta kualitas SDM SMA.

II. TINJAUAN PUSTAKA

2.1 Analisis Multivariat

Multivariat berasal dari kata “multi” yang memiliki arti banyak dan “variat” yang memiliki arti variabel serta diantara variabel tersebut saling berhubungan. Berdasarkan definisi di atas maka analisis multivariat adalah analisis statistika yang terdiri dari banyak variabel serta dianalisis secara bersamaan di setiap pengamatannya sehingga terdapat korelasi antar variabel (Johnson & Wichern, 2015). Data multivariat terdapat di semua cabang ilmu pengetahuan, mulai dari psikologi hingga biologi serta metode analisis multivariat merupakan salah satu bidang statistik yang sangat penting.

Menurut Sitepu, dkk. (2011), metode analisis multivariat secara umum dibagi menjadi dua tipe, yaitu:

a. Metode Dependensi (*Dependence Method)*

Suatu teknik multivariat yang menggunakan analisis ketergantungan untuk menjelaskan suatu variabel terikat (*dependent variable)* serta variabel lainnya sebagai variabel bebas (*independent variable*). Metode ini terdapat empat jenis yaitu analisis regresi berganda, analisis diskriminan berganda, analisis multivariat varians dan analisis korelasi kanonikal.

b. Metode Interdependensi (*Interdependence Method*)

Suatu teknik multivariat yang digunakan untuk menjelaskan semua variabel atau pengelompokkan secara bersamaan serta tidak terdapat variabel terikat ataupun variabel bebas. Metode ini terdapat tiga jenis yaitu analisis faktor, analisis *cluster* dan penskalaan *multidimensional.*

Menurut Chatfield & Collins (2018), data multivariat diperoleh berdasarkan hasil pengukuran terhadap *n* (jumlah individu atau objek) dengan *p* (jumlah variabel) sehingga data tersebut disajikan dalam bentuk matriks $\boldsymbol{X}$ berukuran *n* x *p* , yang dapat ditulis sebagai berikut:

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{ij} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2j} & \dots & x_{2p} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{i1} & x_{12} & \dots & x_{ij} & \dots & x_{ip} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nj} & \dots & x_{np} \end{bmatrix}$$

dengan $i = 1, 2, \dots, n$ dan $j = 1, 2, \dots, p$.

Dalam analisis *cluster*, data multivariat dapat digunakan sebagai input yang dinotasikan dalam bentuk matriks $\boldsymbol{X} = \{x_{ij}\}$, dengan x_{ij} merupakan pengamatan ke-*i* terhadap observasi pada variabel ke-*j*.

2.2 Analisis *Cluster*

Analisis *cluster* adalah salah satu teknik statistika interdependensi untuk membentuk kelompok (*cluster*) dari objek data multivariat (Hardle & Simar, 2019). Analisis *cluster* menganalisis dengan mengelompokkan n objek ke dalam m *cluster* ($m \leq n$) berdasarkan kesamaan karakteristiknya. Tujuan dari analisis *cluster* adalah untuk meringkas data dengan cara mengelompokkan suatu objek berdasarkan karakteristik tertentu diantara objek – objek tersebut. Setelah *n* objek tersebut dikelompokkan kedalam m kelompok berdasarkan *p* variabel maka diperoleh kelompok objek dengan nilai yang relatif sama. Hal tersebut memudahkan dalam hal interpretasi karena objek yang berada dalam satu *cluster* memiliki peluang muncul bersamaan yang cukup tinggi pada satu individu.

Menurut Chrisinta, dkk. (2020), analisis *cluster* adalah analisis yang digunakan untuk mengelompokkan objek-objek berdasarkan ukuran kemiripan atau ketidakmiripan. Semakin terdapat perbedaan karakteristik suatu objek maka

semakin kecil pula kemungkinan untuk berada dalam satu kelompok. Analisis *cluster* mengelompokkan objek – objek berdasarkan jenis data yang dimiliki. Beberapa kelebihan dalam analisis *cluster* adalah dapat mengelompokkan data observasi dalam jumlah yang besar dan variabel yang relatif banyak sehingga data yang akan direduksi lebih mudah dianalisis. Dalam proses analisis *cluster*, pengelompokkan dilakukan dengan mengetahui pola data dari objek pengamatan untuk mendapatkan hasil *cluster* yang optimum. Terdapat dua macam metode dalam analisis *cluster* yaitu metode hierarki yang mengelompokkan suatu pengamatan secara bertahap dan metode non hierarki dilakukan dengan melakukan partisi pada ruang sampel.

Menurut Goreti, dkk. (2017), ciri-ciri *cluster* yang baik adalah memiliki homogenitas (kesamaan) yang tinggi antar anggota dalam satu *cluster* (*within-cluster*) dan heterogenitas (perbedaan) yang tinggi antar *cluster* yang satu dengan *cluster* yang lainnya (*between-cluster*).

Menurut Suhaeni, dkk. (2018), dasar dari pengelompokkan yaitu sifat kemiripan atau ketidakmiripan antar objek. Jika objek berada dalam kelompok yang sama maka terdapat kemiripan dibandingkan objek antar kelompok. Hasil dari analisis *cluster* dapat dipengaruhi oleh objek yang dikelompokkan, ukuran kemiripan atau ketidakmiripin, skala pengukuran serta metode *clustering* yang digunakan. Algoritma pengelompokkan menggunakan ukuran kemiripan yang berguna untuk menggabungkan objek dari suatu data. Untuk data yang bersifat kategorik biasanya menggunakan ukuran kemiripan, sedangkan untuk data yang bersifat numerik menggunakan ukuran ketidakmiripan.

Semakin besar ukuran ketidakmiripan antara dua objek, maka akan semakin besar perbedaan antara kedua objek tersebut, sehingga tidak dapat berada dalam kelompok yang sama (Johnson & Wichern, 2015). Ukuran kemiripan dan ketidakmiripan dapat diukur dengan menggunakan ukuran jarak sehingga diperlukan suatu alat ukur khusus dalam menentukan jarak antara objek pengamatan dalam penelitian. Metode pengukuran jarak antara objek ke-i (x_i)

dengan objek ke-*j* (x_j) yang dapat digunakan berdasarkan karakteristik variabel sebagai berikut:

a. Metode Pengukuran Jarak untuk Variabel Kategorik Biner

Jika variabel yang diamati berupa variabel kategorik biner dengan dua macam karakter yang berbeda (0,1) maka variabel seperti dalam tabel kontengensi berikut:

Tabel 1. Tabel Kontingensi Data Biner

Kategori x_i	**Kategori x_j**		**Total**
	1	0	
1	a	b	a+b
0	c	d	c+d
Total	a+c	b+d	a+b+c+d

Sedangkan untuk pengukuran data biner dalam perhitungan jarak antar variabel dapat menggunakan ukuran sebagai berikut:

Tabel 2. Ukuran Jarak Data Biner

Jenis	**Rumus**
Russel and Rao	$RR(x_i, x_j) = \frac{a}{a+b+c+d}$
Simple Matching	$SM(x_i, x_j) = \frac{a+d}{a+b+c+d}$
Jaccard	$JACCARD(x_i, x_j) = \frac{a}{a+b+c}$
Dice Czekkanowski	$DICE(x_i, x_j) = \frac{2a}{2a+b+c}$

b. Metode Pengukuran Jarak untuk Variabel Kategorik Nominal

Pengukuran jarak dengan variabek nominal memiliki pengukuran konsep yang sama dengan *simple matching coefficient* ataupun *dice*, dengan pada kategorinya dapat memiliki lebih dari dua macam. Jika jumlah variabel sebanyak *m*, maka persamaan untuk pengukuran jarak variabel nominal antara $x_i$dan x_j adalah sebagai berikut:

$$sim(x_i, x_j) = \frac{1}{m}\sum_{l=1}^{m} S_{ijl} \quad (2.\ 1)$$

dengan $S_{ijl} = 1$ jika $x_{il} = x_{jl}$ dan $S_{ijl} = 0$ jika $x_{il} \neq x_{jl}$.

c. Metode Pengukuran Jarak Untuk Variabel Kategorik Ordinal

Konsep pengukuran jarak yang digunakan pada pengamatan dengan variabel ordinal sama dengan metode untuk data numerik, dengan kategorinya dinyatakan sebagai bilangan bulat. Salah satu metode yang dapat digunakan untuk variabel ordinal adalah jarak *manhattan.* Dengan jumlah variabel sebanyak m, maka pengukuran jarak x_i dan x_j pada variabel ordinal dihitung dengan persamaan sebagai berikut:

$$sim(x_i, x_j) = \sum_{i=1}^{m} |x_{il} - x_{jl}| \quad (2.\ 2)$$

d. Metode Pengukuran Jarak untuk Variabel Numerik

Pengamatan dengan jenis data numerik dapat menggunakan jarak Euclidean. Misalkan terdapat dua observasi dengan variabel – variabel berdimensi *m* yaitu $x_i = [x_1, x_2, \dots, x_m]^T$ dan $x_j = [x_1, x_2, \dots, x_m]^T$. Sehingga rumus untuk menghitung jarak Euclidean dapat dihitung dengan persamaan berikut:

$$d_{ij} = \sqrt{(x_i - x_j)^T (x_i - x_j)} \quad (2.\ 3)$$

2.2.1 Normalisasi Data dan Denormalisasi Data

Normalisasi adalah suatu proses penskalaan pada kolom berdasarkan atribut untuk diubah menjadi nilai numerik pada rentang tertentu. Dalam penelitian ini metode normalisasi yang digunakan adalah *Z-score. Z-score* adalah suatu proses normalisasi berdasarkan nilai rata-rata (mean) dan standar deviasi dari data (Nasution, dkk., 2019). Proses normalisasi data diperlukan jika terdapat data di antara variabel-variabel memiliki perbedaan ukuran satuan yang besar serta

dilakukan normalisasi agar menghasilkan pengolahan data yang akurat. Pada umumnya normalisasi *Z-score* dilakukan pada jenis variabel data numerik.

Proses normalisasi tersebut bertujuan untuk mengganti data asli kedalam bentuk lain dengan skala yang sama sehingga memudahkan dalam proses penelitian. Metode normalisasi *Z-score* berguna untuk data yang belum diketahui nilai aktual minimum dan maksimumnya serta digunakan jika rentang antar beberapa variabel sangat jauh. Metode ini memiliki nilai yang stabil terhadap data *outliers*. Perhitungan normalisasi *Z-score* dapat dilihat pada persamaan berikut.

$$Z = \frac{x - \bar{x}}{s} \qquad (2.\ 4)$$

Keterangan:

Z = Nilai data yang baru dari hasil *Z-score*

x = Nilai yang akan dinormalisasi

$\bar{x}$ = Nilai rata-rata setiap kolom

s = Nilai standar deviasi

Denormalisasi adalah proses untuk membangkingkatkan nilai yang telah dinormalisasi menjadi nilai *real* atau asli. Tujuan denormalisasi adalah untuk mempermudah menginterpretasi hasil *output* serta agat mudah dipahami. Perhitungan denormalisasi dapat dilihat pada persamaan berikut

$$x = (Z \times s) + \bar{x} \qquad (2.\ 5)$$

2.2.2 Metode Hierarki

Metode hierarki adalah suatu metode pengelompokkan data yang dimulai dengan mengelompokkan dua atau lebih objek yang memiliki kesamaan paling dekat antar objek. Metode hierarki digunakan jika belum ada informasi jumlah *cluster* yang akan dipilih. Metode ini menghasilkan urutan partisi dengan menggabungkan atau membagi *cluster*. Hasil dari metode ini dapat disajikan dalam bentuk dendogram. Dendogram merupakan *representative* visual dari

seluruh tahapan yang menunjukkan bagaimana *cluster* terbentuk. Metode hierarki secara umum dibagi menjadi dua yaitu metode *agglomerative* (penggabungan) dan metode *divisive* (pemecahan). Terdapat lima metode hierarki sebagai berikut:

a. Pautan Tunggal (*Single Linkage*)
   Metode ini mengelompokkan dua objek yang mempunyai jarak terdekat lebih dahulu. Jadi, pada setiap tahapannya banyak *cluster* yang berkurang satu. Hasil dari metode ini dapat disajikan dalam bentuk suatu dendogram atau diagram pohon. Cabang-cabang tersebut bergabung pada simpul yang posisinya sepanjang suatu sumbu jarak (kemiripan) yang menunjukkan tingkat dimana penggabungan terjadi.

b. Pautan Lengkap (*Complete Linkage)*
   Metode *complete linkage* didasarkan pada jarak maksimum. Metode ini mengelompokkan dua objek yang mempunyai jarak terjauh terlebih dahulu. Kemudian proses dilanjutkan untuk jarak antar variabel yang makin dekat.

c. Pautan Rata-rata (*Average Linkage*)
   Metode *average linkage* merupakan variasi dari metode *single linkage* yaitu menghitung jarak antar dua *cluster.* Metode ini mengelompokkan dua objek berdasarkan jarak rata-rata yang didapat dengan melakukan rata-rata semua jarak antar objek terlebih dahulu. Metode ini menggunakan kriteria rata-rata jarak seluruh individu dalam suatu *cluster* dengan jarak seluruh individu dalam *cluster* yang lain.

d. Metode Ward (*Ward's Method)*
   Metode ward melakukan proses pengelompokkan dengan menggunakan pendekatan analisis varians untuk menghitung jarak antar cluster dengan meminimumkan jumlah kuadrat. Metode ward merupakan bagian dari metode pengelompokkan yang mengelompokkan sebuah objek menjadi satu *cluster* dengan banyaknya *cluster* tidak diketahui. Metode ini didasarkan pada kriteria

Sum Square Error (SSE) dengan ukuran kehomogenan antara dua pengamatan berdasarkan jumlah kuadrat yang paling minimal.

e. Metode Pusat (*Centroid Method)*

Pada metode ini, jarak antara dua *cluster* adalah jarak diantara dua *centroid cluster* tersebut. *Centroid* merupakan rata-rata jarak yang ada pada sebuah *cluster.* Dengan metode ini, setiap terjadi *cluster* yang baru maka akan segera terjadi perhitungan ulang *centroid,* sampai terbentuk *cluster* yang tetap.

2.2.3 Metode Non Hierarki

Analisis *cluster* dengan metode non hierarki merupakan metode *cluster* yang menentukan jumlah *cluster* terlebih dahulu yang diinginkan secara manual. Untuk mengelompokkan n objek kedalam *k cluster* ($k < n$), di mana nilai *k* telah ditentukan terlebih dahulu merupakan tujuan dari metode non hierarki. Metode non hierarki ini dirancang untuk mengelompokkan per item menjadi suatu kumpulan *k cluster.* Dibandingkan metode hierarki, metode non hierarki lebih unggul dalam proses *clustering* jika jumlah objek atau observasi dalam skala ukuran yang besar karena memiliki kecepatan yang lebih tinggi. Metode yang termasuk dalam metode non hierarki diantaranya yaitu *K-Means, K-Modes, K-Medoids.* Terdapat tiga pendekatan untuk menempatkan masing-masing observasi kedalam satu *cluster:*

a. *Sequential Threshold*

Metode ini memulai dengan pemilihan satu *cluster* dan menenmpatkan semua objek yang berada pada jarak tertentu kedalamnya sebagai nilai awal *cluster.* Jika semua objek yang berada pada jarak tertentu telah dimasukkan, *cluster* yang kedua dipilih kemudian ditempatkan ke semua objek yang berjarak tertentu kedalamnya. Setelah itu, *cluster* ketiga dipilih dan proses dilanjutkan seperti yang sebelumnya.

b. *Parallel Threshold*

Metode ini berbanding terbalik dengan metode *Sequential Threshold* karena metode ini memilih sejumlah *cluster* secara bersamaan dan menempatkan objek-objek kedalam *cluster* yang memiliki jarak terdekat *cluster* awal. Pada saat proses berlangsung, jarak antar *cluster* awal dapat ditentukan untuk memasukkan beberapa objek kedalam *cluster-cluster.* Sisa objek-objek tidak akan dikelompokkan jika berada di luar jarak teretentu dari sejumlah *cluster.*

c. *Optimization*

Metode ini hampir sama dengan metode *Sequential Threshold* dan *Parallel Threshold*. Perbedaannya hanya memungkinkan untuk menempatkan kembali objek-objek kedalam *cluster* yang lebih dekat.

2.3 Metode dan Validasi Pengelompokkan (*Clustering*)

Pada tahap pengelompokkan dalam analisis *cluster* dapat dibedakan menurut jenis data berdasarkan variabelnya. Secara umum, pada saat menganalisis menggunakan analisis *cluster* hanya terfokus pada data numerik saja, namun terdapat kasus dengan data kategorik ataupun data yang berskala campuran. Analisis *cluster* data kategorik tidak sama dengan menganalisis data numerik karena terdapat sifat khusus dalam data kategorik, sehingga dalam pengelompokkan data kategorik akan lebih rumit dibandingkan data numerik (Hair *et al*., 2009)

2.3.1 Pengelompokkan Data Numerik

Pengelompokkan data numerik dilakukan berdasarkan pengukuran jarak. Analisis *cluster* pada data numerik menggunakan ukuran ketidakmiripan pada objek pengamatan. Semakin besar perbedaan ukuran jarak maka pada objek

pengamatan akan terdapat sedikit pula kesamaannya begitupun sebaliknya. Dalam penelitian ini akan digunakan analisis *cluster* dengan metode *K-Medoids.*

Menurut Sindi, dkk. (2020), *K-Medoids* atau *Partitioning Around Medoids* (PAM) merupakan salah satu algoritma teknik *cluster* yang dapat digunakan untuk menemukan *medoids* dalam sebuah *cluster* yang termasuk pada titik pusat dari suatu *cluster*. Algoritma *K-Medoids* ini diusulkan pada tahun 1987 yang dikembangkan oleh Leonard Kaufmann dan Peter J. Rousseeuw. *K-Medoids* adalah algoritma dalam analisis *cluster* yang mirip dengan *K-Means*.

Menurut Sihombing, dkk. (2019), perbedaan dari algoritma *K-Medoids* dan *K-Means* adalah algoritma *K-Medoids* menggunakan objek sebagai perwakilan (*medoid*) sebagai pusat *cluster* untuk setiap *cluster*, sedangkan *K-Means* menggunakan nilai rata-rata (*mean*) sebagai pusat *cluster*. Kelebihan dari algoritma *K-Medoids* yaitu tidak sensitif terhadap *outlier*, dapat mengurangi *noise*, dan jika dibandingkan dengan algoritma *K-Means*, *K-Medoids* lebih unggul dalam melakukan klasterisasi dataset heterogen/campuran, pemilihan *cluster*, kompleksitas antar ruang *cluster*, dan waktu eksekusi.

Algoritma *K-Medoids* tidak menentukan nilai rata-rata dari objek dalam *cluster* sebagai titik acuan, tetapi menggunakan *medoid* (*median*) yang merupakan objek yang paling terletak dipusat sebuah *cluster* (Religia dkk., 2020). Dengan demikian, metode partisi masih dapat dilakukan berdasarkan prinsip meminimalkan jumlah dari ketidaksamaan antara setiap objek dan titik acuan yang sesuai (*medoid*). Strategi dasar dari algoritma *K-Medoids* adalah untuk menemukan *cluster k* pada objek *n* dengan terlebih dahulu menemukan objek awal (*medoid*) secara acak sebagai perwakilan untuk setiap *cluster* (Defiyanti, dkk., 2017). Setiap objek yang tersisa dikelompokan dengan *medoid* yang paling mirip. Perhitungan *medoid* dapat menggunakan Persamaan 2.6.

$$medoid_d(S) = \arg\min_{x \in S} \sum_{y=S}^{n} d(x, y) \qquad (2.\ 6)$$

dimana:

S = Set pada data point ($x_1, x_2, \ldots, x_n$)

d = Nilai absolut dari jarak yang digunakan.

Jarak yang dapat digunakan pada variabel dengan jenis data numerik adalah jarak Euclidean. Menurut Nishom (2019), jarak Euclidean adalah salah satu metode pengukuran jarak yang digunakan untuk mengukur jarak dari dua buah titik dalam Euclidean *space* (termasuk bidang Euclidean dua dimensi atau lebih). Perhitungan jarak Euclidean menggunakan persamaan sebagai berikut:

$$d_{ij} = \sqrt{\sum_{a=0}^{n} \left(x_{ia} - x_{ja}\right)^2} \qquad (2.\ 7)$$

Menurut Bhat (2014), tahapan dalam algoritma *K-Medoids* adalah sebagai berikut:

a. Inisialisasi pusat *cluster* sebanyak k (jumlah *cluster*)
b. Hitung setiap objek ke *cluster* terdekat menggunakan persamaan ukuran jarak Euclidean menggunakan Persamaan 2.7
c. Setelah menghitung jarak Euclidean, inisialisasikan pusat *cluster* baru secara acak pada masing-masing objek sebagai kandidat *non medoids.*
d. Hitung jarak setiap objek yang berada pada masing masing *cluster* dengan kandidat *non medoids.*
e. Hitung total simpangan (S) dengan menghitung

   S = total *distance* baru – total *distance* lama

   Jika S < 0 maka tukar objek dengan data *cluster non medoids* untuk membentuk sekumpulan *k* objek baru sebagai *medoids*
f. Ulangi Langkah (c)-(e) hingga tidak terjadi perubahan pada *medoid*, sehingga didapatkan *cluster* beserta anggota masing-masing *cluster.*

Setelah proses pengelompokkan selesai dilakukan, tahapan selanjutnya yaitu validasi pengelompokkan optimum. Dalam tahapan ini akan didapatkan jumlah kelompok yang paling optimum. Metode yang digunakan dalam penelitian ini adalah *sillhoutte coefficient.*

Sillhoutte coefficient merupakan salah satu metode evaluasi yang digunakan untuk menguji kualitas dan kekuatan dari sebuah *cluster*. Menurut Pramesti dkk. (2017) metode ini merupakan gabungan dari dua metode yaitu metode *cohesion* yang berfungsi untuk mengukur seberapa dekat relasi antara objek dalam sebuah *cluster* dan metode *separation* yang berfungsi untuk mengukur seberapa jauh sebuah *cluster* terpisah dengan *cluster* lain. Untuk menghitung nilai *silhouette coefficient*, dapat dilakukan perhitungan dengan mencari nilai maksimal dari nilai *sillhouette coeffecient* global dari jumlah *cluster* 2 sampai jumlah *cluster* $n-1$ seperti pada persamaan berikut:

$$SC = maks_k\ SI\ (k) \tag{2. 8}$$

dengan,

SC = sillhouette coefficient

SI = sillhouette index global

k = jumlah *cluster*

Untuk menghitung nilai *SI* dari sebuah data ke-*i*, ada 2 komponen yaitu a_i dan b_i. Nilai $a_i$adalah rata-rata jarak ke-*i* terhadap semua data lainnya dalam satu *cluster*, sedangkan b_i diperoleh dengan menghitung rata-rata jarak data ke-*i* terhadap semua data dari *cluster* lainnya yang tidak satu *cluster* dengan data ke-*i*, lalu diambil yang terkecil. Berikut persamaan untuk menghitung nilai a_i.

$$a_i{}^j = \frac{1}{m_j - 1} \sum_{r=1, r\neq 1}^{m_j} d\left(x_i{}^j, x_r{}^j\right) \tag{2. 9}$$

dengan,

$a_i{}^j$ = rata-rata jarak data ke-*i* terhadap semua data dalam satu *cluster*

m_j = jumlah data dalam *cluster* ke-*j*

j = *cluster*

i = *index* data ($i = 1,2, \dots, m_j$)

$d\left(x_i{}^j, x_r{}^j\right)$ = jarak data ke-*i* dengan data ke-*r* dalam satu *cluster j*

Kemudian menghitung nilai b_i dengan persamaan sebagai berikut :

$$b_i{}^j = min_{\substack{n=1,\dots k \\ n\neq j}} \left\{ \frac{1}{m_n} \sum_{r=1, r\neq 1}^{m_n} d\left(x_i{}^j, x_r{}^n\right) \right\} \quad (2.\ 10)$$

dengan,

$b_i{}^j$ = rata-rata jarak data ke-*i* terhadap semua data yang tidak dalam satu *cluster* dengan data ke-*i*

m_n = jumlah data dalam *cluster* ke-*n*

j = *cluster*

i = *index* data (i = 1,2, …, m_j)

$d\left(x_i{}^j, x_r{}^n\right)$ = jarak data ke-*i* dengan data ke-*j* dalam satu *cluster n*

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

$$SI_i^j = \frac{b_i^j - a_i^j}{\max\{a_i^j, b_i^j\}} \quad (2.\ 11)$$

dengan,

SI_i^j = *sillhouette index* data ke-i dalam satu *cluster*

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

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

Kemudian, rumus perhitungan untuk mendapatkan nilai SI_j dapat dilihat pada persamaan berikut:

$$SI_j = \frac{1}{m_j} \sum_{i=1}^{m_j} SI_i^j \quad (2.\ 12)$$

dengan,

SI_j = rata-rata *sillhouette index cluster j*

SI_i^j = *sillhouette index* data ke-*i* dalam *cluster j*

m_j = jumlah data dalam *cluster* ke-*j*

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

Selanjutnya adalah rumus perhitungan mendapatkan nilai *SI* global dengan persamaan berikut:

$$SI = \frac{1}{k}\sum_{j=1}^{k} SI_j \qquad (2.\ 13)$$

dengan,

SI = rata-rata *sillhouette index* dari dataset

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

k = jumlah *cluster*

Menurut Kaufman dan Roesseeuw (2009), *silhoutte coeffecient* dapat dilihat berdasarkan kriteria *subjektif* pengukuran pengelompokkan pada Tabel 3.

Tabel 3. Kriteria Pengukuran *Silhoutte Coeffecient*

Nilai *Silhoutte Coeffecient*	**Kriteria**
0,71 – 1,00	Memiliki struktur yang kuat
0,51 – 0,70	Memiliki struktur yang baik
0,26 – 0,50	Memiliki struktur yang lemah
≤ 0,25	Memiliki struktur yang buruk

2.3.2 Pengelompokkan Data Kategorik

Pengelompokkan data kategorik dilakukan berdasarkan pengukuran jarak data kategorik. Analisis *cluster* pada data kategorik menggunakan ukuran ketidakmiripan pada objek pengamatan, kemudian dilakukan tahap pengelompokkan dengan menggunakan metode hierarki ataupun metode non hierarki. Metode ROCK diperkenalkan oleh Guha, Rastologi dan Shim pada tahun 1999. Metode ROCK adalah perkembangan dari metode pengelompokkan hirarki *agglomerative* yang digunakan untuk data kategorik. Metode ini menggunakan konsep *link* (tingkat hubungan) sebagai ukuran kemiripan untuk

membentuk *cluster*-nya (Rumiati, dkk., 2022). Objek pengamatan yang memiliki *link* tinggi akan dikelompokkan dalam satu *cluster*, sedangkan objek pengamatan yang memiliki *link* rendah akan dipisahkan dari pengelompokkan data tersebut. Menurut Safrullah, dkk. (2020), jumlah *link* antar pengamatan bergantung pada nilai *threesold* (θ) yang telah ditentukan. Nilai (θ) merupakan parameter untuk menyatakan adanya *link* antar pengamatan.

Kelebihan metode ROCK adalah dalam menangani *outlier* cukup efektif karena pemangkasan *outlier* memungkinkan untuk membuang yang tidak ada hubungan sehingga titik tersebut tidak berpartisipasi dalam pengelompokkan. Namun, dalam beberapa situasi, *outlier* dapat hadir sebagai *cluster-cluster* yang kecil (Guha, *et al.*, 2000). Tahapan analisis *cluster* data kategorik dengan metode ROCK adalah sebagai berikut:

a) Menghitung *Similaritas*

Menggunakan rumus *Jaccard coefficient.* Ukuran kemiripan antara objek ke-*i* dan objek ke-*j* dihitung dengan persamaan berikut:

$$sim(X_i, X_j) = \frac{|X_i \cap X_j|}{|X_i \cup X_j|}, X_i \neq X_j \qquad (2.\ 14)$$

dengan,

i = 1,2,3, …, n dan j = 1,2,3,…,n

X_i = Himpunan pengamatan ke-i dengan $X_i = \{x_{1i}, x_{2i}, x_{3i}, \dots, x_{m_{kategori}i}\}$

X_j = Himpunan pengamatan ke-j dengan $X_j = \{x_{1j}, x_{2j}, x_{3j}, \dots, x_{m_{kategori}j}\}$

b) Menentukan Tetangga (*Neighbors*)

Pengamatan X_i dan X_j dinyatakan sebagai tetangga jika nilai $sim(X_i, X_j) \geq \theta$ dan dilambangkan dengan nilai 1. Sebaliknya jika $sim(X_i, X_j) \leq \theta$ dan dilambangkan dengan nilai 0. Nilai *threshold* (θ) yang digunakan biasanya berkisar antara 0 sampai 1 yang disesuaikan dengan keadaan data.

c) Menghitung *Link*

Link (X_i, X_j) antar objek diperoleh dari jumlah tetangga yang sama (*common neighbor)* antara X_i dan X_j. Besarnya *link* dipengaruhi oleh nilai *threshold* (θ).

Nilai *threshold* (θ)merupakan parameter yang ditentukan oleh peneliti untuk mengontrol seberapa dekat hubungan antara objek. Nilai *threshold* (θ) yang digunakan biasanya berkisar antara 0 sampai 1.

Metode ROCK menggunakkan informasi tentang *link* sebagai ukuran kemiripan antar objek. Jika terdapat objek pengamatan X_i, X_j dan X_k dimana X_i tetangga dari X_j dan X_j tetangga dari X_k maka X_i dikatakan memiliki *link* dengan X_k walaupun X_i bukan tetangga dari X_k.

Cara menghitung *link* untuk semua kemungkinan pasangan dari *n* objek adalah menggunakan matriks **A**. Matriks **A** merupakan matriks berukuran *n x n* yang bernilai 1 jika X_i dan X_j dinyatakan mirip (tetangga) dan bernilai 0 jika X_i dan X_j tidak mirip (bukan tetangga). Matriks *link* diperoleh dari perkalian matriks tetangga (**A**) dengan matriks tetangga (**A**) itu sendiri. Apabila nilai *link* (X_i, X_j) besar maka kemungkinan besar X_i dan X_j berada pada *cluster* yang sama.

d) Menentukan *Local Heap*

Local heap adalah nilai ukuran kebaikan (*goodness measure)* untuk setiap kelompok dengan kelompok lainnya jika *link* $\neq 0$. *Goodness measure* merupakan persamaan yang menghitung jumlah *link* dibagi dengan kemungkinan *link* yang terbentuk berdasarkan ukuran kelompoknya. *Goodness measure* dapat dihitung dengan persamaan sebagai berikut:

$$g(C_i, C_j) = \frac{link(C_i, C_j)}{(n_i + n_j)^{1+2f(\theta)} - {n_i}^{1+2f(\theta)} - {n_j}^{1+2f(\theta)}} \quad (2.\ 15)$$

dengan,

$link(C_i, C_j) = \sum_{X_i \in C_i, X_j \in C_j} link(C_i, C_j)$

= jumlah *link* semua kemungkinan pasangan objek yang terdapat pada C_i dan C_j

n_i = jumlah anggota dalam *cluster* ke-*i*

n_j = jumlah anggota dalam *cluster* ke-*j*

$f(\theta)$ = $\frac{1-\theta}{1+\theta}$

e) Menentukan *global heap.* *Global heap* adalah nilai maksimum *goodness measure* antar kolom di baris ke-*i.*
f) Ulangi Langkah (d)-(e) hingga mendapatkan nilai maksimum di *global heap* dan *local heap.*
g) Hentikan melakukan pengelompokkan metode ROCK jika jumlah dari *cluster* yang diharapkan sudah terpenuhi dan tidak ada *link* antar kelompok.

Setelah dilakukan tahap pengelompokkan maka akan dilakukan validitas pengelompokan. Pada metode ROCK validitas pengelompokkan dilakukan dengan menghitung nilai *ratio* s_w dan s_b (Wulandari, dkk., 2020).

Jika terdapat sebanyak *n* pengamatan dengan n_k merupakan jumlah pengamatan dengan kategori ke-*k* dimana *k* = 1,2,3, …, K dan $\sum_{k=1}^{K} n_k = n$. Selanjutnya, n_{kc} adalah jumlah pengamatan pada kategori ke-*k* serta kelompok ke-*c* dimana $c = 1,2,3,\dots,C$ dengan C merupakan jumlah kelompok terbentuk, sehingga $nc = \sum_{k=1}^{K} n_{kc}$ merupakan jumlah pengamatan pada kelompok ke-*c* dan $n_k = \sum_{c=1}^{C} n_{kc}$ merupakan jumlah pengamatan pada kategori ke-*k.* Total jumlah pengamatan dapat dituliskan pada persamaan berikut:

$$n = \sum_{c=1}^{C} n_c = \sum_{k=1}^{K} n_k = \sum_{k=1}^{K}\sum_{c=1}^{C} n_{kc} \qquad (2.\ 16)$$

a. *SST* (*Sum Square Total*)

Sum Square Total untuk variabel data kategorik dapat dituliskan pada persamaan berikut:

$$SST = \frac{n}{2} - \frac{1}{2n}\sum_{k=1}^{K} {n_k}^2 \qquad (2.\ 17)$$

b. *SSW* (*Sum Square Within*)

Sum Square Within untuk variabel data kategorik dapat dituliskan pada persamaan berikut:

$$SSW = \sum_{c=1}^{C}\left(\frac{n_c}{2} - \frac{1}{2n_c}\sum_{k=1}^{K} {n_{kc}}^2\right)$$
$$= \frac{n}{2} - \frac{1}{2}\sum_{c=1}^{C}\frac{1}{n_c}\sum_{k=1}^{K} {n_{kc}}^2 \quad (2.\ 18)$$

c. *SSB* (*Sum Square Between*)

Sum Square Between untuk variabel data kategorik dapat dituliskan pada persamaan berikut:

$$SSB = \frac{1}{2}\left(\sum_{c=1}^{C}\frac{1}{n_c}\sum_{k=1}^{K} {n_{kc}}^2\right) - \frac{1}{2n}\sum_{k=1}^{K} {n_k}^2 \quad (2.\ 19)$$

Selanjutnya untuk *MST* (*Mean Square Total*) dapat dihitung menggunakan Persamaan 2.20, *MSW* (*Mean Square Within*) menggunakan Persamaan 2.21, dan *MSB* (*Mean Square Between*) menggunakan Persamaan 2.22 sebagai berikut:

$$MST = \frac{SST}{(n-1)} \quad (2.\ 20)$$

$$MSW = \frac{SSW}{(n-C)} \quad (2.\ 21)$$

$$MSB = \frac{SSB}{(C-1)} \quad (2.\ 22)$$

Kemudian untuk nilai s_w (simpangan baku kelompok) dan s_b(simpangan baku antar kelompok) untuk data kategorik dapat dihitung menggunakan persamaan berikut:

$$S_w = [MSW]^{1/2} \quad (2.\ 23)$$

$$S_b = [MSB]^{1/2} \quad (2.\ 24)$$

Semakin kecil nilai ratio antara s_w dan S_b maka semakin baik pula kinerja pengelompokkan dari hasil *cluster* yang didapatkan sehingga memiliki arti bahwa terdapat homogenitas maksimum dalam kelompok dan heterogenitas maksimum antar kelompok.

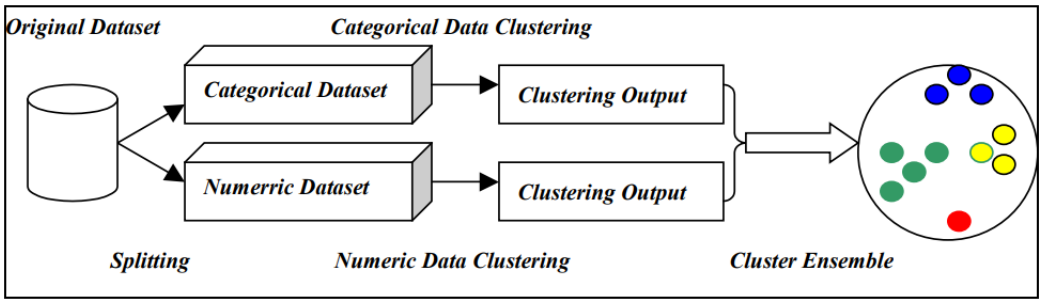
2.4 *Cluster Ensemble*

Pengelompokan data berskala campuran numerik dan kategorik dapat dilakukan dengan membagi data menjadi data murni numerik dan murni kategorik. Misalkan terdapat data dengan variabel berskala campuran sebanyak *m*, dengan $m_{numerik}$ adalah jumlah variabel dengan data murni numerik dan $m_{kategorik}$ adalah jumlah variabel dengan data murni kategorik, sehinga didapatkan $m = m_{numerik} + m_{kategorik}$. Kemudian data numerik dan data kategorik tersebut dipisah dan dikelompokkan sesuai jenis datanya masing-masing. Hasil kelompok tersebut kemudian akan digabungkan menggunakan metode *cluster ensemble* ROCK yang akan diperoleh kelompok akhir (*final cluster).*

Cluster ensemble diperkenalkan pada tahun 2002 dan dikembangkan oleh Strehl dan Gosh. *Cluster ensemble* merupakan sebuah metode yang digunakan dalam analisis *cluster* untuk mengkombinasikan sekumpulan solusi *cluster* (Suhaeni, dkk., 2018). Keunggulan dari metode ini adalah dapat meningkatkan kualitas dan kekuatan pada solusi *cluster.* Para ahli mengembangkan metode ini dengan motivasi untuk mendapatkan hasil dari solusi *cluster* dari metode yang berbeda. Pengelompokkan pada *cluster ensemble* dilakukan dengan menggabungkan beberapa solusi dari metode pengelompokkan hingga didapatkan satu hasil *cluster* yang lebih baik. Input yang dibutuhkan adalah solusi pengelompokkan yang telah diperoleh dari beberapa hasil pengelompokkan tanpa melihat data awal

Cluster ensemble dapat dilakukan dengan dua tahap algoritma. Tahap pertama yaitu membentuk anggota *ensemble* yang anggotanya merupakan solusi dari berbagai metode pengelompokkan yang berbeda. Tahap kedua mengombinasikan seluruh anggota *ensemble* untuk memperoleh satu solusi akhir yang disebut fungsi *consensus.*

Pengelompokkan *ensemble cluster* dengan data campuran dapat menggunakan algoritma CEBMDC (*Cluster Ensemble Based Mixed Data Clustering*) yang ditunjukkan pada gambar berikut:



Gambar 1. Tahapan Analisis dengan Algoritma CEBMDC

Berdasarkan Gambar 1 maka tahapan pengelompokkan menggunakan metode *ensemble* dengan algoritma CEBMDC adalah sebagai berikut:

a. Membagi data menjadi subdata, yaitu murni numerik dan murni kategorik
b. Melakukan pengelompokkan objek yang memiliki variabel numerik dengan algoritma pengelompokkan data numerik, serta melakukan pengelompokkan objek yang memiliki variabel kategorik dengan algoritma pengelompokkan data kategorik.
c. Menggabungkan (*combining*) hasil pengelompokkan dari variabel numerik dan kategorik, yang disebut proses *ensemble.*
d. Melakukan pengelompokkan *ensemble* menggunakan algoritma pengelompokkan data kategorik untuk mendapatkan kelompok akhir (*final cluster*).

III. METODELOGI PENELITIAN

3.1 Waktu dan Tempat Penelitian

Penelitian ini dilakukan pada Semester Ganjil tahun akademik 2022/2023 dengan melakukan penelitian secara studi pustaka di Jurusan Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Lampung.

3.2 Data Penelitian

Data yang digunakan dalam penelitian ini adalah data sekunder yang diperoleh dari *website* resmi Kementerian Pendidikan, Kebudayaan, Riset dan Teknologi (Kemendikbudristek). Kota Bandar Lampung terdiri dari 67 Sekolah Menengah Atas (SMA) Negeri ataupun Swasta, namun dalam penelitian ini hanya digunakan 60 SMA dikarenakan terdapat *missing value*. Data yang digunakan adalah data numerik dan data kategorik. Untuk data numerik yaitu rasio siswa/rombel (X_1), jumlah guru tetap (X_2), jumlah tenaga pendidik (X_3), jumlah ruang kelas (X), jumlah laboratorium (X_5), luas lahan (X_6) dan daya listrik (X_7) sedangkan untuk data kategorik yaitu data akreditasi SMA (X_8), status SMA (X_9), sumber listrik SMA (X_{10}) dan waktu penyelenggaraan SMA (X_{11}). Data kategorik harus diubah terlebih dahulu ke dalam bentuk pengukuran skala nominal. Berikut adalah keterangan dari variabel data kategorik dengan pengukuran skala nominal yang digunakan.

Tabel 4. Indikator Variabel Data Kategorik

Variabel		Kategorisasi
x_8	Akreditasi SMA	1; A 2; B 3; C
x_9	Status SMA	1; Negeri 2; Swasta
x_{10}	Sumber Listrik SMA	1; PLN 2; PLN & Diesel
x_{11}	Waktu Penyelenggaraan SMA	1; Pagi 2; Siang 3; Sehari Penuh (5h/m) 4; Sehari Penuh (6h/m)

3.3 Metode Penelitian

Penelitian ini dilakukan secara studi pustaka dengan mengumpulkan serta mempelajari buku-buku teks, jurnal serta akses internet. Untuk mempermudah perhitungan maka dalam penelitian ini menggunakan program *R.* Adapun tahapan penelitian yang dapat dilakukan adalah sebagai berikut:

1. Mengumpulkan serta memasukkan data kedalam program *R.*
2. Melakukan pemisahan data campuran menjadi dua yaitu data numerik dan data kategorik.
3. Melakukan analisis deskriptif berdasarkan tipe data masing-masing yaitu data numerik dan data kategorik. Analisis deskriptif pada data numerik berdasarkan nilai maksimum, minimum, *mean*, median, standar deviation serta mendeteksi pencilan dengan boxplot dari setiap variabel tersebut sedangkan pada data kategorik dengan diagram lingkaran.
4. Melakukan normalisasi data menggunakan normalisasi *Z-score.*
5. Melakukan pengelompokkan data numerik dengan menggunakan *K-Medoids.* Tahapan *clustering* dengan algoritma *K-Medoids* sebagai berikut:

a) Inisialisasi pusat *cluster* sebanyak *k* (jumlah *cluster*)
b) Hitung setiap objek ke *cluster* terdekat menggunakan persamaan ukuran jarak Euclidean.
c) Setelah menghitung jarak Euclidean, inisialisasikan pusat *cluster* baru secara acak pada masing-masing objek sebagai kandidat *non medoids.*
d) Hitung jarak setiap objek yang berada pada masing masing *cluster* dengan kandidat *non medoids.*
e) Hitung total simpangan (S). Jika S < 0 maka tukar objek dengan data *cluster non medoids* untuk membentuk sekumpulan k objek baru sebagai *medoids*
f) Ulangi Langkah (c)-(e) hingga tidak terjadi perubahan pada *medoid*, sehingga didapatkan *cluster* beserta anggota masing-masing *cluster.*

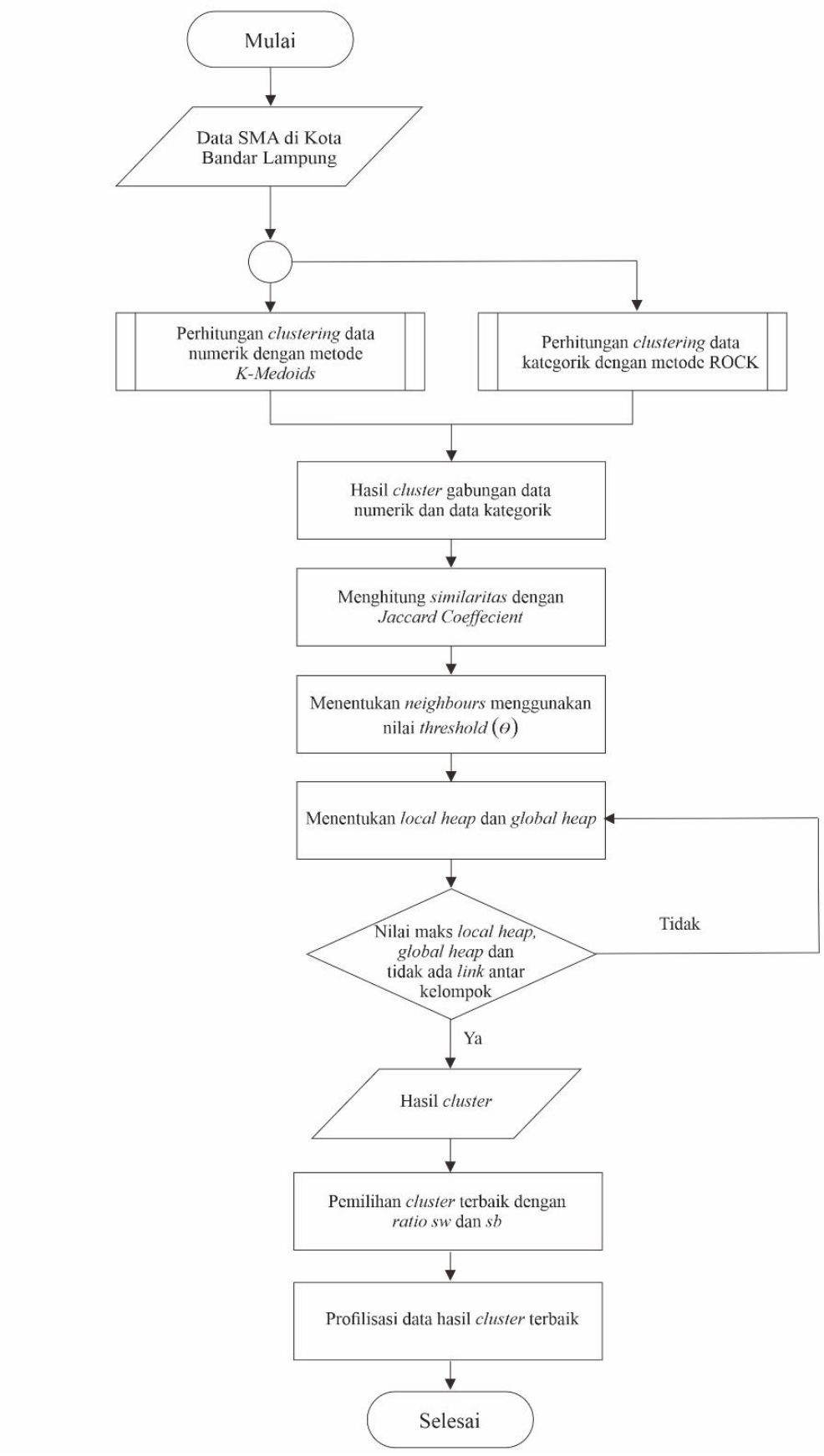
6. Melakukan validasi *cluster* pengelompokkan optimum dengan menghitung *sillhoutte coeffecient.* Semakin nilai . *sillhoutte coeffecient* mendekati nilai 1, maka semakin baik pengelompokkan data dalam satu *cluster.* Sebaliknya jika nilai *sillhoutte index* mendekati nilai -1, maka semakin buruk pengelompokkan data di dalam satu *cluster.*
7. Membandingkan hasil *sillhoutte coeffecient* pada masing-masing *cluster* yang terbentuk untuk menentukan jumlah *cluster* yang optimum.
8. Denormalisasi data hasil pengelompokkan
9. Melakukan profilisasi data hasil pengelompokkan dengan metode *K-Medoids*
10. Melakukan pengelompokkan data kategorik dengan menggunakan *Robust Clustering Using Links* (ROCK). Tahapan *clustering* dengan algoritma ROCK sebagai berikut:
    a) Menghitung *similaritas* menggunakan rumus *Jaccard coefficient*
    b) Menentukan Tetangga (*Neighbors*) menggunakan nilai *threshold* (θ) yang digunakan biasanya berkisar antara 0 sampai 1 yang disesuaikan dengan keadaan data.
    c) Menghitung *link* (X_i, X_j) antar objek diperoleh dari jumlah tetangga yang sama (*common neighbor)* antara X_i dan X_j.
    d) Menentukan *local heap* dan *global heap. Local heap* adalah nilai ukuran kebaikan (*goodness measure)* untuk setiap kelompok dengan kelompok

lainnya jika $link \neq 0$. Menentukan *global heap. Global heap* adalah nilai maksimum *goodness measure* antar kolom di baris ke-*i.*

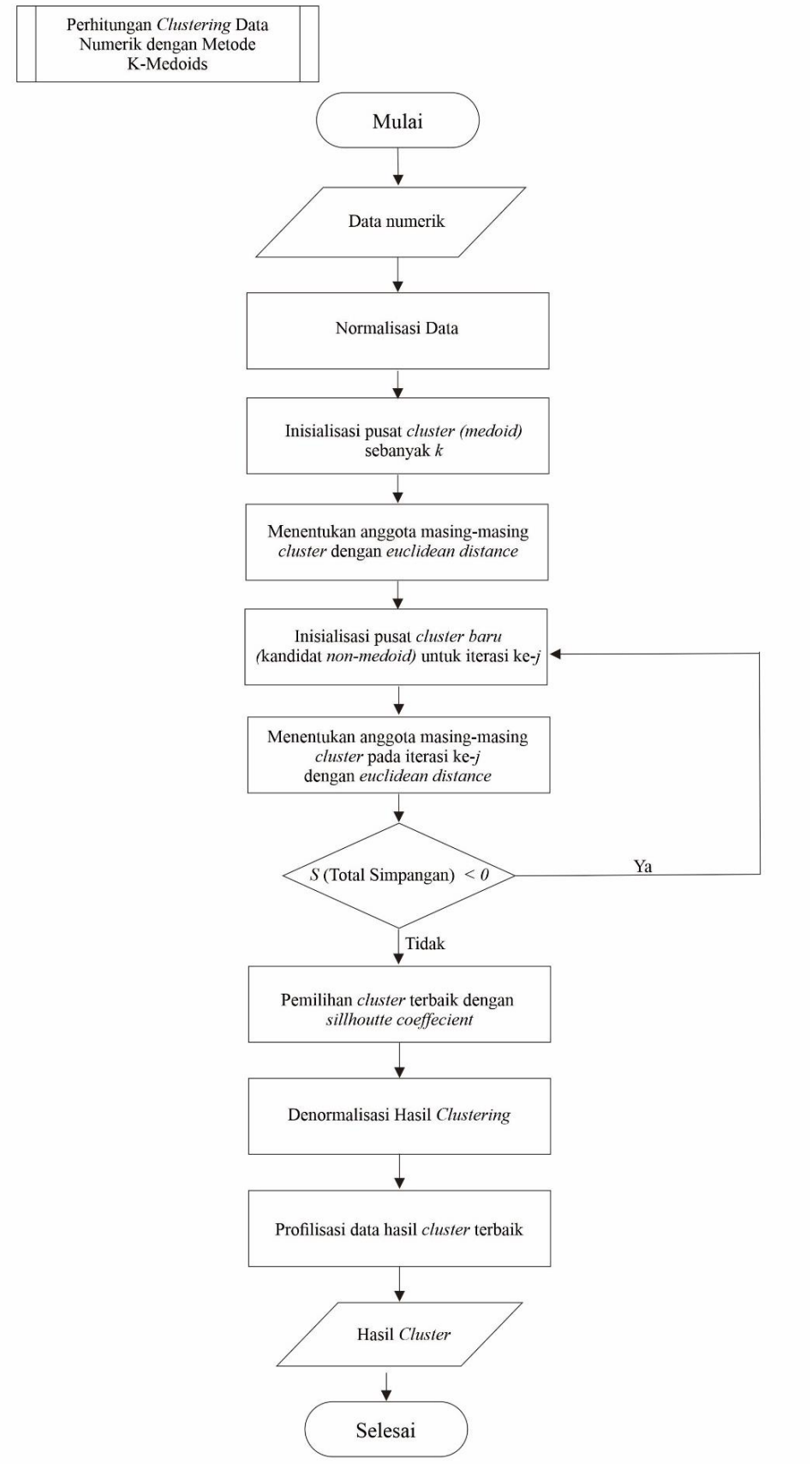
e) Ulangi Langkah (d) hingga mendapatkan nilai maksimum di *global heap* dan *local heap.*

f) Hentikan melakukan pengelompokkan metode ROCK jika jumlah dari *cluster* yang diharapkan sudah terpenuhi dan tidak ada *link* antar kelompok.

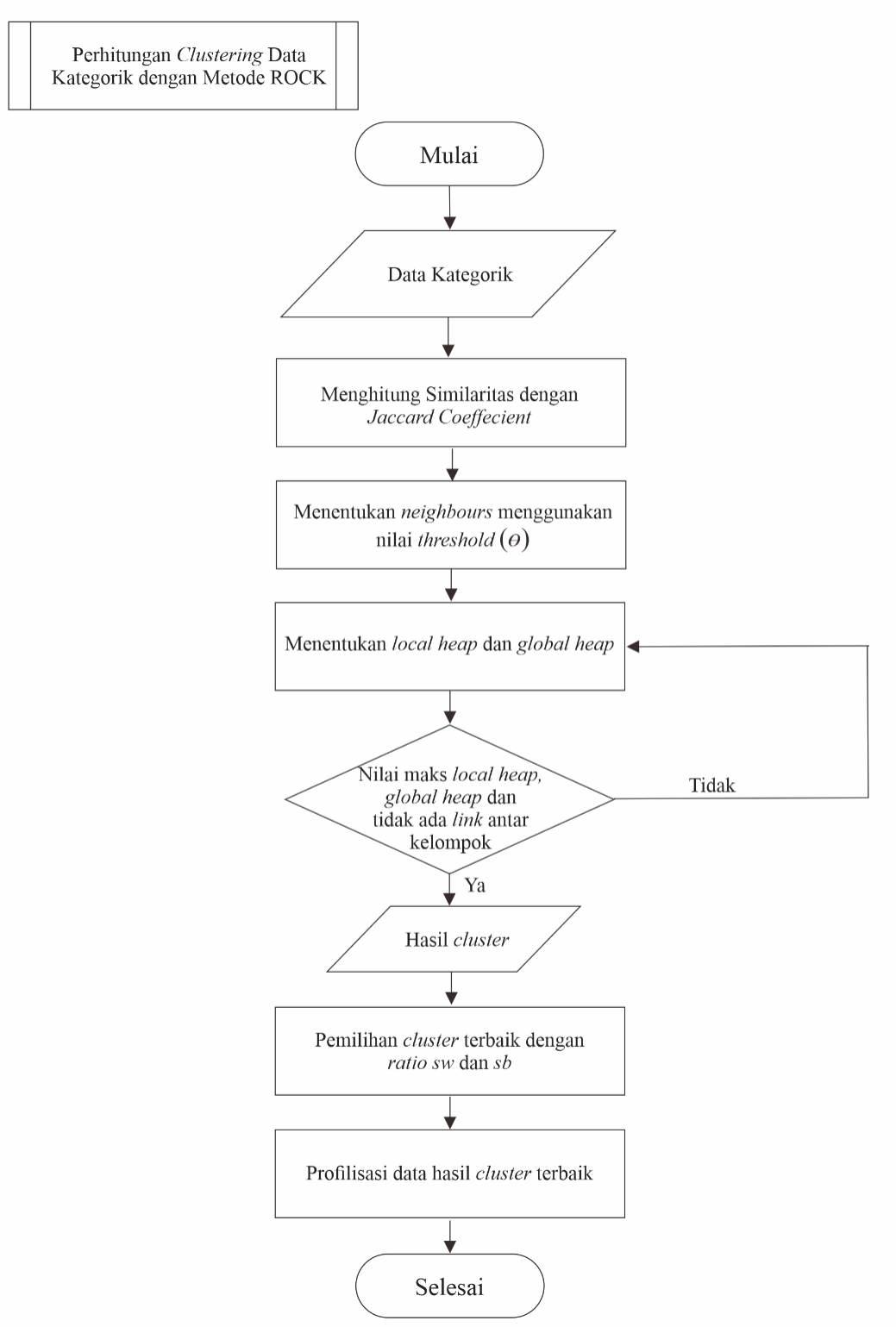
11. Melakukan validasi *cluster* pengelompokkan optimum dengan menghitung nilai ratio s_w dan s_b. Semakin kecil nilai rasio s_w dan s_b maka semakin baik pula kinerja pengelompokkan dari hasil *cluster* yang didapatkan sehingga memiliki arti bahwa terdapat homogenitas maksimum dalam kelompok dan heterogenitas maksimum antar kelompok.
12. Membandingkan hasil ratio s_w dan s_b untuk masing-masing nilai θ dan menentukan jumlah *cluster* yang optimum berdasarkan kriteria ratio s_w dan s_b terkecil.
13. Menggabungkan hasil pengelompokkan data numerik dan data kategorik yang terbaik dengan metode *cluster ensemble.*
14. Melakukan pengelompokkan kembali dengan menggunakan metode *cluster ensemble* pendekatan metode ROCK.
15. Melakukan validasi *cluster* untuk mendapatkan hasil akhir pengelompokkan optimum dengan membandingkan hasil ratio s_w dan s_b untuk mendapatkan jumlah *cluster* yang optimum berdasarkan kriteria ratio s_w dan s_b terkecil.
16. Melakukan analisis karakteristik dari masing-masing hasil *cluster* yang terbentuk dengan metode *ensemble* ROCK berdasarkan sarana, prasarana dan SDM SMA.



Gambar 2. Diagram Alir *Ensemble* ROCK *Clustering*



Gambar 3. Diagram Alir *Clustering* Data Numerik dengan *K-Medoids*



Gambar 4. Diagram Alir *Clustering* Data Kategorik dengan ROCK

V. KESIMPULAN

Berdasarkan hasil analisis dan pembahasan diperoleh kesimpulan sebagai berikut:

1. Analisis *cluster* metode *ensemble Robust Clustering Using Links* (ROCK) dengan algoritma *Cluster Ensemble Based Mixed Data Clustering* (CEBMDC) merupakan metode yang dapat digunakan untuk jenis data campuran.
2. Pengelompokkan Sekolah Menengah Atas (SMA) di Bandar Lampung menghasilkan 2 *cluster* berdasarkan nilai rasio $s_w$dan s_b terkecil sebesar 0.0310 dengan nilai θ sebesar 0.05. Pengelompokkan ini menggunakan gabungan dari pengelompokkan data numerik metode *K-Medoids* yang menghasilkan 2 *cluster* dengan nilai *silhouette coeffecient* sebesar 0.4132 dan pengelompokkan data kategorik metode ROCK menghasilkan 3 *cluster* dengan nilai θ sebesar 0.20 dan nilai *ratio* $s_w$dan s_b terkecil sebesar 0.0444. Dari hasil pengelompokkan dengan metode *ensemble* ROCK diperoleh bahwa *Cluster*-1 yang beranggotakan 25 SMA dari 60 SMA memiliki kualitas sarana dan prasarana serta SDM SMA yang lebih baik daripada *Cluster*-2 yang beranggotakan 35 SMA dari 60 SMA.

DAFTAR PUSTAKA

Ambiani, R. 2021. Penerapan Analisis Cluster Ensemble dengan Metode ROCK Pada Data Indeks Kesejahteraan Masyarakat di Indonesia. Jurusan Matematika FMIPA UNILA, Lampung.

Alvionita, Sutikno, & Suharsono A. 2017. Ensemble ROCK Methods and Ensemble Methods for Clustering of Cross Citrus Accessions Based on Mixed Numerical and Categorical Dataset. *IOP Conference Series: Earth and Environmental Science*. **58**(1): 1-10.

Bhat, A. 2014. K-Medoids Clustering Using Partitioning Around Medoids For Performing Face Recognition. *International Journal of Soft Computing, Mathematics and Control*. **3**(3): 1-12.

Chatfield. C & Collins. A. J. 2018. *Introduction to Multivariate Analysis*. 1th Edition. Routledge, New York.

Chrisinta, D., Sumertajaya, I. M., & Indahwati, I. 2020. Evaluasi Kinerja Metode Cluster Ensemble Dan Latent Class Clustering Pada Peubah Campuran. *Indonesian Journal of Statistics and Its Applications*. **4**(3): 448-461.

Defiyanti, S., Jajuli, M., & Rohmawati, N. 2017. Optimalisasi K-Medoid Dalam Pengklasteran Mahasiswa Pelamar Beasiswa Dengan Cubic Clustering Criterion. *Jurnal Nasional Teknologi dan Sistem Informasi*. **3**(1): 211-218.

Guha, S., Rastologi, R., & Shim, K. 2000. ROCK: A Robust Clustering Algorithm For Categorical Attributes. Proceeding Of The 15th International Conference on Data Engineering.

Goreti, M., Nasution, Y. N., & Wahyuningsih, S. 2017. Perbandingan Hasil Analisis Cluster dengan Menggunakan Metode Single Linkage dan Metode C-Means. *Eksponensial*. **7**(1): 9-16.

Hair, J. F., Black, W. C., Babin, J. B., dan Anderson, E. R. 2009. Multivariate Data Analysis. 7th Edition. Prentice Hall Inc, New Jersey.

Hardle, W. K., & Simar, L. 2019. *Applied multivariate statistical analysis*. 5th Edition. Springer Nature, Switzerland.

Johnson, R. A., & Wichern, D. W. 2015. Applied multivariate statistical analysis. *Statistics*. **6215**(10): 10.

Kaufman, L., & Rousseeuw, P. J. 2009. *Finding Groups in Data: an Introduction to Cluster Analysis.* John Wiley & Sons, New York.

Nasution, D. A., Khotimah, H. H., & Chamidah, N. 2019. Perbandingan Normalisasi Data untuk Klasifikasi Wine Menggunakan Algoritma K-NN. *Journal of Computer Engineering, System and Science*. **4**(1): 78-82.

Nishom, M. 2019. Perbandingan Akurasi Euclidean Distance, Minkowski Distance, dan Manhattan Distance pada Algoritma K-Means Clustering berbasis Chi-Square. *Jurnal Informatika*. **4**(1): 20-24.

Pramesti, D. F., Furqon, M. T., & Dewi, C. 2017. Implementasi Metode K-Medoids Clustering Untuk Pengelompokan Data Potensi Kebakaran Hutan/Lahan Berdasarkan Persebaran Titik Panas (Hotspot). *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*. **1**(9): 723-732.

Reddy, M. J., & Kavitha, B. 2012. Clustering the mixed numerical and categorical dataset using similarity weight and filter method. *International Journal of Database Theory and Application.* **5**(1): 121-134.

Religia, Y., Intani, A. E., & Saputra, A. 2020. Pengelompokan Menggunakan Algoritma K-Medoid Untuk Evaluasi Performa Siswa. *Pelita Teknologi*. **15**(1): 49-55.

Rumiati, A. T., Salsabila, N. Z., Sari, H. J., & Riza, L. F. 2022. Mapping model for target achievement of village SDGs using the Ensemble ROCK method: A case study of Sidoarjo Regency, East Java. *International Journal of Research in Business and Social Science.* **11**(1), 316-327.

Safrullah, S., Wibawa, G. N. A., & Abapihi, B. 2020. Penerapan Analisis Cluster Ensemble dengan Metode Rock untuk Mengelompokkan Data Berskala Campuran. Prosiding Seminar Nasional Statistika| Departemen Statistika FMIPA. Universitas Padjadjaran.

Sharma, S., & Yadav, R. L. 2013. Comparative study of K-means and robust clustering. *International Journal Advanced Computer Research*. **3**(3): 207-210.

Sihombing, R. E., Rachmatin, D., & Dahlan, J. A. 2019. Program Aplikasi Bahasa R Untuk Pengelompokan Objek Menggunakan Metode K-Medoids Clustering. *Jurnal EurekaMatika*. **7**(1): 58-79.

Sindi, S., Ningse, W. R. O., Sihombing, I. A., Zer, F. I. R., & Hartama, D. 2020. Analisis Algoritma K-Medoids Clustering dalam Pengelompokkan Penyebaran Covid-19 di Indonesia. *(JurTI) Jurnal Teknologi Informasi*. **4**(1): 166-173.

Sitepu, R., Irmeilyana, I., & Gultom, B. 2011. Analisis Cluster Terhadap Tingkat Pencemaran Udara Pada Sektor Industri Di Sumatera Selatan. *Jurnal penelitian sains*. **14**(3): 11-17.

Suhaeni, C., Kurnia, A., & Ristiyanti, R. 2018. Perbandingan Hasil Pengelompokan menggunakan Analisis Cluster Berhirarki, K-Means Cluster, dan Cluster Ensemble (Studi Kasus Data Indikator Pelayanan Kesehatan Ibu Hamil). *Jurnal Media Infotama*. **14**(1): 31-38.

Widyadhan, D., Hastuti, R. B., Kharisudin, I., & Fauzi, F. 2021. Perbandingan Analisis Klaster K-Means Dan Average Linkage Untuk Pengklasteran Kemiskinan Di Provinsi Jawa Tengah. *PRISMA, Prosiding Seminar Nasional Matematika*. **4(1)**: 584-594.

Wulandari, L., Farida, Y., Fanani, A., Ulinnuha, N., & Intan, P. K. (2020). Evaluation of Disadvantaged Regions in East Java Based-on the 33 Indicators of the Ministry of Villages, Development of Disadvantaged

Regions, and Transmigration Using the Ensemble ROCK (Robust Clustering Using Link) Method. *Technology and Engineering Systems Journal.* **5**(5): 193-200.