PERBANDINGAN KINERJA ALGORITMA *WARD* DAN ALGORITMA *K-MEANS* DENGAN UJI *SILHOUETTE COEFFICIENT*

SKRIPSI

NURUL AFDHALIAH

H121 16 519

DEPARTEMEN STATISTIKA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS HASANUDDIN

DESEMBER 2020

PERBANDINGAN KINERJA ALGORITMA *WARD* DAN ALGORITMA *K-MEANS* DENGAN UJI *SILHOUETTE COEFFICIENT*

SKRIPSI

Diajukan sebagai salah satu syarat untuk memperoleh gelar Sarjana Sains pada Departemen Statistika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Hasanuddin Makassar

NURUL AFDHALIAH

H 121 16 519

PROGRAM STUDI STATISTIKA DEPARTEMEN STATISTIKA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS HASANUDDIN

MAKASSAR

DESEMBER 2020

LEMBAR PERNYATAAN KEASLIAN

Yang bertanda tangan dibawah ini :

Nama	: Nurul Afdhaliah
NIM	: H121 16 519
Program Studi	: Statistika
Jenjang	: S1

Menyatakan dengan ini bahwa karya tulisan saya berjudul :

PERBANDINGAN KINERJA ALGORITMA *WARD* DAN ALGORITMA *K-MEANS* DENGAN UJI *SILHOUETTE COEFFICIENT*

Adalah karya tulisan saya sendiri dan bukan merupakan pengambilan alihan tulisan orang lain bahwa skripsi yang saya tulis ini benar-benar merupakan hasil karya saya sendiri.

Apabila dikemudian hari terbukti atau dapat dibuktikan bahwa sebagian atau keseluruhan skripsi ini hasil karya orang lain, maka saya bersedia menerima sanksi atas perbuatan tersebut.

`

Makassar, 29 Desember 2020

Yang Menyatakan

Nurul Afdhaliah

NIM H121 16 519

PERBANDINGAN KINERJA ALGORITMA *WARD* DAN ALGORITMA *K-MEANS* DENGAN UJI *SILHOUETTE COEFFICIENT*

Disetujui Oleh:

Pembimbing Utama	Pembimbing Pertama
Dr. Dr. Georgina Maria Tinungki, M.Si.	**Dr. Nurtiti Sunusi, S.Si., M.Si.**
NIP 19620926 198702 2 001	**NIP 197201171997032002**

Ketua Departemen Statistika

Dr. Nurtiti Sunusi, S.Si., M.Si.

NIP 197201171997032002

Pada Tanggal : 29 Desember 2020

LEMBAR PENGESAHAN (TUGAS AKHIR)

PERBANDINGAN KINERJA ALGORITMA *WARD* DAN ALGORITMA *K-MEANS* DENGAN UJI *SILHOUETTE COEFFICIENT*

Disusun dan diajukan oleh

NURUL AFDHALIAH
H 121 16 519

Telah dipertahankan di hadapan Panitia Ujian yang dibentuk dalam rangka Penyelesaian Studi Program Sarjana Program Studi Statistika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Hasanuddin
pada tanggal 29 Desember 2020
dan dinyatakan telah memenuhi syarat kelulusan

Menyetujui,

Pembimbing Utama,	Pembimbing Pendamping,
Dr.Dr. Georgina Maria Tinungki, M.Si	**Dr. Nurtiti Sunusi, S.Si, M.Si**
NIP. 19620926 198702 2001	**NIP. 19720117 199703 2002**

Ketua Progam Studi,

Dr. Nurtiti Sunusi, S.Si, M.Si
NIP. 19720117 199703 2002

KATA PENGANTAR

Bismillahirrahmanirrahim.

Puji syukur kehadirat Allah SWT atas berkat, rahmat dan hidayah- Nya, sehingga penulis dapat menyelesaikan skripsi yang berjudul "**Perbandingan Kinerja Algoritma *Ward* dan Algoritma *K-Means* dengan Uji *Silhouette Coefficient***". Tak lupa pula penulis haturkan salam dan shalawat kepada junjungan kita Nabi Muhammad SAW. Skripsi ini diajukan untuk memenuhi sebagian persyaratan guna memperoleh gelar Sarjana pada Program Studi Statistika Departemen Statistika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Hasanuddin.

Penulis menyadari bahwa penyusunan skripsi ini bukanlah suatu hal yang mudah dan membutuhkan waktu yang singkat. Dalam proses penyusunan skripsi ini, penulis menemukan berbagai kendala dan tantangan. Namun kendala dan tantangan tersebut dapat teratasi berkat usaha keras, tekad yang kuat dan tentunya berkat dukungan tenaga, pikiran dan doa dari berbagai pihak.

Pada kesempatan ini, izinkan penulis mengucapkan terima kasih yang tak terhingga kepada orang tua tercinta, Ayahanda **Drs. H. Mappesabbi** dan Ibunda **Hj. Darmawati, S.Pd**, yang telah melahirkan, membesarkan dan mendidik penulis. Terima kasih telah memberikan kasih sayang yang tak tehingga serta memberikan segala bentuk dukungan yang luar biasa kepada penulis dalam setiap proses penyelesaian skripsi ini, baik itu dukungan moral, materi dan doa yang tiada hentinya selalu diberikan dengan ikhlas kepada penulis. Untuk saudaraku tercinta **Muhammad Awaluddin** yang senantiasa mendoakan, membantu dan memberikan dukungan serta semangat yang tiada hentinya kepada penulis selama ini. Terima kasih telah menjadi saudara sekaligus teman terbaik. Semoga Allah SWT selalu melindungi, memberikan kesehatan, rezeki yang baik dan keselamatan dunia akhirat kepada kedua orang tua dan seluruh keluarga besar penulis.

Penghargaan yang tulus dan ucapan terima kasih dengan penuh keikhlasan juga penulis ucapkan kepada:

1. **Ibu Prof. Dr. Dwia Aries Tina Palubuhu, MA,** selaku Rektor Universitas Hasanuddin beserta seluruh jajarannya;
2. **Bapak Dr. Eng. Amiruddin,** selaku Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Hasanuddin beserta seluruh stafnya;
3. **Ibu Dr. Nurtiti Sunusi, S.Si., M.Si.** selaku Ketua Departemen Statistika sekaligus dosen pembimbing pertama penulis, yang telah meluangkan waktunya untuk membimbing penulis sejak proposal hingga penyelesaian skripsi ini;
4. **Ibu Dr. Dr. Georgina Maria Tinungki, M.Si.** selaku dosen pembimbing utama yang telah meluangkan waktu untuk membimbing penulis dari awal proposal hingga penyelesaian skripsi ini, dan Penasehat Akademik (PA) atas segala masukan, bantuan, nasehat serta motivasi yang diberikan dari penulis masih berstatus Mahasiswa Baru sampai penulis menyelesaikan studi di Departemen Statistika;
5. **Ibu Dra. Nasrah Sirajang, M.Si.** selaku ketua tim penguji dan **Ibu Dr. Erna Tri Herdiani, S.Si., M.Si.** selaku sekretaris tim penguji atas semua saran dan kritikan yang membangun dalam penyempurnaan penyusunan tugas akhir ini serta waktu yang telah diberikan kepada penulis;
6. Seluruh dosen pengajar dan staf Departemen Statistika yang telah membekali ilmu dan kemudahan-kemudahan kepada penulis dalam berbagai hal selama menjadi mahasiswa di Departemen Statistika;
7. Saudari tak sedarahku yang setia nge-gas **Nur Alfianingsih Raja (Inci)**. Terima kasih telah hadir menjadi sahabat dan saudara yang baik, memberikan potongan cerita hidup yang sangat istimewa kepada penulis. Kita punya mimpi bersama, sampai bertemu di puncak kesuksesan itu. Terima kasih karena selalu peduli dan pengertian atas segala hal tentang penulis baik itu di kala suka maupun duka;
8. Support system : **Yyo', Murni, Rizal Cino, Hana, Andi Ria, Yuni,** Terima kasih telah hadir dan membersamai cerita hidup yang istimewa untuk penulis selama ini, selalu meluangkan waktu menjadi pendengar yang baik,

memberikan pengalaman berharga bagi penulis dan selalu mendukung serta memberi motivasi bagi penulis.

9. Terima kasih kepada sahabat-sahabat seperjuangan MINOR (**Ririn, Nidar, Andis, Imma, Mila, Aten, Tari, Cimma, Reza, Shasa,** dan **Wiya**), dan terkhusus **Azman,** Terima kasih selalu setia untuk mendengar dan berbagi kisah kehidupan perkuliahan dan menemani penulis selama mengerjakan skripsi, Terima kasih atas warna warni kehidupan selama ini.
10. Sobat ZEE (**Andi Fathin, Syahrul, Ida, Armin, Andi Hilal, Hilda)** juga **Isal, Ishar** dan **Kak Alling** Terima kasih untuk pertemanan dan persaudaraan hingga saat ini. Bertemu dengan kalian adalah salah satu hal paling indah dari semua rencana Tuhan dalam hidup penulis;
11. Terima kasih untuk saudara-saudari seperjuangan sejak menginjakkan kaki di Universitas Hasanuddin. Saudara saudariku "**A16ORITMA**" dan "**STATISTIKA 2016**". Terima kasih atas semua kisah yang telah kita lalui bersama dan semoga itu abadi dalam ingatan kita masing-masing.
12. Keluarga besar Himpunan Mahasiswa Matematika (**HIMATIKA**), Himpunan Mahasiswa Statistika (**HIMASTAT**). Terima kasih atas kebersamaannya, kesempatan berkarya dan kekeluargaan yang telah diberikan. Salam *Use your mind be the best, Bravo Himatika*!
13. Adik-adik paling pengertian **(Anita, Yaya, A. Rahmat, Ippang, Ikram, dan Syem,)** Terima kasih telah menjadi moodboster untuk penulis, selalu memberikan kerecehan dan kebahagian *unlimited*.
14. Terima kasih kepada teman-teman KKN Posko Desa Songing (**Mega, Putri, Dila, Nayyif, Kak Zulham, Kak Taufik, dan Kak Amin**), Terima kasih untuk cerita 1 Bulannya di Lokasi KKN, telah menjadi teman sekaligus keluarga dalam mengabdi.
15. Semua pihak yang telah membantu secara langsung maupun tidak langsung dalam penyusunan skripsi ini yang tak sempat penulis sebutkan satu per satu.

Akhirnya kepada Allah SWT penulis serahkan segalanya dengan memanjatkan doa yang tiada henti, rasa syukur yang teramat besar penulis haturkan kepada-Nya, atas segala izin dan limpahan berkah-Nya penulis dapat menyelesaikan Strata 1 di Universitas Hasanuddin tercinta. Semoga amal

kebaikan semua pihak yang telah membantu diterima disisi-Nya dan diberikan pahala yang berlipat ganda sesuai dengan amal perbuatannya. Penulis berharap semoga skripsi ini dapat bermanfaat bagi penulis khususnya serta bagi para pembaca pada umumnya. ***Amin Ya Rabbal 'Alamin***.

Terus berjalan dalam juang, panjang umur semangat baik.

Makassar, 29 Desember 2020

Nurul Afdhaliah

PERNYATAAN PERSETUJUAN PUBLIKASI

TUGAS AKHIR UNTUK KEPENTINGAN AKADEMIK

Sebagai civitas akademik Universitas Hasanuddin, saya yang bertanda tangan di bawah ini:

Nama	: Nurul Afdhaliah
NIM	: H 121 16 519
Program Studi	: Statistika
Departemen	: Statistika
Fakultas	: Matematika dan Ilmu Pengetahuan Alam
Jenis karya	: Skripsi

Demi pengembangan ilmu pengetahuan, menyetujui untuk memberikan kepada Universitas Hasanuddin **Hak Prediktor Royalti Non-eksklusif (*Non-exclusiive Royalty- Free Right*)** atas tugas akhir saya yang berjudul:

"Perbandingan Kinerja Algoritma *Ward* dan Algoritma *K-Means* dengan Uji *Silhouette Coefficient*"

Beserta perangkat yang ada (jika diperlukan). Terkait dengan hal di atas, maka pihak universitas berhak menyimpan, mengalih-media/format-kan, mengelola dalam bentuk pangkalan data (*database*), merawat, dan memublikasikan tugas akhir saya selama tetap mencantumkan nama saya sebagai penulis/pencipta dan sebagai pemilik Hak Cipta.

Demikian pernyataan ini saya buat dengan sebenarnya.

Dibuat di Makassar pada tanggal 29 Desember 2020

Yang menyatakan,

Nurul Afdhaliah

ABSTRAK

Analisis *cluster* merupakan suatu teknik analisis multivariat yang bertujuan untuk mengklasifikasikan data observasi ataupun variabel-variabel ke dalam *cluster* sedemikian rupa sehingga masing-masing *cluster* bersifat homogen. Prosedur analisis *cluster* dapat dilakukan dengan dua metode yaitu metode hirarki dan metode nonhirarki. Algoritma *Ward* merupakan salah satu algoritma dari metode hirarki yang bertujuan untuk memperoleh *cluster* yang memiliki varians internal yang sekecil mungkin. Algoritma *K-Means* merupakan salah satu algoritma dari metode non-hirarki dengan menentukan *center* (pusat *cluster*) secara acak dan melakukan pembaharuan *center* sampai kriteria pemberhentian tercapai. Tujuan dari penelitian ini adalah untuk mengelompokan kabupaten/kota menggunakan algoritma *Ward* dan algoritma *K-Means* berdasarkan Kasus Penyakit di sulawesi Selatan tahun 2017. Kemudian membandingkan kedua algoritma tersebut dengan uji *Silhouette Coefficient*. Dari kedua algoritma yang digunakan menghasilkan 2 *cluster* yang terbentuk. Dengan algoritma *Ward* diperoleh *cluster* 1 terdiri dari 22 kabupaten/kota dan *cluster* 2 terdiri dari 2 kabupaten/kota. Untuk algoritma *K-Means* diperoleh *cluster* 1 terdiri dari 20 kabupaten/kota dan *cluster* 2 terdiri dari 4 kabupaten/kota. Berdasarkan uji *silhouette coefficient*, algoritma *Ward* merupakan metode peng*cluster*an terbaik karena memiliki struktur data yang kuat. Dimana seluruh objek dalam *cluster* memilki nilai *silhouette* positif dan nilai rata-rata sebesar 0,69 mendekati angka 1.

Kata Kunci: Analisis *Cluster*, *Ward*, *K-Means*, *Silhouette Coefficient*, Kasus Penyakit.

ABSTRACT

Cluster analysis is a multivariate analysis technique that aims to classify observational data or variables into clusters in such a way that each cluster is homogeneous. The cluster analysis procedure can be carried out by two methods, namely the hierarchical method and the non-hierarchical method. Ward's algorithm is an algorithm of the hierarchical method that aims to obtain clusters that have the smallest possible internal variance. The K-Means algorithm is one of the algorithms of the non-hierarchical method by randomly determining the center (cluster center) and updating the center until the stopping criteria is reached. The purpose of this study is to classify districts / cities using the Ward algorithm and the K-Means algorithm based on disease cases in South Sulawesi in 2017. Then compare the two algorithms with the Silhouette Coefficient test. From the two algorithms used, 2 clusters are formed. With the Ward algorithm, it is obtained that cluster 1 consists of 22 districts / cities and cluster 2 consists of 2 districts / cities. For the K-Means algorithm, it is obtained that cluster 1 consists of 20 districts / cities and cluster 2 consists of 4 districts / cities. Based on the silhouette coefficient test, Ward's algorithm is the best clustering method because it has a strong data structure. Where all objects in the cluster have a positive silhouette value and an average value of 0.69 is close to number 1.

Keywords: Cluster Analysis, Ward, K-Means, Silhouette Coefficient, Desease Cases.

DAFTAR ISI

HALAMAN SAMPUL i
HALAMAN JUDUL i
LEMBAR PERNYATAAN KEOTENTIKAN..... **Error! Bookmark not defined.**
LEMBAR PENGESAHAN PEMBIMBING ii
HALAMAN PENGESAHAN..... **Error! Bookmark not defined.**
KATA PENGANTAR iv
PERNYATAAN PERSETUJUAN PUBLIKASI ix
ABSTRAK x
ABSTRACT xi
DAFTAR ISI xii
DAFTAR GAMBAR xiv
DAFTAR TABEL xv
DAFTAR LAMPIRAN xvi
BAB I PENDAHULUAN 1
1.1. Latar Belakang 1
1.2. Rumusan Masalah 4
1.3. Batasan Masalah 4
1.4. Tujuan Penelitian 4
1.5. Manfaat Penelitian 4
BAB II TINJAUAN PUSTAKA 5
2.1. Analisis Multivariat 5
2.2. Matriks Varian – Kovarian dan Matriks Korelasi 5
2.3. Analisis *Cluster* 7
2.4. Uji *Kaiser Mayer Olkin* 7
2.5. Uji Multikolinearitas 8
2.6. Analisis Komponen Utama 9
2.7. Mengukur Kemiripan atau Ketidakmiripan Antar Objek 10
2.8. Metode Hierarki 10
2.9. Algoritma *Ward* 11

2.10. Metode Non-Hierarki 12
2.11. Pemilihan Metode Terbaik Dengan *Silhoutte Coefficient* 14
2.12. Kesehatan 16
BAB III METODE PENELITIAN 18
3.1. Sumber Data 18
3.2. Deskripsi Variabel 18
3.3. Metode Analisis 18
BAB IV 20
HASIL DAN PEMBAHASAN 20
4.1. Statistika Deskriptif 20
4.2. Uji Kecukupan Sampel 21
4.3. Uji Multikolinearitas 22
4.4. Analisis Komponen Utama 23
4.5. Menghitung Kemiripan atau Ketidakmiripan Antar Objek dengan Jarak Kuadrat *Euclidean* 26
4.6. Penentuan Jumlah *Cluster* Optimum dengan Indeks Validitas *Silhouette* 27
4.7. Melakukan Proses Peng*cluster*an dengan Algoritma *Ward* dan *K-Means* 28
4.7.1 Langkah-langkah Peng*cluster*an Algoritma *Ward* 28
4.7.2 Langkah-langkah Peng*cluster*an Algoritma *K-Means* 31
4.7.3. Hasil Peng*cluster*an dan Anggotanya 34
4.8. Interpretasi *Cluster* 35
4.9. Pemilihan Algoritma Terbaik dengan Uji *Silhouette Coefficient* 37
BAB V 40
PENUTUP 40
5.1. Kesimpulan 40
5.2. Saran 41
DAFTAR PUSTAKA 42

DAFTAR GAMBAR

Gambar 4. 1 Plot *Silhouette* 27

Gambar 4. 2 Dendogram Algoritma *Ward*..... 30

Gambar 4. 3 Plot *Silhouette Coefficient* Untuk Algoritma *Ward*..... 38

Gambar 4. 4 Plot *Silhouette Coefficient* Untuk Algoritma *K-Means*..... 38

DAFTAR TABEL

Tabel 2. 1 Kriteria Pengukuran Uji KMO 8
Tabel 2. 2 Kriteria Pengukuran *Silhouette Coefficient* 15
Tabel 4. 1 Hasil Statistika Deskriptif Tiap Variabel (Jiwa) 20
Tabel 4. 2 Hasil Uji *KMO* 22
Tabel 4. 3 Proporsi Variansi Komulatif 24
Tabel 4. 4 Koefisien Komponen Utama 25
Tabel 4. 5 Nilai atau *Score* Komponen Utama 1 dan Komponen Utama 2 25
Tabel 4. 6 Hasil Jarak Kuadrat *Euclidean* Kabupaten/Kota 27
Tabel 4. 7 Nilai SSE Antar Dua Objek 29
Tabel 4. 8 Nilai *Centroid* pada Tahap Inisialisasi 31
Tabel 4. 9 Hasil Perhitungan Jarak Setiap Objek dengan *Centroid* 32
Tabel 4. 10 Anggota Setiap *Cluster* 33
Tabel 4. 11 Hasil Peng*cluster*an dengan Algortma *Ward* dan Anggotanya 34
Tabel 4. 12 Hasil Peng*cluster*an dengan Algoritma *K-means* dan Anggotanya ... 34
Tabel 4. 13 Rata-rata Variabel *(Centroid)* Setiap *Cluster* 35
Tabel 4. 14 Rata-rata Variabel *(Centroid)* Setiap *Cluster* 36

DAFTAR LAMPIRAN

Lampiran 1. Data Kasus Penyakit di Sulawesi Selatan Tahun 2017 45

Lampiran 2. Matriks Korelasi Antar Variabel 46

Lampiran 3. Matriks Varian Kovarian 47

Lampiran 4. Koefisien Komponen Utama 48

Lampiran 5. Nilai / *Score* Komponen Utama..... 49

Lampiran 6. Jarak Kuadrat *Euclidean* Antar Dua Objek Pengelompokan Kabupaten/Kota di Sulawesi Selatan 50

Lampiran 7. Output Penentuan Jumlah *Cluster* Optimum dengan *Silhouette* 54

Lampiran 8. Nilai SSE Antar Dua Objek Pengelompokan Kabupaten/Kota di Sulawesi Selatan..... 55

Lampiran 9. *Agglomeration Schedule* Algoritma *Ward* 59

Lampiran 10. Dendogram Algoritma *Ward* 60

Lampiran 11. Hasil Perhitungan Jarak Setiap Objek dengan *Centroid* 61

Lampiran 12. Nilai *Centroid* Baru 62

Lampiran 13. *Centroid* Variabel Setiap *Cluster* Untuk Algoritma *Ward* 63

Lampiran 14. *Centroid* Variabel Setiap *Cluster* Untuk Algoritma *K-means*..... 64

Lampiran 15. Grafik Rata-Rata *Silhouette Coefficient* untuk Algoritma *Ward*.... 66

Lampiran 16. Grafik Rata-Rata *Silhouette Coefficient* untuk Algoritma *K-means* 67

Lampiran 17. Sintaks Program *R Studio* untuk Algoritma *Ward* 68

Lampiran 18. Sintaks Program *R Studio* untuk Algoritma *K-Means*..... 70

BAB I
PENDAHULUAN

1.1. Latar Belakang

Seiring berkembangnya zaman, teknologi, dan kehidupan sosial, berdampak pada berbagai aspek di antaranya adalah kesehatan. Indonesia dihadapkan dengan *double burden disease* yaitu penyakit menular (penyakit infeksi) dan penyakit tidak menular (penyakit degeneratif) yang semakin meningkat. Pada masa sekarang, penyakit tidak menular telah menggeser penyakit infeksi sebagai penyakit yang mendominasi dan menjadi penyebab kematian tertinggi. Diabetes Mellitus menjadi penyebab terbesar kerugian kesehatan di Indonesia, sedangkan kasus kematian akibat penyakit tuberkolosis dan diare sudah menurun secara signifikan sejak tahun 1990 (IHME,2018). Penyakit menular menjadi salah satu penyebab utama kematian di dunia. Penyebabnya munculnya penyakit baru (*new emerging disease*) dan munculnya kembali penyakit menular yang lama (*re-emerging disease*) membuat Indonesia menanggung beban berlebih dalam penanggulangan penyakit (*triple burden disease*) (Kemenkes, 2013).

Indonesia saat ini masih menghadapi transisi epidemiologi terkait *triple burden disease,* dengan kata lain Indonesia memiliki beban penyakit ganda. Yang pertama karena telah bergesernya penyakit menular ke arah penyakit tidak menular seperti penyakit jantung, gagal ginjal, diabetes, kanker, dan lainnya. Yang kedua munculnya ancaman penyakit infeksi baru seperti flu burung, ebola, TBC resisten obat. Dan yang terakhir masyarakat masih dihadapkan pada masalah penyakit menular yang belum selesai, yaitu demam berdarah, TBC, malaria, HIV-AIDS, filariasis dan cacingan. Terkait masalah gizi, Indonesia juga tengah menghadapi beban ganda yaitu tentang masalah undernutrisi (gizi kurang, pendek/stunting, dan kurus) juga masalah overnutrisi yakni obesitas/kegemukan. (Menkes, 2018).

Provinsi Sulawesi Selatan merupakan salah satu daerah di Indonesia yang masih menghadapi permasalahan dibidang kesehatan. Berdasarkan data Dinas Kesehatan Provinsi Sulawesi Selatan tercatat penderita beberapa penyakit seperti demam berdarah *dengue* (DBD) di Sulawesi Selatan sejak Januari hingga Oktober

2017 mencapai 1.363 orang yang tersebar di 24 kabupaten/kota, dan data ini masih bersifat sementara. Pada kasus lain, sampai bulan April 2017 ditemukan sekitar 6.135 orang terindikasi penyakit TBC. Jumlah pasien TBC paru yang dinyatakan terkonfirmasi bakteriologis mencapai 246 orang, sedangkan pasien yang terdiagnosis klinis ada 54 orang. Adapun presentase untuk beberapa kasus penyakit lainnya pada tahun 2017 yakni diare sebesar 66,11 persen, pneumonia pada balita sebesar 3,78 persen, obesitas 23,55 persen, dan bayi berat badan lahir rendah sebesar 3,61 persen.

Setiap tahun terdapat kasus masyarakat yang menderita berbagai penyakit, termasuk di provinsi Sulawesi Selatan. Maka untuk mengetahui tingkat kesehatan masyarakat perlu dilakukan pengelompokan kabupaten atau kota berdasarkan karakteristiknya menurut jumlah kasus penyakit yang diderita. Hal ini dilakukan agar pemerintah dapat melakukan penanggulangan atau program yang sesuai dengan tingkat kesehatan di daerah tersebut. Adapun pengelompokan dilakukan dengan cara analisis *cluster* (kelompok) untuk mengidentifikasi permasalahan ini.

Analisis *cluster* merupakan suatu metode yang dipergunakan untuk mengklasifikasi atau mengelompokkan objek (kasus) seperti orang, produk (barang), toko, perusahaan ke dalam *cluster* (kelompok) yang relatif homogen. Objek atau kasus dalam setiap kelompok cenderung mirip satu sama lain dan berbeda jauh (tidak sama) dengan objek dari kelompok yang lainnya. Analisis *cluster* juga disebut analisis klasifikasi dimana setiap objek hanya masuk ke dalam satu kelompok saja, tidak terjadi tumpang tindih (Supranto, 2010).

Prosedur analisis *cluster* dapat dilakukan dengan dua metode yaitu metode hirarki dan metode nonhirarki. Metode hirarki adalah teknik *cluster* yang membentuk tingkatan tertentu seperti struktur pohon dengan proses pengelompokkannya dilakukan secara bertingkat atau bertahap dan hasilnya disajikan dalam bentuk dendogram. Metode ini terdiri dari *Single Linkage, Complete Linkage, Average Linkage*, *Ward* dan *Centroid* (Rafsanjani dkk, 2012). Sedangkan metode non hierarki sering disebut sebagai *K-Means*. Prosedur pada metode non hierarki dimulai dengan memilih sejumlah nilai *cluster* awal sesuai dengan jumlah yang diinginkan, kemudian objek pengamatan digabungkan ke

dalam *cluster-cluster* tersebut. Metode non hierarki ini meliputi metode *sequential threshold, parallel threshold*, dan *optimizing partitioning* (Gudono, 2011).

Tri Febriana Laraswati (2014) sebelumnya telah meneliti menggunakan data jumlah kasus penyakit di provinsi D.I. Yogyakarta tahun 2013 untuk melakukan Perbandingan Kinerja Metode *Complete Linkage*, Metode *Average Linkage*, dan Metode *K-Means* Dalam Menentukan Hasil Analisis *Cluster*. Penelitian ini untuk melakukan pemilihan metode terbaik diantara beberapa metode yang di gunakan juga penggunaan algoritma *K-Means* dilakukan karena algoritma *K-Means* merupakan algoritma dengan proses yang cukup sederhana. Penelitian yang dilakukan oleh Sarfiah (2016) dengan judul "Penerapan Metode *Ward* pada Analisis Gerombol Hierarki" menggunakan jarak kuadrat *Euclidean (Squared Euclidean)* untuk mengukur jumlah perbedaan nilai pada masing-masing variabelnya. Penelitian lainnya tentang algoritma *K-Means* yang dilakukan oleh Muh. Fadil Ilyas (2018) yang berjudul "Pengelompokkan Kabupaten/Kota Di Provinsi Sulawesi Selatan Berdasarkan Indikator Ekonomi Dengan Menggunakan Metode *K-Means Clustering* Non Hirarki". Penelitian tentang algoritma *K-Means* juga dilakukan oleh Ravena Lailatul Rohmah (2019) yang dimana penelitian ini menggunakan tiga metode yaitu *Silhouette Coefficient, Davies Bouldin Index dan Purity* dalam melakukan uji validasi *cluster.*

Berdasarkan latar belakang dan juga beberapa penelitian yang telah dilakukan, penulis tertarik untuk melakukan studi lebih lanjut tentang penerapan metode analisis *cluster* pada bidang kesehatan dengan mengelompokkan kabupaten/kota di Sulawesi Selatan untuk mengetahui tingkat kesehatan menurut jumlah kasus penyakit yang diderita oleh masyarakat sebagai tugas akhir dengan judul **"Perbandingan Kinerja Algoritma *Ward* dan Algoritma *K-Means* dengan Uji *Silhouette Coefficient*"**

1.2. Rumusan Masalah

Berdasarkan uraian latar belakang masalah di atas, maka dapat dirumuskan permasalahan sebagai berikut :

1. Bagaimana peng*cluster*an yang terbentuk menggunakan algoritma *Ward*?
2. Bagaimana peng*cluster*an yang terbentuk menggunakan algoritma *K-Means*?
3. Bagaimana perbandingan algoritma terbaik dari 2 algoritma yang digunakan yaitu algoritma *Ward* dan algoritma *K-Means* menggunakan uji *Silhouette Coefficient?*

1.3. Batasan Masalah

Penulis membatasi penelitian menggunakan jarak kuadrat *Euclidean* untuk mengukur kemiripan antar objek dalam algoritma *Ward* dan algoritma *K-Means* pada data kasus penyakit di Provinsi Sulawesi Selatan tahun 2017.

1.4. Tujuan Penelitian

Berdasarkan rumusan masalah, maka tujuan penelitian ini adalah :

1. Mendapatkan peng*cluster*an yang terbentuk dengan penggunaan algoritma *Ward.*
2. Mendapatkan peng*cluster*an yang terbentuk dengan penggunaan algoritma *K-Means*.
3. Membandingkan hasil peng*cluster*an yang terbentuk dari algoritma *Ward* dan algoritma *K-Means* dengan menggunakan uji *Silhouette Coefficient*.

1.5. Manfaat Penelitian

Hasil penelitian ini diharapkan dapat bermanfaat untuk menambah pengetahuan tentang aplikasi pada ilmu statistik yaitu bidang kesehatan khususnya dengan menggunakan analisis *cluster*. Serta menjadi informasi bagi pemerintah dalam melihat tingkat kesehatan di Provinsi Sulawesi Selatan selanjutnya untuk pengambilan kebijakan.

BAB II
TINJAUAN PUSTAKA

2.1. Analisis Multivariat

Analisis multivariat adalah analisis multi variabel dalam satu atau lebih hubungan. Analisis ini berhubungan dengan semua teknik statistik yang secara simultan menganalisis sejumlah pengukuran pada individu atau objek dalam suatu riset (Santoso, 2014). Menurut Johnson, analisis multivariat mencakup analisis data penelitian yang menggunakan banyak variabel yang dikenakan pengukuran secara bersamaan (Sarwono, 2013).

Selanjutnya dalam analisis multivariat dibagi menjadi dua kategori metode, yaitu metode dependensi dan interdepedensi. Metode dependensi terdapat dua jenis variabel yaitu variabel bebas dan terikat, metode ini terdiri atas beberapa teknik analisis, yaitu regresi berganda, analisis diskriminan, dan manova. Sedangkan metode interdependensi hanya terdapat satu jenis variabel, yaitu variabel bebas, metode ini terdiri atas beberapa teknik analisis, yaitu analisis faktor, analisis *cluster* dan *multidimensional scaling* (Wijaya, 2010).

Data multivariat adalah data yang dikumpulkan dari dua atau lebih observasi dengan mengukur observasi tersebut dengan beberapa karakteristik. Data dalam analisis multivariat dapat dinyatakan dalam bentuk matriks. Ukuran matriks dijelaskan dengan menyatakan banyak baris (garis horizontal) dan banyak kolom (garis vertikal) yang terdapat dalam matriks tersebut. Jika **X** adalah sebuah matriks, maka digunakan x_{ij} untuk menyatakan entri yang terdapat di dalam baris *i* dan kolom *j* dari **X**, *i* = 1, 2,..., *m*; *j* =1, 2, ...,*n*. Jadi matriks berukuran *m* x *n* secara umum dapat dituliskan sebagai berikut :

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \dots & x_{mn} \end{bmatrix}$$

2.2. Matriks Varian – Kovarian dan Matriks Korelasi

Matriks varian-kovarian dari sampel disimbolkan dengan **S**. Matriks varian-kovarian sampel dapat diperoleh dengan cara sebagai berikut:

$$\boldsymbol{S} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\begin{bmatrix}(x_{i1}-\bar{x}_1)\\(x_{i2}-\bar{x}_2)\\\vdots\\(x_{im}-\bar{x}_m)\end{bmatrix}[x_{i1}-\bar{x}_1 \quad x_{i2}-\bar{x}_2 \quad \cdots \quad x_{im}-\bar{x}_m]'\right)$$

$$\boldsymbol{S} = \frac{1}{n-1}\sum_{i=1}^{n}\begin{bmatrix}(x_{i1}-\bar{x}_1)^2 & (x_{i1}-\bar{x}_1)(x_{i2}-\bar{x}_2) & \dots & (x_{i1}-\bar{x}_1)(x_{im}-\bar{x}_m)\\(x_{i2}-\bar{x}_2)(x_{i1}-\bar{x}_1) & (x_{i2}-\bar{x}_2)^2 & \dots & (x_{i2}-\bar{x}_2)(x_{im}-\bar{x}_m)\\\vdots & \vdots & \ddots & \vdots\\(x_{im}-\bar{x}_m)(x_{i1}-\bar{x}_1) & (x_{im}-\bar{x}_m)(x_{i2}-\bar{x}_2) & \dots & (x_{im}-\bar{x}_m)^2\end{bmatrix}$$

$\boldsymbol{S} =$

$$\begin{bmatrix}\frac{1}{n-1}\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)^2 & \frac{1}{n-1}\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)(x_{i2}-\bar{x}_2) & \dots & \frac{1}{n-1}\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)(x_{im}-\bar{x}_m)\\\frac{1}{n-1}\sum_{i=1}^{n}(x_{i2}-\bar{x}_2)(x_{i1}-\bar{x}_1) & \frac{1}{n-1}\sum_{i=1}^{n}(x_{i2}-\bar{x}_2)^2 & \dots & \frac{1}{n-1}\sum_{i=1}^{n}(x_{i2}-\bar{x}_2)(x_{im}-\bar{x}_m)\\\vdots & \vdots & \ddots & \vdots\\\frac{1}{n-1}\sum_{i=1}^{n}(x_{im}-\bar{x}_m)(x_{i1}-\bar{x}_1) & \frac{1}{n-1}\sum_{i=1}^{n}(x_{im}-\bar{x}_m)(x_{i2}-\bar{x}_2) & \dots & \frac{1}{n-1}\sum_{i=1}^{n}(x_{im}-\bar{x}_m)^2\end{bmatrix}$$

$$\boldsymbol{S} = \begin{bmatrix}S_{1^2} & S_{12} & \dots & S_{1m}\\S_{12} & S_{2^2} & \dots & S_{2m}\\\vdots & \vdots & \ddots & \vdots\\S_{1m} & S_{2m} & \dots & S_{m^2}\end{bmatrix} \qquad (2.1)$$

Ukuran keeratan hubungan linear antara variabel random X_i dan X_j adalah koefisien korelasi sampel r_{ij} yang dinyataan dalam varian S_{ii} dan kovarian S_{ij} pada persamaan berikut ini:

$$r_{ij} = \frac{S_{ij}}{\sqrt{S_{ii}}\sqrt{S_{jj}}} \qquad (2.2)$$

Matriks koefisien korelasi adalah matriks simetri berukuran $m\,x\,m$ yang dapat dinyatakan dalam persamaan 2.2 berikut ini:

$$\boldsymbol{R} = \begin{bmatrix}\frac{S_1{}^2}{\sqrt{S_1{}^2}\sqrt{S_1{}^2}} & \frac{S_{12}}{\sqrt{S_1{}^2}\sqrt{S_2{}^2}} & \dots & \frac{S_{1m}}{\sqrt{S_1{}^2}\sqrt{S_m{}^2}}\\\frac{S_{12}}{\sqrt{S_1{}^2}\sqrt{S_1{}^2}} & \frac{S_2{}^2}{\sqrt{S_2{}^2}\sqrt{S_2{}^2}} & \dots & \frac{S_{2m}}{\sqrt{S_m{}^2}\sqrt{S_{22}}}\\\vdots & \vdots & & \vdots\\\frac{S_{1m}}{\sqrt{S_m{}^2}\sqrt{S_1{}^2}} & \frac{S_{2m}}{\sqrt{S_m{}^2}\sqrt{S_2{}^2}} & \ddots & \frac{S_m{}^2}{\sqrt{S_m{}^2}\sqrt{S_m{}^2}}\\ & & \dots & \end{bmatrix}$$

$$= \begin{bmatrix}1 & r_{21} & \dots & r_{1m}\\r_{12} & 1 & \dots & r_{2m}\\\vdots & \vdots & \ddots & \vdots\\r_{1m} & r_{2m} & \dots & 1\end{bmatrix} \qquad (2.3)$$

Apabila semua variabel yang digunakan memiliki unit satuan yang sama maka matriks masukan yang digunakan adalah matriks varian kovarian, dimana ***S*** adalah varian kovarian dari variabel asal (Wardani, 2017).

2.3. Analisis *Cluster*

Analisis *cluster* adalah metode yang tepat untuk mengidentifikasi objek - objek yang homogen ke dalam kelompok – kelompok. *Cluster* yang terbentuk memiliki homogenitas internal yang tinggi dan heterogenitas eksternal yang tinggi (Hair, 2010). Sedangkan menurut Gudono (2011), analisis *cluster* merupakan suatu teknik analisis multivariat yang bertujuan untuk meng*cluster*kan data observasi ataupun variabel-variabel ke dalam *cluster* sedemikian rupa sehingga masing-masing *cluster* bersifat homogen sesuai dengan faktor yang digunakan untuk melakukan peng*cluster*an. Karena yang diinginkan adalah untuk mendapatkan *cluster* yang sehomogen mungkin, maka yang digunakan sebagai dasar untuk meng*cluster*kan adalah kesamaan skor nilai yang dianalisis.

Ciri-ciri analisis *cluster* yang baik yaitu mempunyai :

1. Homogenitas internal (*within cluster*); yaitu kesamaan antar anggota dalam satu *cluster*. Setiap anggota kelompok atau gerombol homogen mempunyai karakteristik tertentu. Hal ini berarti bahwa observasi dalam setiap kelompok sama dengan observasi lain dalam satu kelompok yang sama.
2. Heterogenitas external (*between cluster*); yaitu perbedaan. Setiap kelompok seharusnya berbeda dari kelompok lain dengan karakteristik yang sama. Hal ini berarti bahwa observasi dalam kelompok yang satu seharusnya berbeda dari observasi dalam kelompok lain.

2.4. Uji *Kaiser Mayer Olkin*

Uji kecukupan data diperlukan untuk memastikan bahwa data yang telah diambil sudah cukup secara objektif atau benar telah mewakili dari papulasi. Untuk mengetahui kecukupan data dapat dilakukan uji *Kaiser Mayer Olkin* (KMO). Nilai KMO ini diperoleh dengan membandingkan besarnya koefisien korelasi terobservasi dengan besarnya koefisien korelasi parsial. KMO dirumuskan sebagai berikut :

$$KMO = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} r_{ij}^2}{\sum_{i=1}^{m}\sum_{j=1}^{n} r_{ij}^2 + \sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}^2} \quad (2.4)$$

Dengan :

r_{ij} : koefisien korelasi terobservasi antara variabel i dan j

a_{ij} : koefisien korelasi parsial antara variabel i dan j

Tabel KMO dapat ditunjukkan sebagai berikut :

Tabel 2. 1 Kriteria Pengukuran Uji KMO

Ukuran KMO	Keterangan
≥ 0,90	Baik Sekali
≥ 0,80	Baik
≥ 0,70	Sedang
≥ 0,60	Cukup
≥ 0,50	Kurang
Dibawah 0,50	Ditolak

Dengan demikian, nilai KMO kurang dari 0.5 menandakan bahwa sampel yang diambil tidak dapat mewakili populasi yang ada. (Machfudhoh, 2013).

2.5. Uji Multikolinearitas

Multikolinearitas adalah adanya korelasi atau hubungan yang sangat tinggi antar variabel bebas (Yamin, 2011). Multikolinearitas antar variabel adalah satu pelanggaran asumsi di dalam analisis *cluster* (Hair, 1998). Salah satu cara untuk mendeteksi ada atau tidaknya multikolinearitas adalah dengan menggunakan nilai koefisien korelasi *Pearson*. Multikolinieritas terjadi apabila antar variabel prediktor memiliki nilai koefisien korelasi *Pearson* mencapai atau melebihi 0,8. Jika setelah dideteksi ternyata diketahui bahwa terdapat multikolinearitas antar variabel, maka multikolinearitas dapat diatasi dengan penerapan analisis komponen utama.

2.6. Analisis Komponen Utama

Analisis Komponen Utama (AKU) adalah suatu teknik analisis untuk mentransformasikan variabel asal yang masih berkorelasi satu dengan yang lain menjadi satu set variabel baru yang tidak berkorelasi lagi. Tujuan dari AKU adalah mereduksi dimensi data asal yang semula terdapat variabel bebas menjadi *K* komponen utama. AKU muncul sebagai solusi bagi proses pengumpulan data dimana data tersebut terdiri dari variabel-variabel yang jumlahnya sangat banyak sehingga diperoleh variabel-variabel baru yang jumlahnya lebih sedikit tetapi tetap mampu menjelaskan varian data (Johnson dan Wichern, 2007).

Adapun langkah-langkah dalam analisis komponen utama berdasarkan matriks korelasi :

a. Menghitung matriks varian kovarian **S**
b. Menghitung nilai eigen dari matriks varian kovarian **S** dengan menggunakan rumus sebagai berikut:

$$|\mathbf{S} - \lambda I| = 0 \tag{2.5}$$

c. $\hat{\boldsymbol{e}}_{\boldsymbol{i}}$ adalah vector eigen yang diperoleh dari setiap nilai eigen λ_m yang memenuhi:

$$(\mathbf{S} - \lambda I).\hat{\boldsymbol{e}}_{\boldsymbol{i}} = 0 \tag{2.6}$$

d. Menghitung proporsi varian menggunakan rumus

$$Total\ Varian = \frac{\lambda_i}{\lambda_1 + \lambda_2 + \cdots + \lambda_m}$$

Atau dapat juga dinyatakan dalam besaran persen dengan rumus :

$$Total\ Varian = \frac{\lambda_i}{\sum_{i=1}^{m} \lambda_i} \times 100\% \tag{2.7}$$

e. Menghitung proporsi kumulatif keragaman data asal yang dapat dijelaskan oleh komponen utama ke-*k* dengan menggunakan rumus :

$$Total\ Komulatif\ Varian = \frac{\sum_{j=1}^{n} \lambda_j}{\sum_{i=1}^{m} \lambda_i} \times 100\% \tag{2.8}$$

Kriteria pemilihan *K* (Johnson dan Wichern, 2007) adalah komponen utama yang dianggap dapat menjelaskan keragaman data asal dengan baik jika proporsi kumulatif keragaman data asal yang dijelaskan komponen utama minimal 80%.

2.7. Mengukur Kemiripan atau Ketidakmiripan Antar Objek

Tujuan pokok mengidentifikasi objek yang disajikan oleh data ke dalam *cluster* adalah untuk mengetahui seberapa "dekat" atau seberapa "jauh" objek yang satu dengan objek yang lain. Ukuran kuantitatif yang menunjukkan kedekatan antar objek yang sering digunakan adalah ketidakmiripan (*dissimilarity*), jarak (*distance*), kemiripan (*similarity*) yang secara umum dikenal dengan istilah kedekatan (*proximity*). Objek dikatakan "dekat" ketika kemiripan atau jarak di antara mereka kecil dan kemiripan mereka besar (Everitt, 2011).

Ukuran jarak yang akan digunakan dalam penelitian ini adalah jarak kuadrat *Euclidean (Squared Euclidean Distance)*. *Squared Euclidean Distance* yang merupakan variasi dari jarak *Euclidean*. Jarak kuadrat *Euclidean* adalah jumlah kuadrat perbedaan deviasi di dalam nilai untuk setiap variabel (Hair, 2010). Kalau pada jarak *Euclidean* diakarkan, maka pada jarak kuadrat *Euclidean* akar tersebut dihilangkan. Jarak kuadrat *Euclidean* antara kelompok objek ke-*i* dan kelompok objek ke-*j* dari *m* variabel didefinisikan sebagai berikut:

$$d_{ij} = \sum_{p=1}^{n} \left(x_{ip} - x_{jp}\right)^2 \tag{2.9}$$

dengan :

n = banyaknya variabel

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

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

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

2.8. Metode Hierarki

Hierarki adalah analisis yang peng*cluster*an datanya dilakukan dengan cara mengukur jarak kedekatan pada setiap objek yang kemudian membentuk sebuah dendogram. Metode hierarki memulai pengelompokkan dengan dua atau lebih objek yang mempunyai kesamaan paling dekat. Kemudian diteruskan pada objek yang lain dan seterusnya hingga *cluster* akan membentuk semacam pohon dimana terdapat tingkatan (hierarki) yang jelas antara objek, dari yang paling mirip hingga

yang paling tidak mirip. Dendogram adalah representasi visual dari langkah-langkah dalam analisis *cluster* yang menunjukkan bagaimana *cluster* terbentuk dan nilai koefisien jarak pada setiap langkah. Angka sebelah kanan adalah objek penelitian, dimana objek tersebut dihubungkan oleh garis dengan objek yang lain sehingga pada akhirnya akan membentuk satu *cluster* (Simamora, 2005).

Metode penggerombolan berhirarki ini dibedakan menjadi dua yaitu metode penggabungan (*agglomerative*) dan metode pemisah (*divisive*). Metode *Agglomeratif* dalam pembentukan *cluster* terbagi atas Pautan Tunggal (*Single Linkage*), Pautan Lengkap (*Complete Linkage*), Pautan Rata-rata (*Average Linkage*), algoritma *Ward*, dan *Centroid* (Pusat). Metode *Divisive* beranjak dari sebuah kelompok besar yang terdiri dari semua objek. Objek yang paling besar nilai ketidakmiripannya dipisahkan.

2.9. Algoritma *Ward*

Metode varian sering juga disebut sebagai algoritma *Ward*. Metode ini menggunakan perhitungan yang lengkap dan memaksimumkan homogenitas di dalam satu *cluster*. Homogenitas atau kemiripan pada setiap *cluster* diukur dengan jumlah kuadrat objek-objek di dalam *cluster*, yang lebih dikenal sebagai *Sum of Squares Error* (SSE). Nilai SSE digunakan sebagai fungsi objektif.

Ward mengusulkan penggunaan metode yang didasarkan pada hasil informasi yang minimum dari kenaikan pada jumlah kuadrat deviasi rata-rata *cluster*. Metode *Ward* juga dikenal dengan metode varian minimum dan harus menggunakan jarak kuadrat *Euclidean*. Pada setiap tahap, dua *cluster* yang memiliki kenaikan "*Sum Of Squares Error"* dalam *cluster* yang terkecil digabungkan (Simamora, 2005). Algoritma *Ward* berusaha untuk meminimalkan variasi antar objek dalam suatu *cluster* dan memaksimalkan variasi dengan objek yang ada di *cluster* lainnya.

Agglomerative merupakan prosedur pengelompokan hierarki dimana setiap objek berawal dari *cluster* yang terpisah. *Cluster-cluster* dibentuk dengan mengelompokkan objek ke dalam *cluster* yang semakin banyak objek yang menjadi anggotanya. Proses ini dilanjutkan sampai semua objek menjadi anggota dari *cluster* tunggal. Algoritma *Ward* merupakan bagian dari metode

pengelompokan yang mengelompokkan *m* buah objek ke dalam *m*, *m* − 1, *m* − 2, dan seterusnya sampai menjadi satu *cluster*, dengan banyaknya *cluster* tidak diketahui. SSE hanya dapat dihitung jika *cluster* memiliki elemen lebih dari satu objek. SSE *cluster* yang hanya memiliki satu objek adalah nol.

Algoritma *Ward* ini dihitung berdasarkan persamaan berikut :

$$SSE = \sum_{i=1}^{m} (x_i - \bar{x})'(x_i - \bar{x}) \qquad (2.10)$$

dimana: x_i = vektor kolom yang entrinya nilai objek ke-i dengan i = 1,2,3,…m,

$\bar{x}$ = vektor kolom yang entrinya rata-rata nilai objek dalam gerombol

m = banyaknya objek

Langkah penyelesaian dengan algoritma *Ward*:

a. Dimulai dengan memperhatikan N *cluster* yang mempunyai satu responden per *cluster* (semua responden dinggap sebagai *cluster*).
b. Menentukan ukuran kemiripan dan ketakmiripan antara dua objek menggunakan rumus jarak kuadrat *euclidean.*

$$I_{ij} = SSE_{ij} = \frac{1}{2} \sum_{i=1}^{m} (x_i - \bar{x})'(x_i - \bar{x}) \qquad (2.11)$$

c. *Cluster* pertama dibentuk dengan memilih dua dari N *cluster* yang memiliki nilai SSE terkecil. Hal ini sejalan dengan fungsi tujuannya untuk meminimumkan keheterogenan.
d. Menentukan jumlah *cluster* dan anggotanya dengan metode *Ward.*

2.10. Metode Non-Hierarki

Pada metode non-hierarki, banyaknya *cluster* yang ingin dibentuk harus ditentukan terlebih dahulu. Metode non-hierarki sering disebut *K-Means clustering*. Pusat *cluster* yang dipilih pada metode ini merupakan pusat sementara dengan terus memperbaharui pusat *cluster* sampai kriteria pemberhentian tercapai. Algoritma *K-Means* adalah metode non-hierarki yang bersifat tanpa arahan *(unsupervised)* dan merupakan salah satu metode yang melakukan peng*cluster*an

dengan sistem partisi, hal ini dikarenakan data yang dianalisis tidak mempunyai label *cluster* yang berarti dalam proses peng*cluster*annya tidak mempunyai anggota *cluster* yang pasti. Objek yang sudah masuk ke dalam *cluster* tertentu masih bisa berpindah ke *cluster* yang lain. MacQueen berpendapat bahwa istilah *K-Means* digunakan untuk mendiskripsikan bahwa algoritma *K-Means* menandai setiap objek masuk ke dalam *cluster* yang mempunyai pusat (rata-rata) terdekat (Johnson dan Wichern, 2007).

Pada dasarnya metode *K-Means* berusaha untuk meminimalkan varian dalam satu *cluster* dan memaksimalkan varian antar *cluster*. Ada dua masalah utama pendekatan non hierarki. Pertama, jumlah atau banyaknya *cluster* harus ditentukan terlebih dahulu. Kedua, pemilihan pusat *cluster* tidak menentu (pasti). Seterusnya, hasil peng*cluster*an tergantung pada bagaimana pusat *cluster* dipilih. Banyak program yang dimulai dengan memilih kasus pertama *k* (*k* = jumlah *cluster*) sebagai pusat *cluster* awal. Jadi, hasil peng*cluster*an tergantung pada observasi data. Dibalik segala kekurangan itu, metode ini dapat dilakukan dengan cepat dan sangat bermanfaat kalau jumlah observasi besar (Simamora, 2005).

Langkah-langkah analisis *cluster* metode non-hierarki *K-Means* :

a. Menentukan *k* sebagai jumlah *cluster* yang dibentuk.
   Untuk menentukan banyaknya *cluster k* dilakukan dengan beberapa pertimbangan seperti pertimbangan teoritis dan konsepsual yang mungkin diusulkan untuk menentukan berapa banyak *cluster.*
b. Menentukan *centroid* (titik pusat *cluster*)
   Penentuan *centroid* awal dilakukan secara random/acak dari objek yang tersedia sebanyak *k cluster,* kemudian menghitung *centroid cluster* ke-*i* berikutnya, dengan menggunakan rumus :

$$C_{yq} = \frac{\sum_{q=1}^{m} x_q}{u} \qquad (2.12)$$

   Dimana :

   C_{yq} : Nilai *centroid* dari *cluster* ke-*y* pada variabel ke-q

   x_q : Nilai data pada variabel ke-q

   u : Banyaknya objek dalam *cluster* yang terbentuk

   m : Banyaknya variabel yang diukur

c. Hitung jarak setiap objek masing-masing *centroid* dari masing-masing *cluster*. Untuk menghitung jarak antar objek dengan *centroid* menggunakan jarak kuadrat *Euclidean* dengan rumus :

$$d_{(iy)} = \sum_{q=1}^{m} (x_{iq} - C_{yq})^2. \quad (2.13)$$

Dimana :

$d_{(iy)}$: Jarak kuadrat *Euclidean* antar objek ke - *i* dengan *centroid cluster* ke – *y*

m : Jumlah variabel *cluster*

x_{iq} : Nilai atau data dari objek ke-*i* pada variabel ke-*q*

C_{yq} : Nilai *centroid* dari *cluster* ke-*y* pada variabel ke-*q*

d. Alokasikan masing-masing objek ke dalam *centroid* yang paling dekat.

e. Lakukan iterasi kemudian tentukan posisi *centroid* baru dengan menggunakan persamaan pada poin (b).

f. Ulangi langkah pada poin (c) jika posisi *centroid* baru tidak sama (masih berubah).

2.11. Pemilihan Metode Terbaik Dengan *Silhoutte Coefficient*

Silhouette Coefficient merupakan sebuah metode yang digunakan untuk memvalidasi *cluster* dengan menggabungkan metode separasi dan metode kohesi. Metode separasi adalah sebuah ukuran seberapa jauh atau terpisah sebuah *cluster* dengan yang lainnya dan metode kohesi adalah sebuah ukuran seberapa dekat relasi antara objek dalam sebuah *cluster*. Validasi *cluster* dilakukan untuk mengukur kekuatan dan kualitas hasil *cluster* dari suatu objek terhadap suatu *cluster*.

Tahapan dalam melakukan perhitungan *Silhouette Coefficient* adalah sebagai berikut (Handoyo, 2014):

1. Hitung jarak rata-rata dari suatu objek misalkan objek ke-*i* dengan semua objek lain yang berada di dalam satu *cluster*.

$$a(i) = \frac{1}{[A]-1} \sum j \in_{A, j \neq i} d(i, j) \quad (2.14)$$

2. Hitung rata-rata jarak dari objek ke-*i* tersebut dengan semua objek pada *cluster* lainnya, kemudian ambillah nilai terkecilnya.

$$d(i,C) = \frac{1}{[A]}\sum j \in C\ d(i,j) \tag{2.15}$$

3. Nilai *Silhouette Coefficient*nya adalah:

   Jumlah *s(i)* diperoleh dengan menggabungkan *a(i)* dan *b(i)*:

$$s\begin{cases} 1-\dfrac{a(i)}{b} & jika\ a(i) < b(i), \\ 0 & jika\ a(i) = b(i) \\ \dfrac{b(i)}{a(i)} - 1 & jika\ a(i) > b(i) \end{cases}$$

Sehingga dapat dirumuskan:

$$s(i) = \frac{b(i) - a(i)}{max(a(i), b(i))} \tag{2.16}$$

Nilai rata-rata *Silhouette Coefficient* berada pada interval -1 $\leq$ *s(i)* $\leq$ 1. Nilai *Silhouette Coefficient* dari setiap objek dalam suatu *cluster* merupakan suatu ukuran yang menunjukkan seberapa dekat kemiripan data dikelompokkan di dalam satu *cluster* tersebut. Semakin nilai rata-rata *Silhouette Coefficient* mendekati angka 1 maka semakin baik pengelompokan data di dalam satu *cluster*. Sebaliknya jika nilai rata-rata *Silhouette Coefficient* mendekati angka -1, maka semakin buruk pengelompokan data di dalam satu *cluster* (Kaufman dan Rousseeauw, 2005).

Tabel 2. 2 Kriteria Pengukuran *Silhouette Coefficient*

Silhouette coefficient **(SC)**	**Kriteria**
0.7 < SC $\leq$ 1.0	*Struktur kuat*
0.5 < SC $\leq$ 0.7	*Struktur baik*
0.25 < SC $\leq$ 0.5	*Struktur lemah*
SC $\leq$ 0.25	*Struktur buruk*

2.12. Kesehatan

Derajat kesehatan atau tingkat kesehatan atau status kesehatan adalah skala yang dapat mengukur sehat atau sakitnya keadaan fungsi dan struktur jasmani mental sosial seseorang. Derajat kesehatan kelompok individsju merupakan hasil kumulatif dari derajat kesehatan individu anggota kelompok. Derajat kesehatan masyarakat yang optimal adalah tingkat kondisi kesehatan yang tinggi dari setiap orang atau masyarakat dan harus selalu diusahakan peningkatannya secara terus menerus.

Indikator derajat kesehatan adalah ukuran yang digunakan untuk melihat apakah derajat kesehatan masyarakat sudah optimal, yang dilihat dari unsur kualitas hidup (Dinas Kesehatan, 2013) yaitu :

a. Mortalitas (Angka Kematian)

   Gambaran perkembangan derajat kesehatan masyarakat dilihat dari kejadian kematian dalam masyarakat dari waktu ke waktu. Disamping itu kejadian kematian juga dapat digunakan sebagai indikator dalam penilaian keberhasilan pelayanan kesehatan dan program pembangunan kesehatan lainnya. Angka kematian pada umumnya dapat dihitung dengan melakukan berbagai survei dan penelitian.

b. Morbiditas (Angka Kesakitan)

   Tingkat kesakitan suatu negara juga mencerminkan situasi derajat kesehatan masyarakat yang ada didalamnya. Berikut ini morbiditas penyakit-penyakit menular dan tidak menular yang dapat mengambarkan keadaan derajat kesehatan masyarakat:

   1. Diare

      Penyakit diare merupakan suatu masalah yang mendunia. Penyakit diare jauh lebih banyak terdapat dinegara berkembang daripada negara maju, yaitu 12,5 kali lebih banyak di dalam kasus mortalitas (WHO/EIP, yang tidak dipublikasikan). Penyebab utama penyakit diare adalah infeksi bakteri atau virus. Jalur masuk utama infeksi tersebut melalui feses manusia atau binatang, makanan, air, dan kontak dengan manusia.

2. Pneumonia

Pneumonia adalah infeksi yang menyebabkan paru-paru meradang. Kantung-kantung kemampuan menyerap oksigen menjadi kurang. Kekurangan oksigen membuat sel-sel tubuh tidak bisa bekerja. Gara-gara inilah, selain penyebaran infeksi ke seluruh tubuh, penderita pneumonia bisa meninggal.

Pneumonia merupakan masalah kesehatan di dunia karena angka kematiannya tinggi, tidak saja di negara berkembang tetapi juga dinegara maju seperti Amerika Serikat, Kanada, dan negara-negara Eropa. Di indonesia, pneumonia merupakan penyebab kematian nomor tiga setelah kardiovaskuler dan tuberkulosis. Faktor sosial ekonomi yang rendah mempertinggi angka kematian.

3. TB Paru

Tuberkulosis paru-paru merupakan penyakit infeksi yang menyerang parenkim paru-paru yang disebabkan oleh Mycobacterium tuberculosis. Penyakit ini dapat juga menyebar ke bagian tubuh lain seperti meningen, ginjal, tulang dan nodus limfe. Diseluruh kasus tuberkulosis, sebesar 11% dialami oleh anak-anak di bawah 15 tahun.

4. Berat Badan Lahir Rendah (BBLR)

WHO pada tahun 1961 menyatakan bahwa semua bayi baru lahir yang berat badannya kurang atau sama dengan 2500 gram disebut low birth weight infant (bayi berat badan lahir rendah, BBLR). Cakupan berat bayi lahir rendah yang ditangani adalah berat bayi yang kurang dari 2500 gram yang ditangani di sarana pelayanan kesehatan sesuai tatalaksana berat bayi lahir rendah di suatu wilayah kerja pada kurun waktu tertentu.