SIMULASI PEMILIHAN METODE ANALISIS *CLUSTER* HIRARKI *AGGLOMERATIVE* TERBAIK ANTARA *AVERAGE LINKAGE* DAN WARD PADA DATA YANG MENGANDUNG MASALAH MULTIKOLINEARITAS

(Skripsi)

Oleh

RIZKI AGUNG WIBOWO

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMPUNG
BANDAR LAMPUNG

2020

ABSTRACT

Selection Simulation of The Best Hierarchial Agglomerative Clustering Method between Average Linkage and Ward's on Data Containing Multicollinearity Problems

By

Rizki Agung Wibowo

Multicollinearity is a linear relation (collinearity) that exists between independent variables. In cluster analysis, the effect is different because multicollinearity is a form of implicit weighting. Principal component analysis can be used to reduce the number of variables that correlated into the number of new variables that uncorrelated by maintaining as much variety of data, by using the result of principal component analysis, we can do cluster analysis by average linkage and Ward's methods, then the best method will be chosen based on Dunn and RS indices, it was concluded that Ward's method is better than average linkage based on RS index which means that cluster formed using Ward's has more different characteristics than average linkage method while using Dunn index it can be concluded that average linkage is better than Ward's method which means that cluster formed using average linkage has more compactness than Ward's method.

Keyword : Multicollinearity, Average Linkage Method, Ward's Method, Principal Component Analysis, Dunn Index, RS Index

ABSTRAK

Simulasi Pemilihan Metode Analisis *Cluster* Hirarki *Agglomerative* Terbaik antara *Average Linkage* dan Ward pada Data yang Mengandung Masalah Multikolinearitas

Oleh

Rizki Agung Wibowo

Multikolinearitas adalah hubungan linear yang ada di antara variabel independen, pada analisis klaster efek yang ditimbulkan oleh multikolinearitas berbeda, dikarenakan pada dasarnya multikolinearitas adalah bentuk pembobotan implisit. Analisis komponen utama dapat digunakan untuk mereduksi jumlah himpunan peubah yang banyak dan saling berkorelasi menjadi peubah-peubah baru yang tidak berkorelasi dengan mempertahankan sebanyak mungkin keragaman data tersebut, dengan menggunakan hasil analisis komponen utama dilakukan analisis klaster menggunakan metode *average linkage* dan Ward, yang kemudian akan dipilih metode terbaiknya berdasarkan nilai indeks Dunn dan indeks R-Sq (RS), didapat kesimpulan bahwa metode Ward lebih baik dibandingkan *average linkage* yang ditinjau berdasarkan indeks RS yang berarti klaster yang terbentuk dengan menggunakan metode Ward memiliki karakteristik yang lebih berbeda dibanding dengan metode *average linkage*, sedangkan dengan menggunakan indeks Dunn

didapatkan kesimpulan bahwa metode *average linkage* lebih baik dibandingkan metode Ward yang berarti klaster yang terbentuk dengan menggunakan metode *average linkage* lebih kompak dibanding dengan metode Ward.

Kata kunci : Multikolinearitas, Metode *Average Linkage,* Metode Ward, Analisis Komponen Utama, Indeks Dunn, Indeks RS

SIMULASI PEMILIHAN METODE ANALISIS *CLUSTER* HIRARKI *AGGLOMERATIVE* TERBAIK ANTARA *AVERAGE LINKAGE* DAN WARD PADA DATA YANG MENGANDUNG MASALAH MULTIKOLINEARITAS

Oleh

Rizki Agung Wibowo

Skripsi

Sebagai Salah Satu Syarat untuk Mencapai Gelar
SARJANA MATEMATIKA

pada

Jurusan Matematika
Fakultas Matematika dan Ilmu Pengetahuan Alam

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMPUNG
BANDAR LAMPUNG
2020

Judul Skripsi	: **SIMULASI PEMILIHAN METODE ANALISIS *CLUSTER* HIRARKI *AGGLOMERATIVE* TERBAIK ANTARA *AVERAGE LINKAGE* DAN WARD PADA DATA YANG MENGANDUNG MASALAH MULTIKOLINEARITAS**
Nama Mahasiswa	: Rizki Agung Wibowo
Nomor Pokok Mahasiswa	: 1617031021
Program Studi	: Matematika
Fakultas	: Matematika dan Ilmu Pengetahuan Alam

MENYETUJUI

1. Komisi Pembimbing

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

Dr. Ahmad Faisol, S.Si., M.Sc.
NIP 19800206 200312 1 003

2. Ketua Jurusan Matematika

Prof. Dra. Wamiliana, M.A., Ph.D.
NIP 19631108 198902 2 001

MENGESAHKAN

1. Tim Penguji

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

Sekretaris : **Dr. Ahmad Faisol, S.Si., M.Sc.** .....

Penguji
Bukan Pembimbing : **Drs. Eri Setiawan, M.Si.** .....

2. Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam

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

Tanggal Lulus Ujian Skripsi : **25 Februari 2020**

PERNYATAAN

Yang bertanda tangan di bawah ini ;

Nama : Rizki Agung Wibowo

Nomor Pokok Mahasiswa : 1617031021

Jurusan : Matematika

Dengan ini menyatakan bahwa skripsi saya yang berjudul **"SIMULASI PEMILIHAN METODE ANALISIS *CLUSTER* HIRARKI *AGGLOMERATIVE* TERBAIK ANTARA *AVERAGE LINKAGE* DAN WARD PADA DATA YANG MENGANDUNG MASALAH MULTIKOLINEARITAS"** adalah hasil pekerjaan saya sendiri. Semua hasil tulisan dalam skripsi ini telah mengikuti kaidah penulisan karya ilmiah Universitas Lampung. Apabila kemudian hari terbukti bahwa skripsi ini merupakan hasil salinan atau telah dibuat orang lain, maka saya bersedia menerima sanksi sesuai ketentuan akademik yang berlaku.

Bandar Lampung, Maret 2020

Penulis

Rizki Agung Wibowo
NPM. 1617031021

RIWAYAT HIDUP

Penulis dilahirkan di Pringsewu pada tanggal 23 Mei 1998, sebagai anak pertama dari pasangan Bapak Priyo Satmono dan Ibu Sri Sukarni serta kakak dari Sekar Arum Purbo Kinasih.

Penulis telah menempuh pendidikan di Taman Kanak-kanak (TK) Al-Azhar 7 Hajimena pada tahun 2003-2004, Sekolah Dasar Negeri 1 Rajabasa Raya pada tahun 2004-2010, Sekolah Menengah Pertama Negeri (SMPN) 8 Bandar Lampung pada tahun 2010-2013, dan Sekolah Menengah Atas Negeri (SMAN) 14 Bandar Lampung pada tahun 2013-2016.

Pada tahun 2016 penulis terdaftar sebagai Mahasiswa Program Studi S1 Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung melalui jalur SNMPTN. Selama menjadi Mahasiswa, penulis cukup aktif dalam organisasi kampus antara lain menjadi generasi muda HIMATIKA (Himpunan Mahasiswa Matematika) dan ROIS (Rohani Islam) FMIPA pada periode 2016-2017, Anggota Bidang Keilmuan HIMATIKA periode 2017-2018, Anggota Bidang Hubungan Masyarakat (HUMAS) ROIS FMIPA periode 2017-2018, Ketua Bidang Infokom ROIS FMIPA periode 2018, Kepala Dinas Medinfo BEM FMIPA periode 2019 dan pernah menjadi asisten praktikum Pengantar Teknologi Informasi, Pengantar Analisis Numerik, Simulasi.

Pada bulan Januari 2019 penulis melaksanakan Kuliah Kerja Nyata (KKN) di Kampung Segara Midar, Kecamatan Blambangan Umpu, Kabupaten Way Kanan sebagai bentuk pengabdian mahasiswa dan menjalankan Tri Dharma Perguruan Tinggi. Sebagai bentuk penerapan ilmu yang telah dipelajari, pada bulan Juli 2019 penulis melaksanakan Kerja Praktik (KP) di Badan Pusat Statistik Provinsi Lampung.

KATA MUTIARA

"Hai orang-orang beriman, jadikanlah sabar dan shalat sebagai penolongmu, sesungguhnya Allah beserta orang-orang yang sabar"
(Qs. Al-Baqarah: 153)

"Perumpamaan (nafkah yang dikeluarkan oleh) orang-orang yang menafkahkan hartanya di jalan Allah adalah serupa dengan sebutir benih yang menumbuhkan tujuh bulir, pada tiap-tiap bulir seratus biji. Allah melipat gandakan (ganjaran) bagi siapa yang Dia kehendaki. Dan Allah Maha Luas (karunia-Nya) lagi Maha Mengetahui."
(Qs. Al-Baqarah: 261)

PERSEMBAHAN

Dengan mengucap puji dan syukur kehadirat Allah SWT yang telah memberikan petunjuk dan kemudahan untuk menyelesaikan studiku, kupersembahkan karya kecilku ini untuk:

Ayah dan Ibu tercinta yang selalu mendidik, mendoakan, berkorban, dan hal lain yang tak dapatku ungkapkan dengan kata-kata

Adik ku tersayang

Dosen pembimbing dan penguji yang sangat berjasa dan tidak lelah memberikan arahan serta masukan sehingga peulis dapat menyelesaikan skripsiku

Sahabat dan teman-temanku, Terimakasih atas kebersamaan, do'a dan semangat yang selalu kalian berikan kepadaku.

Universitas Lampung

SANWACANA

Alhamdulillahi Robbil ‘alamin, Puji dan syukur Penulis ucapkan kepada Allah SWT, yang selalu melimpahkan rahmat dan kasih sayang-Nya, sehingga Penulis dapat menyelesaikan skripsi ini. Sholawat serat salam senantiasa tetap tercurah kepada Nabi Muhammad SAW, tuntunan dan tauladan utama bagi seluruh umat manusia.

Skripsi dengan judul “Simulasi Pemilihan Metode Analisis *Cluster* Hirarki *Agglomerative* Terbaik antara *Average Linkage* dan Ward pada Data yang Mengandung Masalah Multikolinearitas” adalah salah satu syarat untuk memperoleh gelar Sarjana Matematika di Universitas Lampung.

Dalam menyelesaikan skripsi ini, banyak pihak yang telah membantu Penulis dalam memberikan bimbingan, dorongan, dan saran-saran. Sehingga dengan segala ketulusan dan kerendahan hati pada kesempatan ini Penulis mengucapkan terimakasih yang sebesar-besarnya kepada:

1. Dr. Khoirin Nisa, M.Si. selaku Dosen Pembimbing 1 utama dan Dosen Pembimbing Akademik yang senantiasa memberikan bimbingan, saran, motivasi, nasehat serta masukan sehingga penulis dapat menyelesaikan perkuliahan dan skripsi ini

2. Dr. Ahmad Faisol, S.Si., M.Sc. selaku Dosen Pembimbing 2 yang telah memberikan masukan dan saran dalam penyelesaian skripsi
3. Drs. Eri Setiawan, M.Si. selaku Dosen Pembahas yang telah memberikan kritik dan saran kepada penulis dalam penyelesaian skripsi
4. Prof. Dra. Wamiliana, MA, Ph.D., selaku Ketua Jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam
5. Dr. Eng. Suripto Dwi Yuwono, M.T. selaku Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung
6. Dosen, staf dan karyawan Jurusan Matematika FMIPA Universitas Lampung yang telah memberikan ilmu dan bantuan kepada penulis
7. Ayah dan Ibu yang tidak pernah lelah memberikan do'a, dukungan, kasih sayang, dan pengorbanan kepada Penulis
8. Adik ku tersayang
9. Teman-teman terbaik Presidium Inti, Rajin Kok dan Butuh Hijrah yang selalu memberikan keceriaan, kebersamaan, dan menjadi tempat berbagi
10. Teman-teman Keluarga Keilmuan HIMATIKA
11. Pimpinan ROIS FMIPA 2018 beserta anggota
12. Pimpinan BEM FMIPA 2019 beserta anggota
13. Teman-teman seperbimbingan Bu Nisa
14. Teman-teman seperjuangan KKN periode 1 tahun 2019 Kampung Segara Midar, Blambangan Umpu, Way Kanan
15. Elita Apriliana dan Junia Rahma N.I. yang bersedia direpotkan oleh penulis
16. Teman-teman Matematika Angkatan 2016.

Penulis juga menyadari bahwa dalam penulisan skripsi ini masih banyak terdapat kekurangan. Oleh karena itu, Penulis mengharapkan saran dan kritik yang membangun guna penelitian selanjutnya agar lebih baik.

Bandar Lampung, Maret 2020

Penulis,

Rizki Agung Wibowo

DAFTAR ISI

Halaman

DAFTAR GAMBAR ..... **vii**

DAFTAR TABEL ..... **viii**

I. PENDAHULUAN
1.1 Latar Belakang dan Masalah 1
1.2 Tujuan Penelitian 4
1.3 Manfaat Penelitian 4

II. TINJAUAN PUSTAKA
2.1 Analisis Klaster 5
2.1.1 Merumuskan Masalah 5
2.1.2 Nilai Z (*Z-Score*) 6
2.1.3 Asumsi pada Analisis Klaster 6
2.1.3.1 Uji (Asumsi) Multikolinearitas 6
2.1.4 Ukuran Kedekatan (Jarak) 8
2.1.5 Analisis Klaster Hirarki 9
2.1.5.1 Metode *Agglomerative* 10
2.1.5.1.1 Metode *Average Linkage* 11
2.1.5.1.2 Metode Ward 12
2.2 Indeks Dunn (*Dunn Index*) 15
2.3 Indeks RS (*R-Squared*) 16
2.4 Analisis Komponen Utama (AKU) 17
2.4.1 Matriks Kovarian ($\boldsymbol{S}$) 19
2.4.2 Nilai Eigen (λ) 20
2.4.3 Vektor Eigen ($\boldsymbol{a}$) 21
2.5 Uji Normal Multivariat Henze Zirkler 22

III. METODOLOGI PENELITIAN
3.1 Waktu dan Tempat Penelitian 24
3.2 Data Penelitian 24
3.3 Metode Penelitian 25

IV. HASIL DAN PEMBAHASAN
4.1 Membangkitkan Data 30
4.2 Standarisasi Data 33
4.3 Uji Asumsi Multikolinearitas 34
4.4 Mengatasi Data Multikolinearitas menggunakan AKU 35
4.5 Analisis Klaster menggunakan Metode *Average Linkage* dan Metode Ward 39
4.5.1 Analisis Klaster menggunakan Metode *Average Linkage* 40
4.5.2 Analisis Klaster menggunakan Metode Ward 43
4.6 Menghitung Indeks Dunn (*Dunn Index*) 48
4.7 Menghitung Indeks RS 51
4.8 Proses Pengulangan Sebanyak 1000 Kali 53
4.9 Analisis Hasil 54

V. KESIMPULAN
5.1 Kesimpulan 56
5.2 Saran 57

DAFTAR PUSTAKA

LAMPIRAN

DAFTAR GAMBAR

Gambar Halaman

1. Klasifikasi Prosedur Pengklasteran.....9
2. Algoritma (*flowchart*) penelitian 29
3. *Scree plot* proporsi nilai eigen pada contoh data yang telah terstandarisasi dengan jumlah objek 10 36
4. Plot komponen utama 1 dan komponen utama 2 pada contoh data yang telah terstandarisasi dengan $n = 10$ 38
5. Dendogram metode *average linkage* (kiri) dan Ward (kanan) dengan menggunakan matriks $\boldsymbol{KU}$ 46
6. Dendogram metode *average linkage* (kiri) dan Ward (kanan) dengan menggunakan hasil analisis komponen utama pada data dengan $n = 20$ 47
7. Dendogram metode *average linkage* (kiri) dan Ward (kanan) dengan menggunakan hasil analisis komponen utama pada data dengan $n = 50$ 47
8. Dendogram metode *average linkage* (kiri) dan Ward (kanan) dengan menggunakan hasil analisis komponen utama pada data dengan $n = 100$ 48
9. Plot rata-rata indeks Dunn.....54
10. Plot rata-rata indeks RS.....55

DAFTAR TABEL

Tabel Halaman

1. Contoh data klaster pertama dibangkitkan dengan $n = 5$ berdistribusi $X_j \sim N(0,1)$ 30
2. Contoh data klaster kedua dibangkitkan dengan $n = 5$ berdistribusi $X_j \sim N(5,2)$ 31
3. Output uji normal multivariat Henze-Zirkler pada data Tabel 1 dan 2 31
4. Penggabungan contoh data pada Tabel 1 dan 2 32
5. Contoh data yang mengandung multikolinearitas dengan $n = 10$ 33
6. Contoh data multikolinearitas dengan $n = 10$ yang telah terstandarisasi 34
7. Output nilai VIF dan kesimpulan pada masing-masing contoh data yang dibangkitkan 34
8. Nilai eigen dan proporsi keragaman yang dijelaskan setiap nilai eigen matriks kovarian pada contoh data yang telah terstandarisasi dengan $n = 10$ 36
9. Output nilai VIF dan kesimpulan pada komponen utama 1 dan 2 39
10. Nilai indeks Dunn hasil analisis klaster pada Sub Bab 4.5 48
11. Indeks RS hasil analisis klaster pada Sub Bab 4.5 51
12. Nilai rata-rata tiap variabel pada klaster yang sama 52
13. Nilai indeks Dunn dan indeks RS hasil analisis klaster pada data yang dilakukan pembangkitan sebanyak 1000 kali pengulangan 53
14. Rata-rata nilai indeks Dunn dan indeks RS hasil analisis klaster pada data yang dilakukan pembangkitan sebanyak 1000 kali pengulangan 54

I. PENDAHULUAN

1.1 Latar Belakang dan Masalah

Menurut Supranto (2004), analisis klaster meneliti seluruh hubungan interdependensi, tidak ada pembedaan variabel bebas, dan tak bebas. Tujuan utama analisis klaster adalah mengklasifikasikan objek (kasus atau elemen) ke dalam kelompok-kelompok yang relatif homogen didasarkan pada suatu set variabel yang dipertimbangkan untuk diteliti. Pada analisis klaster digunakan jarak euclid sebagai alat ukur kedekatan (Rencher, 2002). Semakin kecil besaran jarak suatu objek terhadap objek lain, maka semakin besar kemiripan individu tersebut (Usman dan Nurdin, 2013). Analisis klaster dibagi menjadi dua metode yaitu hirarki dan non hirarki, pada metode hirarki mula-mula setiap objek dianggap sebagai klaster yang kemudian dikelompokkan dengan cara mencari objek-objek yang memiliki jarak terdekat hingga pada akhirnya setiap objek akan menjadi satu klaster besar (Johnson, 1967).

Didalam analisis klaster hirarki terdapat dua metode yaitu *agglomerative* dan *divissive*, pada metode *agglomerative* dibagi kembali menjadi lima metode, yaitu *single linkage, complete linkage, average linkage, Ward's* dan *centroid* (Hardle dan Simar, 2007). Ada dua asumsi yang harus dipenuhi pada analisis klaster,

yaitu sampel representatif dan multikolinearitas antara tiap variabel (Hair dkk., 2014). Multikolinearitas adalah hubungan linear yang ada di antara variabel independen. Multikoliniearitas dapat dilihat dari nilai *Variance Inflation Factor* (VIF), jika nilai VIF melebihi angka 10 maka dapat disimpulkan ada multikolinearitas (Widarjono, 2010).

Menurut Hair dkk. (2014), pada analisis klaster efek yang ditimbulkan oleh multikolinearitas berbeda dengan analisis multivariat yang lainnya, dikarenakan pada dasarnya multikolinearitas adalah bentuk pembobotan implisit pada tiap variabelnya sedangkan pada analisis klaster setiap variabel diberikan bobot yang sama. Masalah multikolinearitas ini dapat diatasi dengan menggunakan analisis komponen utama (AKU) dengan cara mereduksi dimensi suatu data kedalam suatu dimensi seminimal mungkin dengan tetap mempertahankan informasi yang tekandung didalamnya.

Analisis komponen utama (AKU), merupakan analisis tertua dalam APG yang diperkenalkan oleh Karl Pearson tahun 1901, diantaranya dapat digunakan untuk mereduksi jumlah himpunan peubah yang banyak dan saling berkorelasi menjadi peubah-peubah baru yang tidak berkorelasi dengan mempertahankan sebanyak mungkin keragaman data tersebut (Johnson dan Wichern, 2002).

Rujasiri dan Chomtee (2009), dalam jurnal yang berjudul *Comparison of Clustering Techniques for Cluster Analysis* membahas tentang perbandingan *average linkage* dan ward dengan menggunakan data yang tidak mengandung masalah

multikolinearitas dan didapat kesimpulan bahwa metode Ward merupakan metode analisis klaster terbaik, digunakan RMSSTD dan R-Sq sebagai alat ukur untuk menentukan metode analisis klaster terbaiknya. Terdapat alat ukur lainnya yang digunakan untuk menentukan metode analisis klaster terbaik, antara lain indeks Dunn.

Alat ukur indeks Dunn berlandaskan pada fakta bahwa klaster yang terpisah itu biasanya memiliki jarak antar klaster yang besar dan diameter intra klaster yang kecil dengan cara menghitung rasio atau pembagian dari jarak terkecil antar objek pada klaster yang berbeda terhadap jarak intra klaster (Brock dkk., 2008), sedangkan RMSSTD mengukur kehomogenan dari kelompok yang terbentuk pada setiap tahap dengan menghitung standar deviasinya dan indeks R-Sq (RS) mengukur apakah karakteristik antar klaster saling berbeda.

Berdasarkan penelitian yang dilakukan oleh Rujasiri dan Chomtee (2009), penulis termotivasi untuk melakukan penelitian tentang simulasi pemilihan metode terbaik antara dua metode klaster *agglomerative average linkage* dan Ward dalam mengelompokan objek-objek pada data yang mengandung masalah multikolinearitas dan menggunakan indeks Dunn dan indeks RS dalam menentukan metode terbaiknya.

1.2 Tujuan Penelitian

Penelitian ini bertujuan untuk memilih metode terbaik antara dua metode klaster hirarki *agglomerative average linkage* dan Ward dengan menggunakan indeks Dunn dan indeks RS pada data yang mengandung masalah multikolinearitas.

1.3 Manfaat Penelitian

Manfaat yang diperoleh dari penelitian ini adalah:

1. Memberikan sumbangan pemikiran dalam rangka memperluas dan memperdalam pengetahuan ilmu statistika khususnya mengenai analisis klaster hirarki *agglomerative average linkage* dan Ward pada data yang mengandung masalah multikolinearitas
2. Memberikan masukan dan dorongan bagi peneliti yang lain agar dapat mengkaji lebih lanjut tentang analisis klaster hirarki *agglomerative average linkage* dan Ward pada data yang mengandung masalah multikolinearitas.

II. TINJAUAN PUSTAKA

Pada bab ini akan dibahas definisi, asumsi serta contoh tentang analisis klaster, indeks Dunn, RS dan analisis komponen utama (AKU).

2.1 Analisis Klaster

Menurut Supranto (2004), analisis klaster meneliti seluruh hubungan interdependensi, tidak ada pembedaan variabel bebas, dan tak bebas. Tujuan utama analisis klaster adalah mengklasifikasikan objek (kasus atau elemen) ke dalam kelompok-kelompok yang relatif homogen didasarkan pada suatu set variabel yang dipertimbangkan untuk diteliti. Objek atau kasus dalam setiap kelompok cenderung mirip satu sama lain dan berbeda jauh (tidak sama) dengan objek dari klaster lainnya.

2.1.1 Merumuskan Masalah

Hal yang paling penting di dalam masalah analisis klaster adalah pemilihan variabel-variabel yang akan dipergunakan untuk pengklasteran. Pada dasarnya set variabel yang akan dipilih harus menguraikan kemiripan (*similarity*) antara objek. Variabel harus dipilih berdasarkan penelitian sebelumnya, teori atau suatu pertimbangan berkenaan dengan hipotesis yang akan diuji (Supranto, 2004).

2.1.2 Nilai Z (*Z-Score*)

Menurut Lind dkk. (2007), Nilai Z adalah jarak yang bertanda antara sebuah nilai X yang dipilih dari rata-rata (μ) dibagi dengan standar deviasinya (σ). Jadi sebuah nilai Z adalah jarak dari rata-rata diukur dalam unit standar deviasinya dalam bentuk rumus:

$$Z = \frac{X - \mu}{\sigma} \tag{2.1}$$

2.1.3 Asumsi pada Analisis Klaster

Menurut Hair dkk. (2014), Analisis klaster bukanlah teknik statistik inferensia dimana parameter dari sampel dapat menilai dan mewakili suatu populasi. Analisis klaster adalah metode untuk mengukur karakteristik struktural dari serangkaian pengamatan. Pada analisis klaster asumsi seperti normalitas, linearitas dan homoskedastisitas tidak banyak berpengaruh. Ada dua asumsi yang harus dipenuhi pada analisis klaster, yaitu:

1. Sampel representatif
2. Multikolinearitas antara tiap variabel.

Pada penelitian ini hanya difokuskan pada asumsi multikolinearitas.

2.1.3.1 Uji (Asumsi) Multikolinearitas

Multikolinearitas adalah hubungan linear yang ada di antara variabel independen. Multikoliniearitas dapat dilihat dari nilai *Variance Inflation Factor* (VIF).

Rumus untuk menghitung VIF yaitu sebagai berikut:

$$VIF_i = \frac{1}{(1 - R_i^2)} \tag{2.2}$$

Keterangan :

R_i^2 : koefisien determinasi pada variabel i

i : $1,2,3, \dots , p$

Jika nilai VIF melebihi angka 10 maka dapat disimpulkan ada multikolinearitas (Widarjono, 2010).

Pada analisis klaster efek yang ditimbulkan oleh multikolinearitas berbeda, dikarenakan pada dasarnya multikolinearitas adalah bentuk pembobotan implisit, misalkan responden dikelompokan pada 10 variabel, semua pernyataan sikap tentang suatu pelayanan. Ketika multikolinearitas diperiksa, didapat 2 variabel, yang pertama terdiri atas delapan pernyataan dan yang kedua terdiri atas dua pernyataan tersisa. Jika yang dimaksud adalah benar untuk mengelompokan responden pada dimensi pelayanan (dalam hal ini diwakili oleh dua kelompok variabel), maka menggunakan kesepuluh variabel asli akan salah. Karena setiap variabel diberikan bobot yang sama dalam analisis klaster, maka dimensi pertama memiliki peluang empat kali lebih banyak (delapan item dibandingkan dengan dua item) untuk mempengaruhi ukuran kesamaannya. Akibatnya kesamaan akan dipengaruhi secara dominan oleh dimensi pertama dengan delapan item daripada dimensi kedua dengan dua item (Hair dkk., 2014).

2.1.4 Ukuran Kedekatan (Jarak)

Analisis klaster berupaya mengidentifikasi dari vektor-vektor pengamatan yang serupa dan mengelompokkannya menjadi kelompok-kelompok, banyak teknik menggunakan indeks kesamaan atau kedekatan antara setiap pasang pengamatan. Kedekatan atau pendekatan yang biasa digunakan adalah mengukur kemiripan yang dinyatakan dalam jarak antara pasangan objek. Semakin kecil besaran jarak suatu individu terhadap individu lain, maka semakin besar kemiripan individu tersebut, sehingga individu tersebut akan dimasukkan dalam kelompok yang sama (Usman dan Nurdin, 2013).

Pada analisis klaster digunakan jarak Euclid sebagai alat ukur kedekatan, yang didefinisikan sebagai berikut:

$$d(x,y) = \sqrt{(x-y)'(x-y)} = \sqrt{\sum_{j=1}^{p}(x_j - y_j)^2} \tag{2.3}$$

(Rencher, 2002)

Atau dapat ditulis sebagai berikut:

$$d(i,j) = d_{ij} = \sqrt{\sum_{k=1}^{p}(x_{ik} - x_{jk})^2} \tag{2.4}$$

Keterangan :

d_{ij} : Jarak antara objek ke-i dan obyek ke-j

p : Jumlah variabel klaster

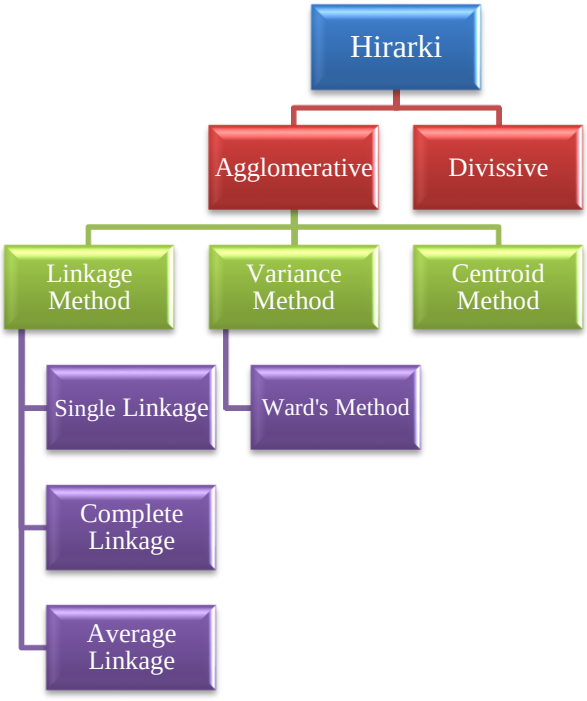
x_{ik} : data dari objek ke-i pada variabel ke-k

x_{jk} : data dari objek ke-j pada variabel ke-k

2.1.5 Analisis Klaster Hirarki

Prosedur pembentukan klaster terbagi menjadi 2, yaitu hirarki dan non-hirarki. Pembentukan klaster hirarki mempunyai sifat sebagai pengembangan suatu hirarki atau struktur mirip pohon bercabang. Metode klaster hirarki merupakan metode pengelompokan yang mana jumlah kelompok yang akan dibuat belum diketahui, teknik ini diproses melalui penggabungan berurutan (*agglomerative*) atau pembagian berurutan (*divissive*).

Teknik *agglomerative* terdiri atas 3 metode, yaitu metode *Linkage*, metode *Variance* dan metode *Centroid*. Metode Linkage terdiri dari metode *Single Linkage*, *Complete Linkage* dan *Average Linkage*, sedangkan metode *Variance* terdiri atas metode Ward.



Gambar 1. Klasifikasi Prosedur Pengklasteran

Cara kerja metode klaster hirarki yaitu, diberikan sekumpulan N item yang akan di klaster, dan sebuah matriks $N\ x\ N$ yang menyatakan jarak antar item pada N:

1. Mulai dengan membuat klaster sebanyak N, masing-masing klaster mempunyai sebuah item
2. Cari sepasang klaster yang jaraknya terdekat dan dijadikan sebuah klaster baru
3. Hitung jarak antar klaster yang baru dengan masing-masing klaster yang lainnya.

Ulangi langkah 2 dan 3 sampai semua item menjadi sebuah klaster dengan N item (Johnson, 1967).

2.1.5.1 Metode *Agglomerative*

Metode *agglomerative* dimulai dengan menganggap bahwa tiap objek adalah sebuah klaster, kemudian dua objek dengan jarak terdekat digabungkan menjadi satu klaster, selanjutnya objek ketiga akan bergabung dengan klaster yang ada atau akan bersama dengan objek lain membentuk klaster berikutnya, dan seterusnya hingga terbentuk satu klaster yang terdiri dari keseluruhan objek. Metode hirarki *agglomerative* terbagi menjadi lima metode dalam pembentukan klaster yaitu (Hardle dan Simar, 2007):

1. Pautan tunggal (*Single Linkage*)
2. Pautan lengkap (*Complete Linkage*)
3. Pautan rata-rata (*Average Linkage*)
4. Metode Ward (*Ward's Method*)
5. Metode centroid (pusat)

2.1.5.1.1 Metode *Average Linkage*

Menurut Johnson dan Wichern (2007), pada pendekatan *average linkage*, jarak antara dua klaster didefinisikan sebagai jarak rata-rata antara semua anggota didalam satu klaster dengan semua anggota pada klaster lain. *Average linkage* menggunakan jarak terdekat dan metode ini dapat digunakan untuk mengelompokan objek atau variabel

$$d_{(UV)W} = \frac{\sum_i \sum_k d_{ik}}{N_{(UV)} N_W} \tag{2.5}$$

Keterangan :

d_{ik} : jarak antar objek i pada klaster (UV) dan objek k pada klaster W

$N_{(UV)}$: jumlah item pada klaster (UV)

N_W : jumlah item pada klaster W

Contoh 1:

Diketahui matriks jarak euclid antara 4 objek sebagai berikut:

$$D = d_{ij} = \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \end{matrix} \begin{bmatrix} 0 & 2 & 3 & 4 \\ 2 & 0 & 1 & 5 \\ 3 & 1 & 0 & 6 \\ 4 & 5 & 6 & 0 \end{bmatrix}$$

Langkah-langkah pengklasteran adalah sebagai berikut:

1. Jarak objek yang berdekatan adalah objek 2 dan 3 yaitu sebesar 1, maka gabungkan objek 2 dan 3 pada satu klaster (23)

2. Hitung jarak klaster (23) dengan objek lain

   - $d_{(23)1} = \frac{d_{21}+d_{31}}{2} = \frac{2+3}{2} = 2{,}5$
   - $d_{(23)4} = \frac{d_{24}+d_{34}}{2} = \frac{5+6}{2} = 5{,}5$
   - $d_{(23)(23)} = \frac{d_{22}+d_{23}+d_{32}+d_{33}}{4} = \frac{0+1+1+0}{4} = 0{,}5$

   Didapat matriks jarak baru

$$\begin{matrix} 23 \\ 1 \\ 4 \end{matrix}\begin{bmatrix} 0{,}5 & 2{,}5 & 5{,}5 \\ 2{,}5 & 0 & 4 \\ 5{,}5 & 4 & 0 \end{bmatrix}$$

3. Dipilih jarak klaster terdekat yaitu klaster (23) dan objek 1 yang digabungkan menjadi klaster (231), kemudian dilakukan perhitungan jarak antara klaster (231) dan objek 4

   - $d_{(231)4} = \frac{d_{24}+d_{34}+d_{14}}{3} = \frac{5+6+4}{3} = 5$
   - $d_{(231)(231)} = \frac{d_{22}+d_{23}+d_{21}+d_{32}+d_{33}+d_{31}+d_{12}+d_{13}+d_{11}}{9}$

     $d_{(231)(231)} = \frac{0+1+2+1+0+3+2+3+0}{9}$

     $d_{(231)(231)} = 1{,}333$

   Diperoleh matriks jarak baru

$$\begin{matrix} 231 \\ 4 \end{matrix}\begin{bmatrix} 1{,}333 & 5 \\ 5 & 0 \end{bmatrix}$$

2.1.5.1.2 Metode Ward

Metode Ward atau yang biasa disebut metode jumlah kuadrat tambahan, menggunakan jarak kuadrat dalam satu klaster dan jarak kuadrat antar klaster. Menurut Johnson dan Wichern (2007), metode Ward mempertimbangkan pengelompokan secara hirarki berdasarkan meminimalkan informasi yang hilang

dalam menggabungkan dua grup. Metode Ward didasarkan pada kriteria *sum of square error* (SSE) dengan ukuran kehomogenan antara dua objek berdasarkan jumlah kuadrat kesalahan yang paling minimal, SSE hanya dapat dihitung jika klaster memiliki elemen lebih dari satu objek. Formula untuk menghitung SSE adalah sebagai berikut:

$$SSE = \sum_{i=1}^{N} (\boldsymbol{y_i} - \bar{\boldsymbol{y}})'(\boldsymbol{y_i} - \bar{\boldsymbol{y}}) \tag{2.6}$$

Keterangan:

y_i : vektor kolom yang entrinya nilai rata-rata objek i

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

$\bar{y}$: vektor kolom yang entrinya rata-rata nilai objek dalam klaster

N : banyaknya objek

Jika *A B* adalah klaster yang diperoleh dengan menggabungkan klaster A dan B, maka jumlah jarak antar klaster adalah sebagai berikut:

$$SSE_A = \sum_{i=1}^{n_A} (\boldsymbol{y_i} - \bar{\boldsymbol{y}}_{\boldsymbol{A}})'(\boldsymbol{y_i} - \bar{\boldsymbol{y}}_{\boldsymbol{A}}) \tag{2.7}$$

$$SSE_B = \sum_{i=1}^{n_B} (\boldsymbol{y_i} - \bar{\boldsymbol{y}}_{\boldsymbol{B}})'(\boldsymbol{y_i} - \bar{\boldsymbol{y}}_{\boldsymbol{B}}) \tag{2.8}$$

$$SSE_{AB} = \sum_{i=1}^{n_{AB}} (\boldsymbol{y_i} - \bar{\boldsymbol{y}}_{\boldsymbol{AB}})'(\boldsymbol{y_i} - \bar{\boldsymbol{y}}_{\boldsymbol{AB}}) \tag{2.9}$$

Dengan SSE_A, SSE_B dan SSE_{AB} adalah jumlah kesalahan klaster A, B dan AB. $\bar{y}_A$, $\bar{y}_B$ dan $\bar{y}_{AB}$ adalah vektor kolom yang entrinya rata-rata nilai objek dari klaster A, B dan AB. n_A, n_B dan n_{AB} adalah banyaknya klaster pada objek A, B dan AB.

Metode Ward menggabungkan dua klaster A dan B dengan meminimalkan peningkatan SSE, didefinisikan sebagai jarak antara klaster A dan klaster B, yaitu sebagai berikut:

$$I_{AB} = SSE_{AB} - (SSE_A + SSE_B) \tag{2.10}$$

Dapat ditunjukan bahwa peningkatan I_{AB} pada (2.10) memiliki bentuk dua persamaan ekuivalen sebagai berikut:

$$I_{AB} = n_A(\bar{\boldsymbol{y}}_{\boldsymbol{A}} - \bar{\boldsymbol{y}}_{\boldsymbol{AB}})'(\bar{\boldsymbol{y}}_{\boldsymbol{A}} - \bar{\boldsymbol{y}}_{\boldsymbol{AB}}) + n_B(\bar{\boldsymbol{y}}_{\boldsymbol{B}} - \bar{\boldsymbol{y}}_{\boldsymbol{AB}})'(\bar{\boldsymbol{y}}_{\boldsymbol{B}} - \bar{\boldsymbol{y}}_{\boldsymbol{AB}}) \tag{2.11}$$

$$I_{AB} = \frac{n_A n_B}{n_A + n_B}(\bar{\boldsymbol{y}}_{\boldsymbol{A}} - \bar{\boldsymbol{y}}_{\boldsymbol{B}})'(\bar{\boldsymbol{y}}_{\boldsymbol{A}} - \bar{\boldsymbol{y}}_{\boldsymbol{B}}) \tag{2.12}$$

Dari persamaan (2.12), meminimalkan peningkatan SSE adalah ekuivalen untuk meminimalkan jarak antar objek. Jika A hanya terdiri atas y_i dan B hanya terdiri atas y_j, maka SSE_A dan SSE_B bernilai nol dan (2.10) dan (2.12) dapat direduksi menjadi

$$I_{ij} = SSE_{AB} = \frac{1}{2}(\boldsymbol{y_i} - \boldsymbol{y_j})'(\boldsymbol{y_i} - \boldsymbol{y_j}) \tag{2.13}$$

$$I_{ij} = SSE_{AB} = \frac{1}{2}d^2(\boldsymbol{y_i}, \boldsymbol{y_j}) \tag{2.14}$$

Dengan I_{ij} adalah jarak antar objek i dan objek j, y_i adalah nilai rata-rata objek i dan y_j adalah nilai rata-rata objek j. $d^2(y_i, y_j)$ adalah jarak Euclid kuadrat antar objek i dan objek j.

Menurut Rencher (2002), jarak antara objek AB dan objek C dengan metode Ward yaitu sebagai berikut:

$$I_{(AB)C} = \frac{n_A + n_C}{n_{AB} + n_C} I_{AC} + \frac{n_B + n_C}{n_{AB} + n_C} I_{BC} - \frac{n_C}{n_{AB} + n_C} I_{AB} \tag{2.15}$$

Dengan $I_{(AB)C}$ adalah jarak antara klaster AB dan klaster C, I_{AC} adalah jarak antara klaster A dan klaster C, I_{BC} adalah jarak antara klaster B dan klaster C, I_{AB} adalah jarak antara klaster A dan klaster B dan $n_A, n_{B,} n_C$ adalah banyaknya objek pada klaster ke-A, ke-B dan ke-C.

2.2 Indeks Dunn (*Dunn Index*)

Indeks Dunn adalah salah satu pengukuran validitas klaster yang diajukan oleh J.C. Dunn. Ukuran validitas klaster ini berlandaskan pada fakta bahwa klaster yang terpisah itu biasanya memiliki jarak antar klaster yang besar dan diameter intra klaster yang kecil (Satoto, Khotimah dan Muhammad, 2015).

Indeks Dunn adalah rasio atau pembagian dari jarak terkecil antar objek pada klaster yang berbeda terhadap jarak intra klaster terbesar (Brock dkk., 2008).

Indeks Dunn dapat dituliskan sebagai berikut:

$$D = \min_{j=i+1,\ldots,n_c} \left(\frac{d(c_i, c_j)}{\max_{k=1,\ldots,n_c}(diam(c_k))} \right) \tag{2.16}$$

Dimana nilai $d(c_i, c_j)$ dan $diam(c_k)$ didefinisikan sebagai berikut:

$$d(c_i, c_j) = \min_{y \in c_j}^{x \in c_i}(d(x, y)) \tag{2.17}$$

$$diam(c_k) = \max_{y \in c_i}^{x \in c_i}(d(x, y)) \tag{2.18}$$

Nilai pada indeks Dunn ini jika nilainya semakin besar, maka hasil klaster akan semakin baik. Indeks Dunn memiliki rentang nilai dari nol sampai tak hingga.

2.3 Indeks RS (*R-Squared*)

Menurut Sharma (1996), indeks RS dapat didefinisikan sebagai berikut:

$$RS = \frac{SS_B}{SS_T} = \frac{SS_T - SS_W}{SS_T} = \frac{\left\{\sum_{j=1}^{n}(x_j - \bar{x})^2\right\} - \left\{\sum_{i=1}^{n_c}\sum_{j=1}^{r_i}(x_{ij} - \bar{x})^2\right\}}{\left\{\sum_{j=1}^{n}(x_j - \bar{x})^2\right\}} \tag{2.19}$$

Keterangan:

x_j : data ke – j pada variabel

x_{ij} : data ke – j pada variabel pada masing-masing klaster ke – i

Semakin besar nilai RS maka klaster yang dihasilkan akan semakin baik. RS memiliki rentang nilai dari nol sampai satu.

2.4 Analisis Komponen Utama (AKU)

Menurut Johnson dan Wichern (2007), analisis komponen utama (AKU), merupakan analisis tertua dalam APG yang diperkenalkan oleh Karl Pearson tahun 1901, yang biasanya digunakan untuk:

1. Identifikasi peubah baru yang mendasari data peubah ganda
2. Mereduksi jumlah himpunan peubah yang banyak dan saling berkorelasi menjadi peubah-peubah baru yang tidak berkorelasi dengan mempertahankan sebanyak mungkin keragaman data tersebut
3. Menghilangkan peubah-peubah asal yang tidak memberi informasi yang penting.

Peubah baru yang terbentuk adalah peubah yang merupakan kombinasi linear dari peubah asal, jumlah kuadrat koefisien dalam kombinasi linear tersebut bernilai satu, dan tidak saling berkorelasi dan ragamnya terurut dari yang terbesar ke yang terkecil. Sebelum mencari KU dilakukan penguraian pada peubah-peubah asal dengan penguraian nilai singular (PNS), yaitu:

$$\boldsymbol{X_{n\times p}} = \boldsymbol{U_{n\times r}}\, \boldsymbol{L_{r\times r}}\, \boldsymbol{A'_{r\times p}} \quad ; rank(X) = r \tag{2.20}$$

Matriks $\boldsymbol{L}$ adalah matriks diagonal yang unsur-unsur diagonalnya merupakan akar kuadrat dari akar ciri $\boldsymbol{X'X}$ dan $\boldsymbol{A}$ merupakan matriks yang kolom-kolomnya adalah vektor ciri dari $\boldsymbol{X'X}$ yang berpadanan dengan akar ciri λ dimana $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \ldots \geq \lambda_r > 0$.

Matriks $\boldsymbol{A}$ dan $\boldsymbol{U}$ merupakan matriks dengan kolom ortonormal dimana $\boldsymbol{U'U} = \boldsymbol{A'A} = I$ sedangkan matriks

$$\boldsymbol{U} = \left[\frac{Xa_1}{\sqrt{\lambda_1}}, \frac{Xa_2}{\sqrt{\lambda_2}}, \frac{Xa_3}{\sqrt{\lambda_3}}, \dots, \frac{Xa_r}{\sqrt{\lambda_r}}\right] \tag{2.21}$$

dimana $\boldsymbol{a_i}$ merupakan vektor ciri yang berkaitan dengan λ_i.

Perhatikan kombinasi linear berikut:

$Y_1 = \boldsymbol{a_1'X} = a_{11}X_1 + a_{12}X_2 + a_{13}X_3 + \ \dots + a_{1p}X_p$

$Y_2 = \boldsymbol{a_2'X} = a_{21}X_1 + a_{22}X_2 + a_{23}X_3 + \ \dots + a_{2p}X_p$

$Y_3 = \boldsymbol{a_3'X} = a_{31}X_1 + a_{32}X_2 + a_{33}X_3 + \ \dots + a_{3p}X_p$

$\vdots$

$Y_p = \boldsymbol{a_p'X} = a_{p1}X_1 + a_{p2}X_2 + a_{p3}X_3 + \ \dots + a_{pp}X_p$

dan

$$Var(Y_i) = \boldsymbol{a_i'\,\Sigma\,a_i} \quad ; i = 1,2,3, \dots, p \tag{2.22}$$

maka

$$Cov(Y_i, Y_k) = \boldsymbol{a_i'\,\Sigma\,a_k} \quad ; i,k = 1,2,3, \dots, p \tag{2.23}$$

Dimana $\boldsymbol{\Sigma}$ adalah matriks kovarians atau dapat diganti dengan matriks korelasi ρ dan $\boldsymbol{a_i}$ merupakan vektor ciri yang berkaitan dengan λ_i. Komponen utama yang terbentuk adalah kombinasi linear yang tidak saling berkorelasi $Y_1, Y_2, Y_3, \dots, Y_p$ yang variansinya pada (2.22) sebesar mungkin. Secara umum, komponen utama ke-*i* adalah kombinasi linear $\boldsymbol{a_i'\,X}$ yang memaksimumkan $Var(\boldsymbol{a_i'X})$ terhadap kendala $\boldsymbol{a_i'\,a_i} = \boldsymbol{1}$ dan $Cov(\boldsymbol{a_i'X}, \boldsymbol{a_k'X}) = \boldsymbol{0}$ untuk $k < i$ yang berarti

- Komponen utama 1 adalah kombinasi linear $\boldsymbol{a_1'\,X}$ yang memaksimumkan $Var(\boldsymbol{a_1'X})$ terhadap kendala $\boldsymbol{a_1'\,a_1 = 1}$
- Komponen utama 2 adalah kombinasi linear $\boldsymbol{a_2'\,X}$ yang memaksimumkan $Var(\boldsymbol{a_2'X})$ terhadap kendala $\boldsymbol{a_2'\,a_2 = 1}$ dan $Cov(\boldsymbol{a_1'X, a_2'X}) = \boldsymbol{0}$
- Komponen utama i adalah kombinasi linear $\boldsymbol{a_i'\,X}$ yang memaksimumkan $Var(\boldsymbol{a_i'X})$ terhadap kendala $\boldsymbol{a_i'\,a_i = 1}$ dan $Cov(\boldsymbol{a_i'X, a_k'X}) = \boldsymbol{0}$ untuk $k < i$

(Johnson dan Wichern, 2007).

Adapun langkah-langkah yang dilakukan dalam analisis komponen utama, yaitu:

1. Menghitung matriks kovarian ($\boldsymbol{\Sigma}$)
2. Menghitung Nilai Eigen (λ)
3. Hitung vektor eigen ($\boldsymbol{a}$)
4. Hitung komponen utama ($\boldsymbol{a_i'X}$)

2.4.1 Matriks Kovarian ($\boldsymbol{S}$)

Matriks kovarian ($\boldsymbol{S}$) berlaku untuk sampel dapat ditulis sebagai berikut:

$$\mathbf{S} = \begin{bmatrix} s_{X_1X_1} & s_{X_1X_2} & \cdots & s_{X_1X_p} \\ s_{X_2X_1} & s_{X_2X_2} & \cdots & s_{X_2X_p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{X_pX_1} & s_{X_pX_2} & \cdots & s_{X_pX_p} \end{bmatrix}$$

$s_{X_2X_1} = s_{X_1X_2}$

Secara teori kovarian didefinisikan sebagai berikut:

$$s_{XY} = Cov(X, Y) = E(XY) - E(X)E(Y) \tag{2.24}$$

$$s_{XY} = Cov(X,Y) = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{N-1} \quad (2.25)$$

Dengan X dan Y adalah variabel acak (Hardle dan Simar, 2007).

Berdasarkan persamaan (2.25) diatas $s_{X_1X_1}$, $s_{X_2X_1}$, $s_{X_pX_1}$, $s_{X_pX_p}$dapat dijabarkan sebagai berikut:

- $s_{X_1X_1} = \frac{\sum_{i=1}^{N}(X_{1i}-\overline{X_1})^2}{N-1}$
- $s_{X_2X_1} = \frac{\sum_{i=1}^{N}(X_{2i}-\overline{X_2})(X_{1i}-\overline{X_1})}{N-1}$
- $s_{X_pX_1} = \frac{\sum_{i=1}^{N}(X_{pi}-\overline{X_p})(X_{1i}-\overline{X_1})}{N-1}$
- $s_{X_pX_2} = \frac{\sum_{i=1}^{N}(X_{pi}-\overline{X_p})(X_{2i}-\overline{X_2})}{N-1}$
- $s_{X_pX_p} = \frac{\sum_{i=1}^{N}(X_{pi}-\overline{X_p})^2}{N-1}$

2.4.2 Nilai Eigen (λ)

Semua matriks persegi $\boldsymbol{X}$, nilai eigen λ dan vektor eigen ($\boldsymbol{a}$) dapat ditemukan sedemikian rupa

$$\boldsymbol{Xa} = \lambda \boldsymbol{a} \quad (2.26)$$

Untuk mendapatkan nilai λ dan $\boldsymbol{x}$ dapat ditulis sebagai berikut:

$$(\boldsymbol{X} - \lambda \boldsymbol{I})\boldsymbol{a} = \boldsymbol{0} \quad (2.27)$$

Berdasarkan persamaan (2.27) didapat persamaan baru untuk mendapatkan nilai λ yaitu:

$$|\boldsymbol{X} - \lambda \boldsymbol{I}| = 0 \quad (2.28)$$

Persamaan (2.28) dapat disebut persamaan karakteristik, jika $\boldsymbol{X}$ berordo $n\,x\,n$ maka akan didapat n akar yang berarti terdapat n nilai eigen yaitu $\lambda_1, \lambda_2, \dots, \lambda_n$ (Rencher, 2002).

2.4.3 Vektor Eigen ($\boldsymbol{a}$)

Vektor eigen dapat dihitung dengan menggunakan persamaan (2.27).

Contoh 2:

Diketahui matriks kovarian $\boldsymbol{S}$, akan dihitung nilai eigen dan vektor eigen

$$\mathbf{S} = \begin{bmatrix} 1 & -2 \\ 1 & 4 \end{bmatrix}$$

- Menghitung Nilai eigen (λ)

  $|\mathbf{S} - \lambda \boldsymbol{I}| = 0$

  $\left|\begin{bmatrix} 1 & -2 \\ 1 & 4 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}\right| = 0$

  $\left|\begin{bmatrix} 1-\lambda & -2 \\ 1 & 4-\lambda \end{bmatrix}\right| = 0$

  $\lambda^2 - 5\lambda + 6 = 0$

  $(\lambda - 2)(\lambda - 3) = 0$

  Didapat $\lambda_1 = 2$ dan $\lambda_2 = 3$

- Menghitung Vektor Eigen ($\boldsymbol{a}$)

  - Untuk $\lambda_1 = 2$

    $(\mathbf{S} - \lambda_1 \boldsymbol{I})\boldsymbol{a} = \mathbf{0}$

    $\left(\begin{bmatrix} 1 & -2 \\ 1 & 4 \end{bmatrix} - \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}\right)\begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = 0$

$$\left(\begin{bmatrix}-1 & -2\\ 1 & 2\end{bmatrix}\right)\begin{bmatrix}a_1\\ a_2\end{bmatrix}=0$$

$$-a_1-2a_2=0$$

$$a_1+2a_2=0$$

Maka nilai $a_1=-2$ dan $a_2=1$ atau dapat ditulis

$$\boldsymbol{a}=\begin{bmatrix}-2\\ 1\end{bmatrix}$$

- Untuk $\lambda_2=3$

$$(\mathbf{S}-\lambda_2\boldsymbol{I})\boldsymbol{a}=\mathbf{0}$$

$$\left(\begin{bmatrix}1 & -2\\ 1 & 4\end{bmatrix}-\begin{bmatrix}3 & 0\\ 0 & 3\end{bmatrix}\right)\begin{bmatrix}a_1\\ a_2\end{bmatrix}=0$$

$$\left(\begin{bmatrix}-2 & -2\\ 1 & 1\end{bmatrix}\right)\begin{bmatrix}a_1\\ a_2\end{bmatrix}=0$$

$$-2a_1-2a_2=0$$

$$a_1+a_2=0$$

Maka nilai $a_1=1$ dan $a_2=-1$ atau dapat ditulis

$$\boldsymbol{a}=\begin{bmatrix}1\\ -1\end{bmatrix}$$

2.5 Uji Normal Multivariat Henze-Zirkler

Menurut Korkmaz dkk. (2014), uji Henze-Zirkler berdasarkan pada jarak fungsional non negatif yang mengukur jarak antara dua fungsi distribusi. Statistik uji dati Henze-Zirkler normal multivariat memiliki persamaan sebagai berikut:

$$HZ=\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}e^{-\frac{\beta^2}{2}D_{ij}}-2\left(1+\beta^2\right)^{-\frac{p}{2}}\sum_{i=1}^{n}e^{-\frac{\beta^2}{2(1+\beta^2)}D_i}+n(1+2\beta^2)^{-\frac{p}{2}} \qquad (2.29)$$

Keterangan:

p = Jumlah variabel atau dimensi data

β = $\frac{1}{\sqrt{2}}\left(\frac{n(2p+1)}{4}\right)^{\frac{1}{p+4}}$

D_{ij} = $(x_i - x_j)'\boldsymbol{S^{-1}}(x_i - x_j)$

D_i = $(x_i - \bar{x})'\boldsymbol{S^{-1}}(x_i - \bar{x}) = m_{ii}$

D_{ij} adalah jarak Mahalanobis antara objek i dan j dan D_i adalah jarak Mahalanobis kuadrat pada objek i.

III. METODOLOGI PENELITIAN

3.1 Waktu dan Tempat Penelitian

Penelitian ini dilaksanakan pada semester ganjil tahun ajaran 2019/2020 bertempat di Jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung.

3.2 Data Penelitian

Data yang digunakan dalam penelitian ini adalah hasil dari membangkitkan data yang mengandung masalah multikolinearitas dengan menggunakan *software* RStudio versi 1.2.1335 dan menggunakan beberapa *package* yang disediakan oleh *software* RStudio.

Menurut Kibria dan Muniz (2009), untuk mendapatkan data yang mengandung multikolinearitas pada setiap himpunan data X_{ij} dibangkitkan menggunakan simulasi Monte Carlo dengan persamaan sebagai berikut:

$$X_{ij} = \sqrt{(1-\rho^2)}\, x_{ij} + \rho x_{ip} \tag{3.1}$$

Keterangan:

ρ : korelasi antar dua variabel yang ditentukan

x_{ij} : data yang dibangkitkan berdistribusi normal dengan μ dan σ ditentukan

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

j : $1, 2, 3, \dots, p$

3.3 Metode Penelitian

Penelitian ini dilakukan secara studi pustaka yaitu mempelajari buku-buku teks, jurnal serta akses internet yang menunjang proses penelitian. Adapun langkah-langkah penelitian yang dilakukan sebagai berikut:

1. Membangkitkan data $\boldsymbol{X_i} \sim N_p(\boldsymbol{\mu_i}, \boldsymbol{\Sigma_i})$, $\boldsymbol{X_i} = [X_1, X_2, X_3, \dots, X_p]^t$ adalah vektor pengamatan dengan $i = 1, 2, 3, \dots$ dibangkitkan secara acak yang kemudian dikonversi menjadi data multikolinearitas dengan ketentuan:

    a. Data ke-1 dibangkitkan dengan $n = 10$ yang membentuk dua klaster. Klaster pertama dengan $n = 5$ berdistribusi $X_j \sim N(0,1)$ sehingga $\boldsymbol{X_1} \sim N_4(\boldsymbol{\mu_1}, \boldsymbol{\Sigma_1})$ dengan $\boldsymbol{\mu_1} = [0,0,0,0]^t$ dan $\boldsymbol{\Sigma_1} = 1\mathbf{I_4}$ dimana $j = 1, 2, 3, 4$. Klaster kedua dengan $n = 5$ objek berdistribusi $X_j \sim N(5,2)$ sehingga $\boldsymbol{X_2} \sim N_4(\boldsymbol{\mu_2}, \boldsymbol{\Sigma_2})$ dengan $\boldsymbol{\mu_2} = [5,5,5,5]^t$ dan $\boldsymbol{\Sigma_2} = 2\mathbf{I_4}$ dimana $j = 1, 2, 3, 4$ yang kemudian gabungan data $\boldsymbol{X_1}$ dan $\boldsymbol{X_2}$ dikonversi menjadi data multikolinearitas dengan menggunakan persamaan (3.1)

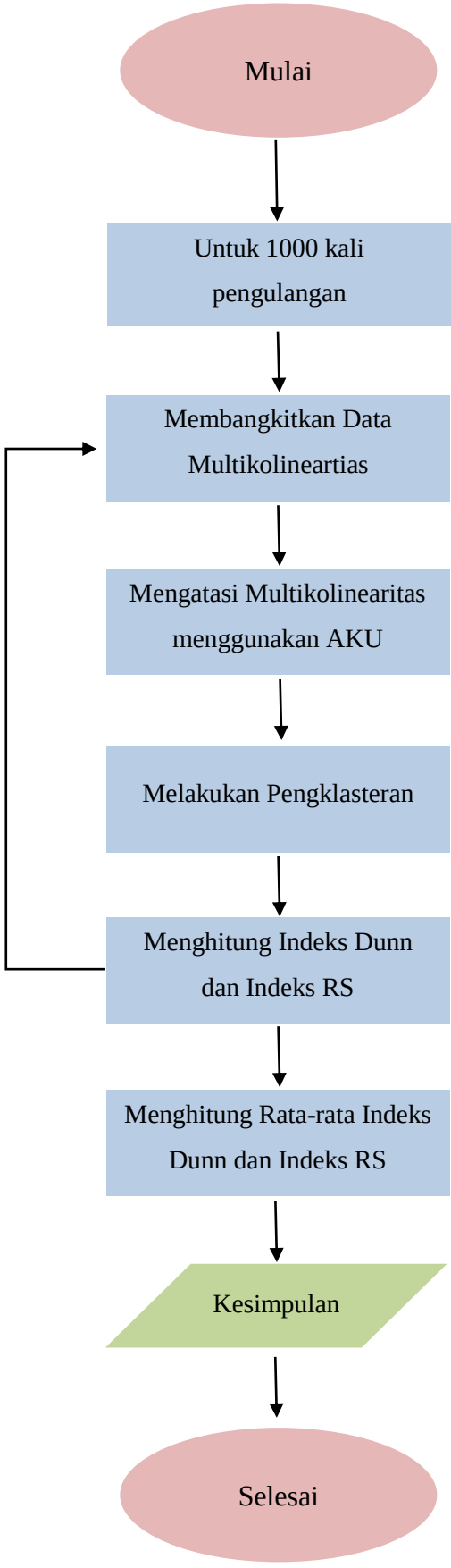
b. Data ke-2 dibangkitkan dengan $n = 20$ yang membentuk tiga klaster. Klaster pertama dengan $n = 7$, berdistribusi sama dengan klaster pertama pada poin a. Klaster kedua dengan $n = 7$, berdistribusi sama dengan klaster kedua pada poin a. Klaster ketiga dengan $n = 6$ berdistribusi $X_j \sim N(8,3)$ sehingga $\boldsymbol{X_3} \sim N_4(\boldsymbol{\mu_3}, \boldsymbol{\Sigma_3})$ dengan $\boldsymbol{\mu_3} = [8,8,8,8]^t$ dan $\boldsymbol{\Sigma_3} = 3\mathbf{I_4}$ dimana $j = 1, 2, 3, 4$ yang kemudian gabungan data $\boldsymbol{X_1}$, $\boldsymbol{X_2}$ dan $\boldsymbol{X_3}$ dikonversi menjadi data multikolinearitas dengan menggunakan persamaan (3.1)

c. Data ke-3 dibangkitkan dengan $n = 50$ yang membentuk empat klaster. Klaster pertama dengan $n = 12$, berdistribusi sama dengan klaster pertama pada poin a. Klaster kedua dengan $n = 12$, berdistribusi sama dengan klaster kedua pada poin a. Klaster ketiga dengan $n = 13$, berdistribusi sama dengan klaster ketiga pada poin b. Klaster keempat dengan $n = 13$ berdistribusi $X_j \sim N(10,4)$ sehingga $\boldsymbol{X_4} \sim N_4(\boldsymbol{\mu_4}, \boldsymbol{\Sigma_4})$ dengan $\boldsymbol{\mu_4} = [10,10,10,10]^t$dan $\boldsymbol{\Sigma_4} = 4\mathbf{I_4}$ dimana $j = 1, 2, 3, 4$ yang kemudian gabungan data $\boldsymbol{X_1}$, $\boldsymbol{X_2}$, $\boldsymbol{X_3}$ dan $\boldsymbol{X_4}$ dikonversi menjadi data multikolinearitas dengan menggunakan persamaan (3.1)

d. Data ke-4 dibangkitkan dengan $n = 100$ yang membentuk lima klaster. Klaster pertama dengan $n = 20$, berdistribusi sama dengan klaster pertama pada poin a. Klaster kedua dengan $n = 20$, berdistribusi sama dengan klaster kedua pada poin a. Klaster ketiga dengan $n = 20$, berdistribusi sama dengan klaster ketiga pada poin b. Klaster keempat dengan $n = 20$, berdistribusi sama dengan klaster keempat pada poin c. Klaster kelima dengan $n = 20$ berdistribusi $X_j \sim N(13,3)$ sehingga $\boldsymbol{X_5} \sim N_4(\boldsymbol{\mu_5}, \boldsymbol{\Sigma_5})$ dengan $\boldsymbol{\mu_5} = [13,13,13,13]^t$dan $\boldsymbol{\Sigma_5} = 3\mathbf{I_4}$ dimana $j = 1, 2, 3, 4$ yang kemudian gabungan data $\boldsymbol{X_1}$, $\boldsymbol{X_2}$, $\boldsymbol{X_3}$, $\boldsymbol{X_4}$ dan $\boldsymbol{X_5}$ dikonversi menjadi data multikolinearitas dengan menggunakan persamaan (3.1)

2. Standarisasi data kedalam bentuk nilai Z
3. Melakukan uji asumsi multikolinearitas (VIF)
4. Mengatasi data yang mengandung multikolinearitas menggunakan analisis komponen utama (AKU)
5. Melakukan pengklasteran dengan menggunakan metode *average linkage* dan metode Ward dengan menggunakan data hasil analisis komponen utama
6. Menghitung dan mencatat indeks Dunn pada tiap metode
7. Menghitung dan mencatat indeks RS pada tiap metode
8. Mengulang langkah 1 (satu) sampai langkah 7 (tujuh) sebanyak 1000 (seribu) kali pengulangan

9. Melakukan evaluasi indeks Dunn dan indeks RS dengan menghitung rata-ratanya
10. Analisis hasil



Gambar 2. Algoritma (*flowchart*) penelitian

V. KESIMPULAN

5.1 Kesimpulan

Berdasarkan hasil dan pembahasan yang telah dijelaskan pada Bab IV, maka dapat diambil kesimpulan sebagai berikut:

1. Analisis klaster metode *average linkage* dan Ward pada data yang mengandung multikolinearitas dapat diatasi dengan analisis komponen utama

2. Berdasarkan nilai indeks Dunn, metode *average linkage* memberikan hasil yang lebih baik dibandingkan metode Ward dalam pengklasteran data. Ukuran indeks Dunn berlandaskan pada fakta bahwa klaster yang terpisah itu biasanya memiliki jarak antar klaster yang besar dan diameter intra klaster yang kecil, yang berarti klaster-klaster yang dibentuk oleh metode *average linkage* memiliki jarak antar klaster yang lebih besar, diameter intra klaster yang lebih kecil dan lebih kompak dibandingkan metode Ward

3. Berdasarkan nilai indeks RS, metode Ward memberikan hasil yang lebih baik dibandingkan metode *average linkage* dalam pengklasteran data. Indeks RS mengukur apakah karakteristik antar klaster saling berbeda, yang berarti klaster yang terbentuk dengan menggunakan metode Ward memiliki karakteristik yang lebih berbeda dibanding dengan metode *average linkage.*

5.2 Saran

Pada penelitian ini digunakan indeks Dunn yang senstitif terhadap data outlier sehingga jika pada klaster terdapat data outlier, maka akan ada kemungkinan terjadinya kesalahan dalam menyimpulkan metode klaster mana yang paling baik dalam mengelompokan objek-objeknya (Yatskiv dan Gusarova, 2005). Berdasarkan hal ini, untuk penelitian selanjutnya dapat menggunakan indeks lain seperti *connectivity* dan *silhoutte width* sebagai alat ukur validitas analisis klaster *agglomerative* metode *average linkage* dan metode Ward. Selain itu, pada penelitian ini digunakan jumlah objek (*n*) sebanyak 10, 20, 50, 100, dan jumlah klaster sebanyak 2, 3, 4, dan 5, oleh karena itu untuk penelitian selanjutnya dapat menggunakan ukuran sampel dan jumlah klaster yang lebih bervariasi.

DAFTAR PUSTAKA

Brock,G., Datta, S., Datta, S. dan Pihur, V. 2008. clValid: An R Package for Cluster Validation. *Journal of Statistical Software*. **25**: 4.

Hair, J.F., Black, W.C., Babin, B.J. dan Anderson, R.E. 2014. *Multivariate Data Analysis*. 7th Edition. Pearson Education Limited, England.

Hardle, W. dan Simar, L. 2007. *Applied Multivariate Statistical Analysis*. 2nd Edition. Springer Berlin Heidelberg, New York.

Johnson, S.C. 1967. *Hierarchical Clustering Schemes*. Prentice-Hall, Inc., New Jersey.

Johnson, R. dan Wichern, D. 2007. *Applied Multivariate Analysis*. 6th Edition. Prentice Hall Inc., New Jersey.

Kibria, B. dan Muniz, G. 2009. On Some Ridge Regression Estimator: An Emprical Comparison. *Communication in Statistics – Simulation and Computation*. **38**: 621-630.

Korkmaz, S., Goksuluk, D. dan Zararsiz, G. 2014. MVN: An R Package for Assessing Multivariate Normality. *The R Journal*. **6**: 2.

Lind, D.A., Marchal, W.G. dan Wathen, S.A. 2007. *Teknik-teknik Statistika dalam Bisnis dan Ekonomi Menggunakan Kelompok Data, Edisi 13*. Diterjemahkan oleh Chriswan Sungkono. Salemba Empat, Jakarta.

Rencher, A.C. 2002. *Methods of Multivariate Analysis.* 2nd Edition. A John Wiley & Sons, Inc. Publication, Canada.

Rujasiri, P. dan Chomtee, B. 2009. Comparison of Clustering Techniques for Cluster Analysis. *Kasetsart Journal (Natural Science)*. **43**: 378-388.

Satoto, B.D., Khotimah, B.K. dan Muhammad, A. 2015. Pengelompokan Tingkat Kesehatan Masyarakat Menggunakan *Shelf Organizing Maps* dengan *Cluster Validation Idb* dan *I-Dunn*. *Seminar Nasional Aplikasi Teknologi Informasi (SNATi) 2015*.

Sharma, S. 1996. *Applied Multivariate Techniques*. A John Wiley & Sons, Inc., Canada.

Supranto. 2004. *Analisis Multivariat: Arti dan Interpretasi*. Rineka Cipta, Jakarta.

Usman, H. dan Nurdin, S. 2013. *Aplikasi Teknik Multivariate untuk Riset Pemasaran*. PT. Raja Grafindo Persada, Jakarta.

Widarjono, A. 2010. *Analisis Statistika Multivariat Terapan*. UPP STIM YKPN, Yogyakarta.

Yatskiv, I. dan Gusarova, L. 2005. The Methods of Cluster Analysis Result Validation. *Proceedings of International Conference RelStat'04*. **6**: 77.