

**PENERAPAN ANALISIS *CLUSTER ENSEMBLE*
DENGAN METODE ROCK PADA DATA INDEKS KESEJAHTERAAN
MASYARAKAT DI INDONESIA**

(Skripsi)

Oleh

RESTU AMBIANI

1757031002



**JURUSAN MATEMATIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMPUNG
BANDAR LAMPUNG
2021**

ABSTRACT

APPLICATION OF ENSEMBLE CLUSTERING ANALYSIS WITH ROCK METHOD ON COMMUNITY WELFARE INDEX DATA IN INDONESIA

By

RESTU AMBIANI

Cluster analysis is a multivariate technique with interdependence methods that generally focus on one data scale. In certain cases, were found 2 data scales in the study with cluster analysis. Transformation method is considered inappropriate for carrying out the analysis, so cluster ensemble analysis was developed to solve this problem. This study uses data on indicators of community welfare in Indonesia. This research used ensemble ROCK method. Numerical data analysis using agglomerative hierarchical method. and categorical analysis using ROCK method. The results of numerical data clustering analysis were obtained by 4 cluster consisting of 30 provinces in the first cluster, 1 province in the second cluster, 2 provinces in the third cluster, and 1 province in the fourth cluster. The results of categorical data clustering analysis were obtained by 3 cluster consisting of 10 provinces in the first cluster, 10 provinces in the second cluster, and 14 provinces in the third cluster. Cluster ensemble of mixed data generate 2 cluster with the smallest ratio is 4.2×10^{-2} . With 19 provinces consist in first cluster and 15 provinces in the second cluster.

Keywords: Cluster Analysis, Indicators of Community Welfare, Agglomerative Hierarchical, Ensemble ROCK.

ABSTRAK

PENERAPAN ANALISIS *CLUSTER ENSEMBLE* DENGAN METODE ROCK PADA DATA INDEKS KESEJAHTERAAN MASYARAKAT DI INDONESIA

Oleh

RESTU AMBIANI

Analisis *cluster* merupakan salah satu teknik multivariat metode interdependen yang umumnya terfokus pada satu skala data. Pada kasus tertentu ditemukan 2 skala data dalam penelitian menggunakan analisis *cluster*. Metode transformasi dianggap tidak cukup tepat untuk melakukan analisis, maka dikembangkan metode analisis *cluster ensemble* untuk menyelesaikan permasalahan ini. Penelitian ini menggunakan data indikator kesejahteraan rakyat di Indonesia. Metode yang digunakan dalam penelitian ini adalah metode *ensemble* ROCK. Analisis data numerik menggunakan metode hirarki *agglomerative* dan analisis data kategorik menggunakan metode ROCK. Hasil analisis pengelompokan data numerik diperoleh sebanyak 4 kelompok terdiri dari 30 provinsi dalam kelompok 1, 1 provinsi kelompok 2, 2 provinsi kelompok 3, dan 1 provinsi kelompok 4. Hasil analisis data kategorik diperoleh sebanyak 3 kelompok terdiri dari 10 provinsi dalam kelompok 1, 10 provinsi dalam kelompok 2, dan 14 provinsi dalam kelompok 3. Pengelompokan data campuran menghasilkan 2 kelompok dengan nilai rasio terkecil yaitu 4.2×10^{-2} . Dengan 19 provinsi dalam kelompok 1 dan 15 provinsi dalam kelompok 2.

Kata Kunci: Analisis Klaster, Indikator Kesejahteraan Rakyat, Hirarki *Agglomerative*, *Ensemble* ROCK.

**PENERAPAN ANALISIS *CLUSTER ENSEMBLE*
DENGAN METODE ROCK PADA DATA INDEKS KESEJAHTERAAN
MASYARAKAT DI INDONESIA**

Oleh

RESTU AMBIANI

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
2021**

Judul Skripsi : **PENERAPAN ANALISIS *CLUSTER ENSEMBLE* DENGAN METODE ROCK PADA DATA INDEKS KESEJAHTERAAN MASYARAKAT DI INDONESIA**

Nama Mahasiswa : **Restu Ambiani**

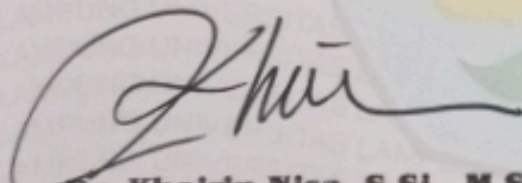
Nomor Pokok Mahasiswa : 1757031002

Jurusan : Matematika

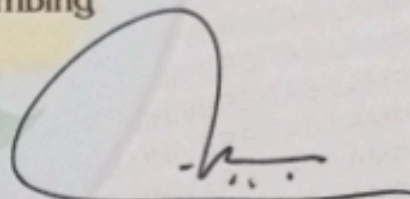
Fakultas : Matematika dan Ilmu Pengetahuan Alam

MENYETUJUI

1. Komisi Pembimbing

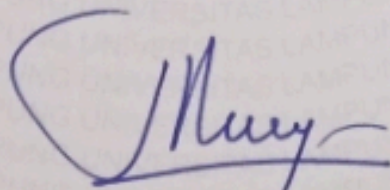


Dr. Khoirin Nisa, S.Si., M.Si.
NIP 19740726 200003 2 001



Sublan Saidi, S.Si., M.Si.
NIP 19800821 200812 1 001

2. Ketua Jurusan Matematika

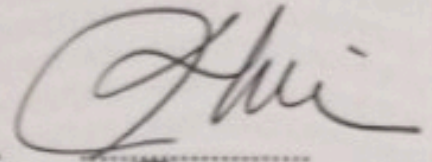


Dr. Aang Nuryaman, S.Si., M.Si.
NIP 19740316 200501 1 001

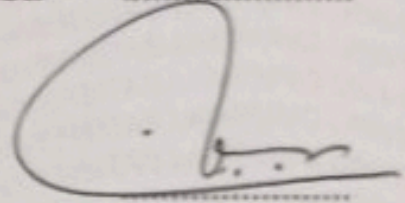
MENGESAHKAN

1. Tim Penguji

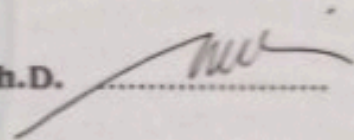
Ketua : Dr. Khoirin Nisa, S.Si., M.Si.



Sekretaris : Sublan Saidi, S.Si., M.Si.



Penguji
Bukan Pembimbing : Ir. Netti Herawati, M.Sc., Ph.D.



2. Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam



Dr. Eng. Satripto Dwi Yuwono, M.T.
NIP. 19746705 200003 1 001

Tanggal Lulus Ujian Skripsi : 15 Desember 2021

PERNYATAAN SKRIPSI MAHASISWA

Saya yang bertanda tangan dibawah ini:

Nama : Restu Ambiani

Nomor Pokok Mahasiswa : 1757031002

Judul : **PENERAPAN ANALISIS *CLUSTER*
ENSEMBLE DENGAN METODE ROCK
PADA DATA INDEKS
KESEJAHTERAAN MASYARAKAT DI
INDONESIA**

Jurusan : **Matematika**

Dengan ini menyatakan bahwa skripsi ini adalah hasil pekerjaan saya sendiri, bukan hasil orang lain, dan semua hasil tulisan yang tertuang dalam skripsi ini telah mengikuti kaidah penulisan karya ilmiah Universitas Lampung. Apabila kemudian hari terbukti bahwa skripsi ini merupakan hasil salinan atau dibuat oleh orang lain, maka saya bersedia menerima sanksi sesuai dengan ketentuan akademik yang berlaku.

Bandar Lampung, 15 Desember 2021
Penulis,



Restu Ambiani
NPM. 1757031002

RIWAYAT HIDUP

Penulis bernama lengkap Restu Ambiani, lahir di Cilegon pada tanggal 3 Oktober 1999. Penulis merupakan anak ketiga dari empat bersaudara yang lahir dari pasangan Bapak Tibyani dan Ibu Tatu Ukhro.

Penulis menempuh pendidikan sekolah dasar di SD Negeri Samang Raya I Kecamatan Citangkil Kota Cilegon pada tahun 2005-2011, kemudian melanjutkan ke jenjang sekolah menengah pertama di SMP Negeri 4 Cilegon Kecamatan Ciwandan Kota Cilegon pada tahun 2011-2014, dan jenjang sekolah menengah atas di SMA Islam Terpadu Raudhatul Jannah Kota Cilegon pada tahun 2014-2017. Pada tahun 2017, Penulis diterima sebagai mahasiswa S1 di Jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung melalui jalur Seleksi Mandiri Masuk Perguruan Tinggi Negeri (SMMPTN).

Selama menempuh pendidikan di Universitas Lampung penulis aktif di Organisasi tingkat Fakultas dengan bidang Sosial dan Pengabdian Masyarakat Badan Eksekutif Mahasiswa (BEM) FMIPA Unila pada tahun 2018-2020. Pada tahun 2020, Penulis juga aktif di organisasi luar dalam bidang Sosial dan Pengabdian di Himpunan Mahasiswa Banten Lampung, penulis melaksanakan Kuliah Kerja Nyata (KKN) di Tri Tunggal Jaya, Kecamatan Banjar Margo, Kabupaten Tulang Bawang, sebagai bentuk pengabdian kepada masyarakat, dan pada tahun yang sama Penulis Praktik Kerja Lapangan (PKL) di Kantor Badan Pusat Statistik Kota Bandar Lampung, sebagai bentuk aplikasi bidang ilmu di dunia kerja.

KATA INSPIRASI

*Successful people don't fear failure but understand
that it's necessary to learn and grow from"*

(Robert Kiyosaki)

"If Opportunity doesn't knock, build a door."

(Milton Berle)

*"Sesungguhnya sesudah kesulitan itu ada kemudahan. Maka apabila kamu telah
selesai (dari suatu urusan), kerjakanlah dengan sungguh-sungguh (urusan yang
lainnya). Dan hanya kepada Tuhanmu lah hendaknya kamu berharap."*

(QS. Al-Insyirah : 6-8)

"If you can dream it, you can do it."

(Walt Disney)

"Life has no remote, get up and change it yourself."

(Mark A Cooper)

PERSEMBAHAN

Alhamdulillahirobbil'alamin,

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

Diri Sendiri

Terima kasih untuk tetap bertahan dan tidak menyerah dengan keadaan serta terus berusaha untuk mengembangkan diri dan menjadi pribadi yang lebih baik.

Mama, Ayah, Kakak, dan Adik

Terima kasih selalu memberikan doa dan dukungan terus menerus kepadaku disaat mulai ingin menyerah. Terima kasih yang sebesar-besarnya atas semua kesabaran, kasih sayang, waktu, dan pengorbanan yang kalian berikan kepadaku.

Dosen

Terima kasih kepada dosen-dosen Pembimbing dan Pembahas yang telah sangat sabar dalam membimbing dan memberikan masukan ide ide yang membangun sehingga aku dapat menyelesaikan skripsi ini. Terima kasih atas ilmu dan pengetahuan yang telah kalian berikan.

Sahabat - sahabatku

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

Terima kasih telah memberikan warna dan kenangan indah dalam hidupku

SANWACANA

Puji dan syukur penulis panjatkan kehadiran Allah SWT, yang telah memberikan rahmat dan hidayah-Nya kepada penulis, sehingga penulis dapat menyelesaikan skripsi yang berjudul “Penerapan Analisis *Cluster Ensemble* Dengan Metode ROCK Pada Data Indeks Kesejahteraan Masyarakat Di Indonesia”.

Penulis menyadari bahwa skripsi ini tidak akan terwujud tanpa adanya bantuan, bimbingan, dan doa dari berbagai pihak sehingga skripsi ini dapat terselesaikan. Dengan segala kerendahan hati penulis mengucapkan terima kasih kepada:

1. Ibu Dr. Khoirin Nisa, S.Si., M.Si., selaku dosen Pembimbing I yang selalu memberikan motivasi, bimbingan, pengarahan, dan saran kepada penulis dalam menyelesaikan skripsi serta telah banyak membimbing selama proses perkuliahan.
2. Bapak Subian Saidi, S.Si., M.Si., selaku Pembimbing II, terimakasih atas kesediaan memberikan bimbingan, saran, dan kritik dalam proses penyelesaian skripsi ini.
3. Ibu Ir. Netti Herawati, M.Sc., Ph.D., selaku Pembahas skripsi yang telah memberikan evaluasi dan saran perbaikan untuk skripsi penulis.
4. Ibu Dr. Khoirin Nisa, S.Si., M.Si., selaku pembimbing akademik, terimakasih atas bimbingan dan pembelajarannya selama ini.
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. Suropto Dwi Yuwono, M.T., selaku Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung.
7. Seluruh dosen, staf dan karyawan Jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung.

8. Ayah, mama, kakak, adik dan seluruh keluarga besar yang selalu memberi dukungan, motivasi, semangat, dan perhatian kepada penulis sehingga dapat menyelesaikan skripsi ini.
9. Sahabat-sahabat seperjuangan: Suci Anugrah, Indah Oktasari, Anindiya Handayani, Sayyidah, Amalia Puspita, Bella Pratiwi, Ajeng Ayu, Alexandra Florecita, Azzakiah Nupus, Indah Suciati, dan Epmi Annisa yang selalu memberikan semangat serta dukungan kepada penulis.
10. Teman-temanku Himpunan Mahasiswa Banten 2017, terima kasih telah memberikan keceriaan di tanah rantauan ini.
11. Teman-teman “Lambetika C” terima kasih atas kebersamaan dan keceriaan selama masa perkuliahan.
12. Teman-teman Matematika 2017, terima kasih atas kebersamaannya.
13. Seluruh pihak yang telah membantu dan terlibat dalam menyelesaikan skripsi ini.

Penulis menyadari bahwa dalam penulisan skripsi ini masih jauh dari kesempurnaan, oleh karena itu kritik dan saran sangat penulis harapkan. Semoga skripsi ini dapat bermanfaat bagi penulis dan bagi para pembaca.

Bandar Lampung, 15 Desember 2021
Penulis

Restu Ambiani

DAFTAR ISI

	Halaman
DAFTAR TABEL	vii
DAFTAR GAMBAR	viii
I. PENDAHULUAN	
1.1. Latar Belakang dan Masalah	1
1.2. Tujuan Penelitian	2
1.3. Manfaat Penelitian	2
II. TINJAUAN PUSTAKA	
2.1. Data	3
2.2. Analisis <i>Cluster</i>	3
2.3. Metode Pengelompokan	7
2.3.1. <i>Clustering</i> Data Kategorik	7
2.3.2. <i>Clustering</i> Data Numerik	9
2.3.2.1. Metode <i>Single Linkage</i>	11
2.3.2.2. Metode <i>Complete Linkage</i>	11
2.3.2.3. Metode <i>Average Linkage</i>	12
2.3.3. <i>Clustering</i> Data Campuran	14
III. METODOLOGI PENELITIAN	
3.1. Waktu dan Tempat Penelitian	17
3.2. Data Penelitian	17
3.3. Metode Penelitian	17
IV. HASIL DAN PEMBAHASAN	
4.1. Data Indikator Kesejahteraan Rakyat	19
4.1.1. Data Numerik	19
4.1.2. Data Kategorik	21
4.2. Pengelompokan dengan Metode <i>Ensemble ROCK</i>	23
4.2.1. Pengelompokan Data Numerik	23
4.2.2. Pengelompokan Data Kategorik	27
4.2.3. Pengelompokan Data Campuran	30

V. SIMPULAN DAN SARAN

5.1. Simpulan	35
5.2. Saran	35

DAFTAR PUSTAKA	37
-----------------------------	-----------

LAMPIRAN	39
-----------------------	-----------

DAFTAR TABEL

Tabel	Halaman
1. Kontingensi Jarak Data Biner	5
2. Ukuran Jarak Data Biner	5
3. Statistika Deskriptif Variabel Numerik	19
4. Nilai R^2	24
5. Nilai Rasio S_W dan S_B	25
6. Hasil <i>Clustering</i> Numerik	26
7. Nilai Rasio Metode ROCK	29
8. Hasil <i>Clustering</i> Kategorik	29
9. Hasil <i>Clustering</i> Numerik dan Kategorik	30
10. Rasio <i>Ensemble</i> ROCK	31
11. Hasil ROCK <i>Ensemble Clustering</i>	32

DAFTAR GAMBAR

Gambar	Halaman
1. <i>Flow Chart</i> Pengelompokkan <i>Ensemble</i>	15
2. <i>Flow Chart</i> Algoritma CEBMDC	15
3. Persentase Kepadatan Penduduk	21
4. Persentase Penduduk Miskin	22
5. Plot R^2 Numerik	24
6. Dendogram Pengelompokkan Numerik	26
7. Plot Rasio Metode ROCK	28
8. Plot Rasio <i>Ensemble</i> ROCK	31
9. Kepadatan Penduduk <i>Cluster 1</i>	32
10. Penduduk Miskin <i>Cluster 1</i>	33
11. Kepadatan Penduduk <i>Cluster 2</i>	33
12. Penduduk Miskin <i>Cluster 2</i>	34

I. PENDAHULUAN

1.1 Latar Belakang dan Masalah

Analisis *cluster* merupakan salah satu teknik dalam analisis multivariat metode interdependen yang tidak membedakan variabel bebas dan tak bebas dalam mengelompokkan suatu data. Mengelompokkan objek-objek pengamatan menjadi beberapa kelompok yang mengacu pada karakteristik yang dipertimbangkan untuk diteliti merupakan tujuan utama dari analisis *cluster* (Supranto, 2004). Dalam analisis *cluster*, objek yang berada dalam kelompok yang sama merupakan objek yang paling dekat kesamaannya, serta mempunyai kemiripan satu dengan yang lain (Johnson & Wichern, 2007).

Pengelompokkan data pada umumnya terfokus pada pengelompokan objek dengan variabel data berskala numerik atau kontinu, sedangkan untuk data kategorik kurang difokuskan dan mulai berkembang sekitar tahun 1990-an. Pada umumnya analisis *cluster* hanya dikembangkan untuk memproses salah satu data yaitu data kategorik atau data numerik saja. Permasalahan yang seringkali terjadi dalam analisis *cluster* adalah jenis data yang berskala campuran yaitu data kategorik dan numerik. Transformasi data merupakan suatu metode yang seringkali digunakan untuk mengelompokkan data berskala campuran dengan cara mentransformasi data kategorik menjadi numerik maupun data numerik menjadi data kategorik. Kelebihan metode transformasi adalah dapat mengurangi kompleksitas dalam komputasi, akan tetapi transformasi dinilai tidak cukup tepat untuk *clustering* data karena adanya beberapa informasi yang hilang ketika suatu data dilakukan transformasi (Dewangan, *et al.*, 2010).

Berdasarkan kelemahan pada *clustering* data menggunakan metode transformasi maka dikembangkan metode *cluster ensemble* untuk data berskala campuran. Menurut Zengyou, *et al.* (2002), *cluster ensemble* adalah suatu metode yang digunakan untuk menjalankan beberapa algoritma *clustering* yang berbeda guna mendapatkan bagian yang sama dari data yang bertujuan untuk menyatukan hasil dari hasil-hasil *clustering* individual.

Berdasarkan uraian diatas maka dalam penelitian ini akan dikaji mengenai penerapan analisis *cluster ensemble* dengan metode ROCK pada data indeks kesejahteraan masyarakat di Indonesia.

1.2 Tujuan Penelitian

Tujuan dari penelitian ini adalah mengetahui hasil *cluster* yang terbentuk dengan metode *ensemble ROCK* untuk data berskala campuran serta karakteristik dari hasil *cluster* yang terbentuk pada tingkat kesejahteraan provinsi di Indonesia.

1.3 Manfaat Penelitian

Manfaat penelitian ini adalah mengetahui dan memahami analisis *cluster* untuk data berskala campuran dengan menggunakan metode *ROCK* serta memberikan referensi bagi pembaca yang tertarik dalam analisis *cluster* dengan data campuran.

II. TINJAUAN PUSTAKA

2.1 Data

Data merupakan fakta yang dapat diolah menjadi suatu informasi. Menurut Anderson & Scolve (1974), data yang digunakan dalam analisis statistik berdasarkan jenis variabelnya dikelompokkan menjadi dua, yaitu data numerik dan data kategorik. Data numerik merupakan data yang variabelnya bersifat kuantitatif, data numerik merupakan data yang menghasilkan informasi numerik. Data numerik dapat dikelompokkan menjadi dua, yaitu data diskrit dan kontinu. Data kategorik merupakan data yang variabelnya bersifat kualitatif, data kategorik merupakan data dari hasil pengklasifikasian atau pengelompokkan dari suatu data.

2.2 Analisis *Cluster*

Analisis *cluster* merupakan suatu teknik analisis metode interdependen yang tidak membedakan variabel bebas dan tak bebas dalam analisis multivariat. Analisis *cluster* bertujuan untuk membuat *cluster* atau mengelompokkan data observasi ataupun variabel-variabel ke dalam *cluster* sedemikian rupa sehingga masing-masing *cluster* bersifat homogen sesuai dengan faktor yang digunakan untuk melakukan *clustering*. Untuk mendapatkan *cluster* yang sehomogen mungkin, maka yang digunakan sebagai dasar untuk melakukannya adalah kesamaan skor nilai yang dianalisis.

Menurut Gudono (2011), data mengenai ukuran kesamaan dapat dianalisis dengan analisis *cluster* sehingga dapat ditentukan siapa yang masuk *cluster* mana.

Analisis *cluster* merupakan suatu teknik analisis statistik yang ditujukan untuk menempatkan sekumpulan objek ke dalam dua atau lebih kelompok berdasarkan kesamaan-kesamaan objek atas dasar berbagai karakteristik (Simamora, 2005).

Menurut Han & Kamber (2001), analisis *cluster* adalah suatu teknik *data mining* untuk mengelompokkan himpunan objek (*dataset*) ke dalam beberapa *cluster* hanya berdasarkan kemiripan karakteristik dari atribut yang dimiliki oleh data objek sedemikian sehingga data objek yang berada di dalam *cluster* yang sama memiliki kemiripan satu sama lain tetapi tidak mirip dengan data objek yang berada dalam *cluster* yang berbeda.

Hasil analisis *cluster* dipengaruhi oleh objek yang dikelompokkan, variabel yang diamati, ukuran kemiripan atau ketidakmiripan yang digunakan, skala ukuran yang digunakan, serta metode *clustering* yang digunakan. Beberapa manfaat analisis *cluster* adalah eksplorasi data variabel ganda, reduksi data, dan prediksi keadaan objek. Ukuran kemiripan dan ketidakmiripan merupakan hal yang sangat mendasar dalam analisis *cluster*. Algoritma pengelompokkan menggunakan ukuran kemiripan dan ketidakmiripan digunakan untuk menggabungkan atau memisahkan objek dari suatu data. Ukuran kemiripan biasa digunakan untuk menganalisis data kategori, sedangkan ukuran ketidakmiripan untuk menganalisis data numerik.

Ukuran ketidakmiripan antar objek ke- i dengan objek ke- j (d_{ij}), merupakan fungsi yang memiliki sifat-sifat sebagai berikut $d_{ij} \geq 0$, $d_{ii} = 0$, $d_{ij} = d_{ji}$, dan $d_{ik} + d_{jk} \geq d_{ij}$, $\forall i, j$, dan k . Semakin besar nilai ukuran ketidakmiripan antara dua objek maka semakin besar perbedaan antar kedua objek tersebut, sehingga makin cenderung untuk tidak berada dalam kelompok yang sama (Jhonson & Wichern, 2007).

Ukuran kemiripan dan ketidakmiripan biasanya diukur dengan jarak. Menurut Sharma (1996), salah satu faktor yang berpengaruh terhadap hasil *cluster* yang terbentuk adalah jarak antar objek pengamatan. Berdasarkan faktor tersebut dibutuhkan alat ukur untuk menentukan jarak antar objek pengamatan. Berikut ini merupakan metode pengukuran jarak berdasarkan karakteristik variabel yang dikelompokkan,

1. Metode Pengukuran Jarak untuk Variabel Kategorik Biner

Jika variabel yang diamati berupa variabel biner yang hanya memiliki dua macam karakter yang berbeda (0,1), maka variabel yang diamati dapat dibentuk suatu tabel kontingensi seperti pada Tabel 1. Perhitungan ukuran jarak antara variabel x_i dan x_j untuk pengukuran data biner dapat menggunakan beberapa ukuran seperti pada Tabel 2.

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

Tabel 2. Ukuran Jarak Data Biner

Jenis	Rumus
<i>Simple Matching</i>	$SM(x_i, x_j) = \frac{a + d}{a + b + c + d}$
<i>Jaccard</i>	$JACCARD(x_i, x_j) = \frac{a}{a + b + c}$
<i>Russel and Rao</i>	$RR(x_i, x_j) = \frac{a}{a + b + c + d}$
<i>Dice Czekanowski, Sørensen</i>	$DICE(x_i, x_j) = \frac{2a}{2a + b + c}$

2. Metode Pengukuran Jarak untuk Variabel Kategorik Nominal

Pengukuran jarak untuk data variabel nominal memiliki konsep yang sama dengan *simple matching coefficient* maupun *dice*, dimana kategorinya dapat lebih dari dua macam. Dengan jumlah variabel sebanyak m , maka rumus untuk pengukuran jarak variabel nominal antara x_i dan x_j ditunjukkan pada persamaan 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 $x_{il} \neq x_{jl}$. Perhitungan ukuran jarak antara variabel x_i dan x_j untuk pengukuran data biner dapat menggunakan beberapa ukuran sebagai berikut.

3. Metode Pengukuran Jarak untuk Variabel Kategorik Ordinal

Pengukuran jarak untuk data variabel ordinal memiliki konsep yang sama dengan metode untuk data numerik, dimana kategorinya dinyatakan sebagai suatu bilangan bulat. Salah satu metode yang dapat digunakan untuk variabel ordinal adalah jarak *manhattan*. Dengan jumlah variabel sebanyak m , maka rumus untuk pengukuran jarak x_i dan x_j pada variabel nominal ditunjukkan pada persamaan berikut

$$sim(x_i, x_j) = \sum_{l=1}^m |x_{il} - x_{jl}| \quad (2.2)$$

4. Metode Pengukuran Jarak untuk Variabel Numerik

Pada variabel yang memiliki jenis data numerik maka jarak yang dapat digunakan adalah jarak Euclidean. Misalkan terdapat dua observasi dengan variabel-variabel berdimensi m yaitu $x = [x_1, x_2, \dots, x_m]^T$ dan $y = [y_1, y_2, \dots, y_m]^T$. Berikut adalah perhitungan jarak Euclidean yang mengukur jarak antara observasi x dan y sebagai berikut

$$d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \dots + (x_m - y_m)^2}$$

$$d(x, y) = \sqrt{(x - y)^T (x - y)} \quad (2.3)$$

Menurut Hair, *et al.* (2009), suatu *cluster* dapat dinyatakan bahwa *cluster* tersebut merupakan *cluster* yang baik adalah *cluster* yang mempunyai:

1. Homogenitas (kesamaan) yang tinggi antara anggota dalam satu *cluster* (*within-cluster*).
2. Heterogenitas (perbedaan) yang tinggi antara *cluster* yang satu dengan *cluster* yang lain (*between cluster*).

2.3 Metode Pengelompokan

Dalam analisis *cluster*, tahapan pengelompokan dibedakan menurut jenis data yang terdapat dalam variabel. Pada umumnya analisis *cluster* hanya terfokus dengan variabel yang jenis datanya numerik, akan tetapi ada kasus dengan data kategorik bahkan ada kasus dengan data campuran numerik dan kategorik. Analisis *cluster* data kategorik tidak dapat diperlakukan sama dengan data numerik. Hal tersebut dikarenakan sifat khusus data kategorik, sehingga pengelompokan data pada data kategorik lebih rumit dibandingkan dengan pengelompokan data numerik (Hair, *et al.*, 2009).

2.3.1 Clustering Data Kategorik

Clustering data kategorik dilakukan dengan menggunakan ukuran kemiripan atau jarak untuk data berskala kategorik kemudian dapat dilakukan *clustering* dengan menggunakan metode hirarki maupun *non-hirarki*. Menurut Guha, *et al.*, (2002), metode *clustering* hirarki dan *non-hirarki* dinilai tidak tepat digunakan pada data kategorik sehingga dikembangkan metode ROCK untuk *clustering* data kategorik tersebut.

Metode *clustering* yang digunakan untuk tipe data kategorik adalah algoritma ROCK. ROCK pertama kali diperkenalkan oleh Guha, Rastogi, & Shim pada tahun 1999. Metode ROCK menggunakan konsep *link* sebagai ukuran kemiripan

untuk membentuk *cluster*. Metode ROCK dapat menangani *outlier* dengan cukup efektif, pemangkasan *outlier* memungkinkan untuk membuang yang tidak ada hubungan, sehingga titik tersebut tidak berpartisipasi dalam pengelompokan. Namun, dalam beberapa situasi *outlier* dapat hadir sebagai *cluster-cluster* yang kecil (Guha, *et al.*, 2002). *Clustering* untuk data kategorik dengan algoritma ROCK dilakukan dengan tiga langkah. Adapun langkahnya yaitu sebagai berikut:

1. Menghitung similaritas menggunakan rumus Jaccard (Rahayu, 2013). Ukuran kemiripan antara pasangan objek ke- i dan objek ke- j dihitung dengan rumusan yang didefinisikan pada persamaan 2.4

$$sim(X_i, X_j) = \frac{|X_i \cap X_j|}{|X_i \cup X_j|}, i \neq j \quad (2.4)$$

dengan :

$$i = 1, 2, 3, \dots, n$$

$$j = 1, 2, 3, \dots, n$$

$$X_i = \text{himpunan pengamatan ke-}i \text{ dengan } X_i = \{x_{1i}, x_{2i}, \dots, x_{m_i}\}$$

$$X_j = \text{himpunan pengamatan ke-}j \text{ dengan } X_j = \{x_{1j}, x_{2j}, \dots, x_{m_j}\}$$

$$|X| = \text{bilangan kardinal atau jumlah anggota dari himpunan } X$$

2. Menentukan tetangga. Pengamatan dinyatakan sebagai tetangga jika nilai $sim(X_i, X_j) \geq \theta$.
3. Menghitung *link* antar objek pengamatan. Besarnya *link* dipengaruhi oleh nilai *threshold* (θ) yang merupakan parameter yang ditentukan oleh pengguna yang dapat digunakan untuk mengontrol seberapa dekat hubungan antara objek. Besarnya nilai θ yang diinputkan adalah $0 < \theta < 1$.

Metode ROCK menggunakan 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 dikatakan X_i memiliki *link* dengan X_k walaupun X_i bukan tetangga dari X_k . Cara untuk menghitung *link* untuk semua kemungkinan pasangan dari n objek dapat menggunakan matriks. Matriks A

merupakan matriks berukuran $n \times n$ yang bernilai 1 jika X_i dan X_j dinyatakan mirip dan bernilai 0 jika X_i dan X_j tidak mirip.

Jumlah *link* antar pasangan X_i dan X_j diperoleh dari hasil kali antara baris ke X_i dan kolom ke X_j dari matriks A. Jika link antara X_i dan X_j semakin besar maka semakin besar pula kemungkinan X_i dan X_j berada dalam satu *cluster* yang sama (Dutta, *et al.*, 2005).

Penggabungan *cluster* dengan algoritma ROCK didasarkan pada ukuran kebaikan (*goodness measure*) antar *cluster* seperti persamaan 2.4. *Goodness measure* merupakan suatu persamaan yang menghitung jumlah *link* dibagi dengan kemungkinan *link* yang terbentuk berdasarkan ukuran *cluster*-nya (Tyagi & Sharma, 2012).

$$g(C_i, C_j) = \frac{\text{link}[C_i, C_j]}{(n_i + n_j)^{1+2f(\theta)} - n_i^{1+2f(\theta)} - n_j^{1+2f(\theta)}} \quad (2.5)$$

dengan:

$\text{link}[C_i, C_j] = \sum_{X_i \in C_i, X_j \in C_j} \text{link}(X_i, X_j)$ = jumlah *link* dari 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

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

2.3.2 Clustering Data Numerik

Clustering data numerik didasarkan pada jarak atau ukuran ketidakmiripan data. Hasil dari pengelompokkan disajikan dalam bentuk dendogram atau diagram pohon yang memudahkan penelusuran objek sehingga penelusuran informasi menjadi lebih mudah. Dendogram umumnya digunakan untuk membantu memperjelas proses hirarki tersebut (Hair, *et al.*, 2009).

Terdapat dua teknik pengelompokan dalam analisis hirarki. Pengelompokan hirarki diawali dengan pengelompokan dua atau lebih objek yang memiliki kemiripan terdekat, lalu berlanjut dengan objek yang memiliki kemiripan terdekat pada posisi selanjutnya dan seterusnya. Pengelompokan ini disebut dengan pengelompokan dengan teknik pembagian. Algoritma AGNES pertama kali diperkenalkan oleh Kufmann dan Rousseeuw pada tahun 1960. AGNES merupakan algoritma *agglomerative hierarchial clustering* yang cukup populer untuk memproses data numerik (Han & Kamber, 2001). Algoritma AGNES diawali dengan menghitung jarak antar objek, setiap objek difungsikan sebagai *cluster*, kemudian secara bertahap digabungkan sebagai pasangan *cluster* terdekat berdasarkan ukuran jarak dan metode *agglomerative* yang digunakan hingga semua *cluster* tergabung dalam satu *cluster*.

Pada variabel dengan jenis data numerik maka jarak yang dapat digunakan adalah jarak Euclidean. Jarak Euclidean digunakan dalam mengukur jumlah kuadrat perbedaan nilai pada masing-masing variabel.

$$d_i = \sqrt{\sum_{k=1}^p (X_{ik} - X_{jk})^2} \quad (2.6)$$

dengan,

- d_i = jarak antar objek ke-i dan objek ke-j
- p = jumlah variabel *cluster*
- X_{ik} = data dari subjek ke-i pada variabel ke-k
- X_{jk} = data dari subjek ke-j pada variabel ke-k

Metode *agglomerative* yang digunakan yaitu metode *single linkage*, *complete linkage*, dan *average linkage*. Metode *agglomerative* merupakan suatu ukuran terhitung yang digunakan oleh algoritma *clustering hierarchial agglomerative* untuk menggabungkan dua *cluster* C_i dan C_j . Untuk menggabungkan dua *cluster* pada tahap pertama, dimana setiap *cluster* hanya memiliki objek tunggal, diperlukan ukuran ketidakmiripan antar objek. Jika suatu pengelompokan terdiri dari beberapa *cluster* sebelumnya, maka diperlukan ukuran ketidakmiripan antar

cluster, *cluster* dengan ukuran ketidakmiripan terkecil digabungkan dalam suatu *cluster* baru.

Didefinisikan d_{ij} menyatakan ukuran ketidakmiripan antara *cluster*-i dengan *cluster*-j dan $d_{k(i,j)}$ menyatakan ukuran ketidakmiripan antara *cluster* k dengan *cluster* (i,j) yang merupakan gabungan dari *cluster* ke-i dan *cluster* ke-j. jarak antar objek i dan j, maka beberapa teknik *agglomerative* antara *cluster* dinyatakan sebagai berikut (Jhonson & Wichern, 2007).

2.3.2.1 Metode *Single Linkage*

Metode ini merupakan suatu metode penggabungan yang didasarkan pada jarak terkecil atau jarak terdekat. Jika terdapat dua objek yang terpisah dengan jarak yang kecil maka kedua objek tersebut termasuk dalam *cluster* yang sama. Ukuran jarak yang digunakan adalah

$$d_{k(i,j)} = \min (d_{ki}, d_{kj}) \quad (2.7)$$

2.3.2.2 Metode *Complete Linkage*

Metode ini merupakan metode yang definisinya berlawanan dengan metode *single linkage* dikarenakan metode ini menggunakan jarak terbesar sebagai acuan dalam penentuan objek yang akan masuk ke dalam suatu *cluster*. Ukuran jarak yang digunakan adalah

$$d_{k(i,j)} = \max (d_{ki}, d_{kj}) \quad (2.8)$$

2.3.2.3 Metode Average Linkage

Cara kerja metode *average linkage* tidak jauh berbeda dengan kedua metode lainnya. Kriteria yang digunakan pada metode ini adalah rata-rata jarak seluruh objek dalam suatu *cluster* dengan jarak seluruh objek pada *cluster* lain.

Dengan n_i dan n_j merupakan jumlah objek pengamatan dalam *cluster* ke- i dan ke- j . Ukuran jarak yang digunakan adalah

$$d_{k(i,j)} = \frac{n_i}{n_i+n_j} d_{ki} + \frac{n_j}{n_i+n_j} d_{kj} \quad (2.9)$$

Keuntungan penggunaan metode hirarki dalam analisis *cluster* adalah mempercepat pengolahan dan menghemat waktu karena data yang diolah membentuk hirarki atau membentuk tingkatan tersendiri sehingga mempermudah dalam penafsiran, namun kelemahan dari metode ini adalah seringkali terdapat kesalahan pada data *outlier*, perbedaan ukuran jarak yang digunakan, dan terdapatnya variabel yang tidak relevan.

Menentukan jumlah *cluster* optimum merupakan tahapan penting setelah proses *clustering*, tahapan ini disebut dengan tahap validasi *clustering* (Halkidi, *et al.*, 2001). Indeks R^2 merupakan salah satu indeks yang digunakan untuk menentukan jumlah *cluster* optimum pada pengelompokkan hirarki (Sharma, 1996). Indeks R^2 didapatkan dari berbagai macam perhitungan keragaman data, baik keragaman *cluster*, keragaman total, maupun keragaman antar *cluster*. Indeks validitas untuk menentukan jumlah *cluster* optimum pada pengelompokkan hirarki dapat dituliskan sebagai berikut,

1. *Sum of Square Total (SST)*

$$SST = \sum_{l=1}^{m_{num}} \sum_{i=1}^{n_l} (x_{il} - \bar{x}_l)^2 \quad (2.10)$$

2. *Sum of Square Within Group (SSW)*

$$SSW = \sum_{c=1}^C \sum_{l=1}^{m_{num}} \sum_{i=1}^{n_c} (x_{ilc} - \bar{x}_{lc})^2 \quad (2.11)$$

3. *Sum of Square Between Group (SSB)*

$$SSB = SST - SSW \quad (2.12)$$

dengan:

- m_{num} = Jumlah Variabel numerik dalam pengamatan
- C = Jumlah *Cluster* yang dibentuk dalam pengamatan
- n = Total jumlah objek pengamatan
- n_c = Jumlah anggota pada *cluster* ke- c untuk $c = 1, 2, \dots, C$
- \bar{x}_l = Rata-rata keseluruhan objek pada variabel ke- l untuk
 $l = 1, 2, \dots, m_{num}$
- \bar{x}_{lc} = Rata-rata variabel ke- l pada *cluster* ke- c untuk $c = 1, 2, \dots, C$

R^2 *cluster* baru adalah rasio dari SSB dan SST . SSB adalah ukuran perbedaan antar *cluster*, dimana $SST = SSB + SSW$. Semakin besar variasi antar *cluster* maka semakin homogen variasi di dalam masing-masing *cluster*, dan sebaliknya. Dari konsep diatas maka R^2 dapat didefinisikan sebagai ukuran perbedaan antar *cluster*, dengan nilai berkisaran dari 0 sampai 1. Nilai $R^2 = 0$ menunjukkan bahwa tidak terdapat perbedaan antara *cluster*, sedangkan nilai $R^2 = 1$ menunjukkan adanya perbedaan yang signifikan antar *cluster* yang terbentuk. Nilai R^2 dirumuskan pada persamaan

$$R^2 = \frac{SB}{ST} = \frac{[ST-SW]}{ST} \quad (2.13)$$

Penentuan jumlah *cluster* yang terbentuk dapat dilihat dari plot nilai indeks R^2 . Jika terdapat kenaikan nilai indeks R^2 yang sangat tajam pada suatu *cluster*, maka pada titik tersebut merupakan jumlah *cluster* yang optimum. Selain nilai R^2 , penentuan jumlah *cluster* optimum dapat dilakukan dengan melihat nilai F atau perbandingan antara SSB dan SSW yang tertinggi.

Berbeda dengan metode hirarki, metode *non* hirarki dimulai dengan terlebih dahulu jumlah *cluster* yang diinginkan (2 *cluster*, 3 *cluster* atau yang lain).

Setelah jumlah *cluster* diketahui, selanjutnya dilakukan proses pengelompokan. Salah satu contoh metode *non* hirarki adalah K-Means *Cluster* (Hair, *et al.*, 2009). Metode *non* hirarki memiliki keuntungan yaitu efisiensi dalam melakukan analisis sampel dalam ukuran yang lebih besar, serta hanya memiliki sedikit kelemahan pada data *outlier* dan ukuran jarak. Kelemahan metode tersebut adalah titik yang dihasilkan acak. Metode pengelompokan hirarki dan *non* hirarki merupakan metode yang saling melengkapi dalam penyelesaian permasalahan. Akan tetapi, jika dalam pengelompokan objek tidak terdapat informasi apriori tentang *cluster* yang terbentuk, maka metode yang digunakan metode hirarki. Dengan kata lain, hasil dari metode hirarki dapat digunakan sebagai informasi awal untuk pengelompokan objek pada metode *non* hirarki.

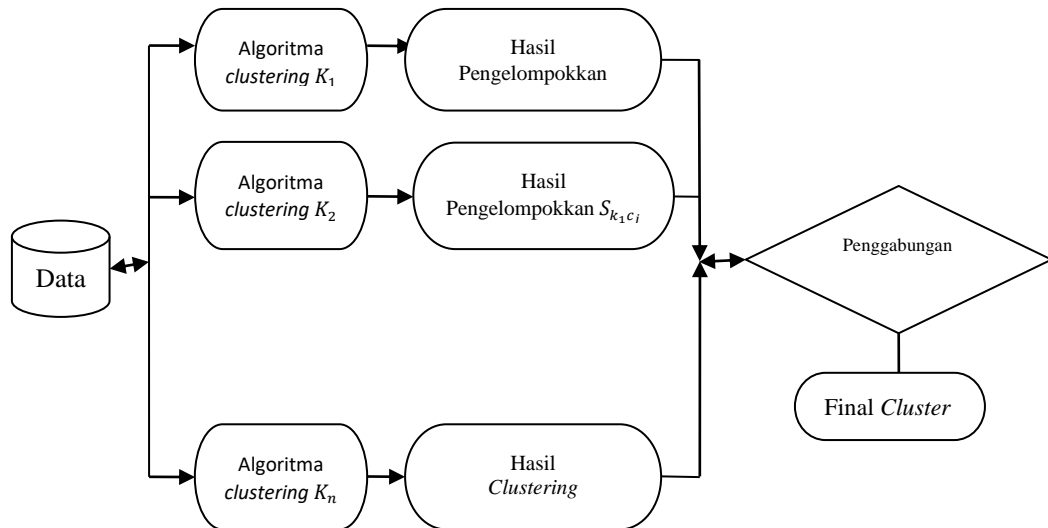
2.3.3 *Clustering Data Campuran*

Pengelompokan *ensemble* merupakan pengelompokan yang menggabungkan beberapa algoritma berbeda untuk mendapatkan partisi umum dari sebuah data, yang bertujuan untuk menggabungkan hasil dari pengelompokan individu (Suguna & Selvi, 2012).

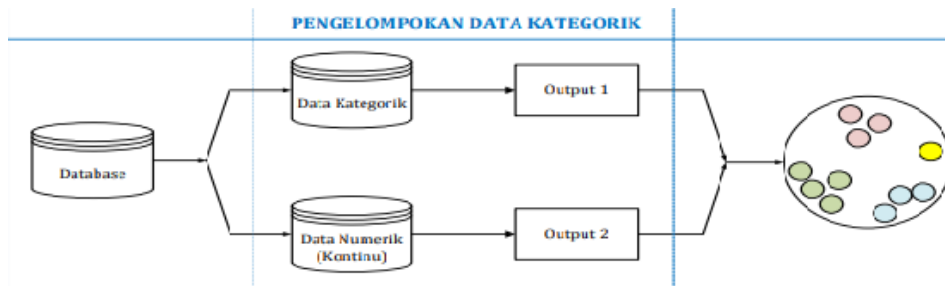
Tujuan dari pengelompokan *ensemble* adalah untuk menggabungkan hasil pengelompokan dari beberapa algoritma pengelompokan yang lebih baik dan *robust* (Yoon, *et al.*, 2006). Pengelompokan *ensemble* terdiri atas dua tahap algoritma yaitu:

1. Melakukan pengelompokan dengan beberapa algoritma dan menyimpan hasil pengelompokan tersebut.
2. Menggunakan fungsi konsensus untuk menentukan *final cluster* dari *cluster-cluster* hasil tahap pertama.

Skema pengelompokan *ensemble* secara umum ditunjukkan pada Gambar 1, sedangkan skema pengelompokan *ensemble* untuk data campuran dengan algoritma CEBMDC secara umum ditunjukkan pada Gambar 2.



Gambar 1. *Flow Chart Pengelompokan ensemble*



Gambar 2. *Flow Chart Algoritma CEBMDC*

Menurut (Zengyou, *et al.*, 2005) langkah-langkah dalam analisis data campuran menggunakan metode pengelompokan *ensemble* yang disebut Algoritma CEBMDC memiliki tahapan sebagai berikut.

1. Membagi data menjadi dua bagian, yaitu murni numerik dan murni kategorik.
2. Melakukan pengelompokan objek dengan variabel numerik dengan algoritma pengelompokan data numerik, serta melakukan pengelompokan objek dengan variabel kategorik dengan algoritma pengelompokan data kategorik.

3. Menggabungkan (*combining*) hasil pengelompokan dari variabel numerik dan kategorik, yang disebut proses *ensemble*.
4. Melakukan pengelompokan *ensemble* menggunakan algoritma pengelompokan data kategorik untuk mendapatkan *cluster* akhir (*final cluster*).

III. METODOLOGI PENELITIAN

3.1 Waktu dan Tempat Penelitian

Penelitian ini dilakukan pada semester ganjil tahun akademik 2021/2022 dan bertempat di Jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung

3.2 Data Penelitian

Data yang digunakan pada penelitian ini adalah data sekunder yang diperoleh dari data Publikasi Badan Pusat Statistik. Data yang digunakan adalah data indikator kesejahteraan rakyat, variabel yang digunakan yaitu indeks gizi usia 0-23 bulan, indeks gizi usia 0-59 bulan, tingkat pengangguran terbuka, angka melek huruf, angka partisipasi sekolah, angka harapan hidup, dan angka morbiditas sebagai variabel numerik dan kepadatan penduduk serta persentase penduduk miskin sebagai variabel kategorik.

3.3 Metode Penelitian

Penelitian ini dilakukan secara studi pustaka, yaitu dengan mengumpulkan serta mempelajari sumber-sumber pustaka yang di peroleh dari buku-buku teks, jurnal, dan internet guna menunjang penulisan penelitian ini. Untuk mempermudah perhitungan, penelitian ini menggunakan *software R*. Adapun tahap-tahap penelitian yang dilakukan adalah sebagai berikut:

1. Mengumpulkan serta memasukkan data ke dalam *software R*
2. Membagi data menjadi dua bagian, yaitu data numerik dan data kategorik.
3. Melakukan pengelompokkan untuk data numerik dengan menggunakan metode hirarki *agglomerative* dan jarak yang digunakan untuk metode hirarki adalah jarak Euclidean.
4. Menentukan k optimum dengan menggunakan R^2 .
5. Memilih pengelompokkan terbaik dari data numerik didasarkan pada nilai rasio S_W dan S_B terkecil. Semakin kecil nilai rasio S_W dan S_B berarti terdapat homogenitas maksimum dalam *cluster*, dan heterogenitas maksimum antar *cluster*.
6. Melakukan pengelompokkan data kategorik menggunakan metode *ROCK*.
7. Memilih pengelompokkan terbaik dari data kategorik didasarkan pada nilai rasio S_W dan S_B terkecil.
8. Menggabungkan hasil dari pengelompokkan terbaik (tahapan *ensemble*)
9. Melakukan pengelompokkan ulang dengan menggunakan metode *ROCK*.
10. Mendapatkan keluaran dari tahapan *ensemble ROCK*.
11. Menganalisis hasil keluaran dari pengelompokkan dengan metode *ensemble ROCK*.

V. SIMPULAN DAN SARAN

5.1 Simpulan

Berdasarkan hasil analisis dan pembahasan yang telah diuraikan pada bab sebelumnya, maka dapat diambil kesimpulan sebagai berikut.

1. Algoritma CEBMDC yang dilakukan untuk menganalisis metode *ensemble* ROCK dapat digunakan untuk menangani permasalahan data dalam melakukan pengelompokkan untuk data berskala campuran numerik dan kategorik.
2. Pengelompokkan data indikator kesejahteraan rakyat di Indonesia menggunakan metode *ensemble* ROCK dengan nilai θ sebesar 0.15 menghasilkan 2 *cluster* dengan nilai rasio sebesar 0.042 ini berarti simpangan baku dalam *cluster* bernilai 0.042 kali dari simpangan baku antar *cluster*. Analisis ini menggunakan metode kombinasi pengelompokkan data numerik dengan metode hirarki *agglomerative* dengan menggunakan jarak Euclidean dengan menggunakan metode *single linkage* dan pengelompokkan data kategorik menggunakan metode ROCK dengan nilai θ sebesar 0.01.

5.2 Saran

Saran yang dapat diambil dari penelitian ini yaitu, untuk akademisi dan penelitian selanjutnya yang tertarik dengan metode *ensemble* ROCK dapat melakukan pengembangan penelitian dengan saran sebagai berikut.

1. Pengelompokkan data numerik dalam penelitian ini hanya menggunakan metode *agglomerative* dengan jarak Euclidean dengan teknik *single linkage*, *complete linkage*, dan *average linkage*, sehingga masih terdapat beberapa pengelompokkan untuk data numerik lainnya seperti metode Ward dan metode lain yang dapat digunakan untuk melakukan pengelompokkan data numerik.
2. Pengelompokkan data kategorik pada penelitian ini menggunakan metode ROCK sehingga dapat dilakukan penelitian lain dengan menggunakan metode K-Modes dan lainnya.
3. Pengelompokkan *ensemble* pada penelitian ini menggunakan algoritma CEBMDC, sehingga penelitian lain yang akan menggunakan metode *ensemble* dapat menggunakan algoritma lain seperti *consensus function*.

DAFTAR PUSTAKA

- Anderson, T.W. & Scolve, S.L. 1974. *Introductory Statistical Analysis*. Houghton Mifflin., Boston.
- Dewangan, R.R., Sharma, L.K., & Akassapu, A.K. 2010. Fuzzy Clustering Technique for Numerical and Categorical Dataset. *International Journal on Computer Science and Engineering*. **1**(13): 75-80
- Dutta, M., Mahanta, A.K., & Pujari, A.K. 2005. QROCK : A Quick Version of the ROCK Algorithm for Clustering of Categorical Data. *Proceedings of the 15 IEEE International Conference on Data Engineering*.
- Gudono. 2011. *Analisis Data Multivariat*. Edisi kedua. BPFE., Yogyakarta.
- Guha, S., Rastogi, R., & Shim, K. 2002. ROCK: A Robust Clustering Algorithm for Categorical Attributes. *Journal of Information System*. **(25)**5:345-366.
- Hair, J.F., Black, W.C., Babin, J.B., & Anderson, E.R. 2009. *Multivariate Data Analysis*. 7th Edition. Prentice Hall, Inc., New Jersey.
- Halkidi, M., Batistakis, Y., & Vazirgiannis, M. 2001. On Clustering Validation Techniques. *Journal of Intelligent Information Systems*. **(17)**2:107-145
- Han, J., & Kamber, M. 2001. *Data Mining : Concepts and Techniques*. Academic Press., USA.
- Johnson, R.A. & Wichern, D.A. 2007. *Applied Multivariate Statistical Analysis*. 6th Edition. Pearson Education, Inc., New Jersey.

- Rahayu, D.P. 2013. Analisis Karakteristik Cluster dengan Menggunakan Cluster Ensemble. *Jurnal Matematika, Sains, dan Teknologi* (14)1:1-10
- Sharma, S. 1996. *Applied Multivariate Technique*. John Willey and Sons, Inc., New York.
- Simamora, B. 2005. *Analisis Multivariat Pemasaran*. Edisi pertama. PT. Gramedia Pustaka Utama., Jakarta.
- Suguna, J., & Selvi, M.A. 2012. Ensemble Fuzzy Clustering for Mixed Numerical and Categorical Data. *International Journal of Computer Application*. (42)3: 19-23
- Supranto. 2004. *Analisis Multivariat dan Interpretasi*. Rineka Cipta, Jakarta.
- Tyagi, A., & Sharma, S. 2012. Implementation of Rock Clustering Algorithm for the Optimization of Query Searching Time. *International Journal on Computer Science and Engineering*. (4)5:809-815
- Yoon, H.S., Ahn, S.Y., Lee, S.H., Cho, S.B., & Kim, J.H. 2006. Heterogeneous Clustering Ensemble Method For Combining Different Cluster Results. *Proceedings of Data Mining for Biomedical Application*, Singapore : 9 April 2006. Page 82- 92
- Zengyou, H., Xiaofei, X., & Shengcun, D 2002. Squeezer : An Efficient Algorithm for Clustering Categorical Data. *Journal of Computer Science and Tcehnology*. 17(5):611-625.
- Zengyou, H., Xiaofei, X., & Shengcun, D. 2005. Clustering Mixed Numeric and Categorical Data : A Cluster Ensamble Approach. *Computer Science*. Cornell University. <https://arxiv.org/abs/cs/0509011>. Diakses pada tanggal 27 Juli 2021.