

**IMPLEMENTASI ALGORITMA *CENTROID LINKAGE* DAN *K-MEDOIDS* DALAM
MENGELOMPOKKAN KABUPATEN/KOTA DI SULAWESI SELATAN
BERDASARKAN INDIKATOR PENDIDIKAN**

SKRIPSI



NUR ALFIANINGSIH RAJA

H 121 16 517

DEPARTEMEN STATISTIKA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS HASANUDDIN

MAKASSAR

2020

**IMPLEMENTASI ALGORITMA *CENTROID LINKAGE* DAN *K-MEDOIDS* DALAM
MENGELOMPOKKAN KABUPATEN/KOTA DI SULAWESI SELATAN
BERDASARKAN INDIKATOR PENDIDIKAN**

SKRIPSI

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

NUR ALFIANINGSIH RAJA

H 121 16 517

**PROGRAM STUDI STATISTIKA DEPARTEMEN STATISTIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS HASANUDDIN**

MAKASSAR

2020

LEMBAR PERNYATAAN KEOTENTIKAN

Yang bertanda tangan dibawah ini :

Nama : Nur Alfianingsih Raja
NIM : H12 16 517
Program Studi : Statistika
Jenjang : S1

Menyatakan dengan ini bahwa karya tulisan saya berjudul

**Implementasi Algoritma *Centroid Linkage* dan *K-Medoids* dalam
Mengelompokkan Kabupaten/Kota di Sulawesi Selatan berdasarkan
Indikator Pendidikan**

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, 30 Desember 2020



Nur Alfianingsih Raja
NIM H121 16 517

**IMPLEMENTASI ALGORITMA *CENTROID LINKAGE* DAN *K-MEDOIDS* DALAM
MENGELOMPOKKAN KABUPATEN/KOTA DI SULAWESI SELATAN
BERDASARKAN INDIKATOR PENDIDIKAN**

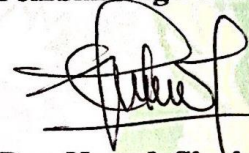
Disetujui Oleh:

Pembimbing Utama,



**Dr. Dr. Georgina M Tinungki, M.Si.
NIP. 19620926 198702 2 001**

Pembimbing Pertama,



**Dra. Nasrah Sirajang, M.Si.
NIP. 19650519 199303 2 002**

Ketua Departemen Statistika,




**Dr. Nurtli Samusi, S.Si., M.Si.
NIP. 19720117 199703 2002**

Pada Tanggal : 30 Desember 2020

HALAMAN PENGESAHAN

**IMPLEMENTASI ALGORITMA *CENTROID LINKAGE* DAN *K-MEDOIDS* DALAM
MENGELOMPOKKAN KABUPATEN/KOTA DI SULAWESI SELATAN
BERDASARKAN INDIKATOR PENDIDIKAN**

Disusun dan diajukan oleh

NUR ALFIANINGSIH RAJA

H 121 16 517

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 30 Desember 2020
dan dinyatakan telah memenuhi syarat kelulusan.

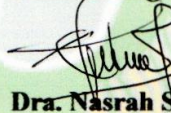
Menyetujui,

Pembimbing Utama,



Dr. Dr. Georgina Maria Tinungki, M.Si.
NIP. 19620926 198702 2001

Pembimbing Pendamping,



Dra. Nasrah Sirajang, M.Si.
NIP. 19650519 199303 2 002

Ketua Departemen Statistika,



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 “**Implementasi Algoritma Centroid Linkage dan K-Medoids dalam Mengelompokkan Kabupaten/Kota di Sulawesi Selatan berdasarkan Indikator Pendidikan**”. 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. Rajamuddin, M.Si** dan Ibunda **Ramlah**, yang telah melahirkan, membesarkan dan mendidik penulis. Terima kasih telah memberikan kasih sayang yang tak terhingga 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. Semoga Allah SWT selalu melindungi, memberikan kesehatan, rezeki yang baik dan keselamatan dunia akhirat kepada kedua orang tua penulis. Terima kasih kepada saudara kandung penulis, kakak **Nur Azizah Raja** dan saudara sepupu penulis, Adik **Resfianti** yang senantiasa mendoakan, membantu dan memberikan dukungan serta semangat yang tiada hentinya kepada penulis selama ini. Terima kasih telah menjadi saudara sekaligus teman terbaik;

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, M.Si.** selaku Ketua Departemen Statistika, serta segenap dosen pengajar dan staf Departemen Statistika yang telah membekali ilmu dan kemudahan-kemudahan kepada penulis dalam berbagai hal selama menjadi mahasiswa di Departemen Statistika;
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, pembimbing PKM dan Penasehat Akademik (PA) penulis selama menempuh pendidikan di Universitas Hasanuddin;
5. **Ibu Dra. Nasrah Sirajang, M.Si.** selaku dosen pembimbing pertama penulis, yang juga telah meluangkan waktunya untuk membimbing penulis sejak proposal hingga penyelesaian skripsi ini;
6. **Bapak Drs. Raupong, M.Si.** dan **Bapak Siswanto, S.Si., M.Si.** selaku tim penguji atas semua saran dan kritikan yang membangun dalam penyempurnaan penyusunan tugas akhir ini serta waktu yang telah diberikan kepada Penulis;
7. Terima kasih kepada sahabat-sahabat seperjuangan MINOR (**Dilla, Wiya, Nidar, Imma, Shasa, Cimma, Reza, Mila, Andis, Ririn, Aten, dan Tari**), dan GxPersona (**Nidar, Wana, Asma dan Reima**). Terima kasih telah memberikan potongan cerita hidup yang sangat istimewa kepada penulis selama kurang lebih 4 tahun terakhir ini. Bertemu dengan kalian adalah salah satu hal paling indah dari semua rencana Tuhan dalam hidup penulis. Kita punya mimpi bersama, sampai ketemu di puncak kesuksesan itu. Terima kasih karena selalu peduli dan pengertian atas segala hal tentang penulis baik itu di kala suka maupun duka;

8. 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. Terkhusus saudaraku **Azman**, terima kasih selalu setia untuk mendengar, berbagi kehidupan perkuliahan, membantu dan menemani penulis selama mengerjakan skripsi.
9. Keluarga besar Badan Eksekutif Mahasiswa Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Hasanuddin (**BEM FMIPA UNHAS**), 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!*
10. Terima kasih kepada teman-teman KKN Posko Desa Pa’bumbungan (**Ma’un, Ridha, Tari, Rina, Riri, Nina, Pite’, Gori, Uja, Ibe, Ciwang, Asta**), yang telah menjadi teman sekaligus keluarga dalam mengabdikan.
11. 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, 30 Desember 2020



Nur Alfianingsih Raja

**PERNYATAAN PERSETUJUAN PUBLIKASI
TUGAS AKHIR UNTUK KEPENTINGAN AKADEMIK**

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

Nama : Nur Alfianingsih Raja
NIM : H 121 16 517
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-exclusive Royalty- Free Right*)** atas tugas akhir saya yang berjudul:


“Implementasi Algoritma *Centroid Linkage* dan *K-Medoids* dalam Mengelompokkan Kabupaten/Kota di Sulawesi Selatan berdasarkan Indikator Pendidikan”

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 30 Desember 2020

Yang menyatakan,


Nur Alfianingsih Raja

ABSTRAK

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. Penelitian ini menggunakan algoritma *Centroid linkage* yang berguna untuk membentuk kelompok berdasarkan jarak antar *centroid*nya dan algoritma *K-Medoids* yang didasarkan pada penggunaan objek yang paling terpusat (*medoid*) untuk mengelompokkan kabupaten/kota dan memperoleh hasil perbandingan berdasarkan data indikator pendidikan di Sulawesi Selatan. Implementasi Algoritma *Centroid Linkage* dan *K-Medoids* pada data indikator pendidikan di Sulawesi selatan tahun 2018 yakni pengelompokkan kabupaten/kota di Sulawesi Selatan yang menghasilkan 2 *cluster* dengan *cluster* 1 sebanyak 21 kabupaten/kota, *cluster* 2 sebanyak 3. Untuk menentukan metode terbaik dilihat dari nilai rasio Simpangan Baku dalam *Cluster* (S_W) dan Simpangan Baku antar *Cluster* (S_B) yaitu memiliki nilai rasio simpangan baku (S) yang sama pada algoritma *Centroid Linkage* dan *K-Medoids* sebesar 104.967.

Kata Kunci: Analisis *Cluster*, *Centroid Linkage*, *K-Medoids*, Simpangan baku.

ABSTRACT

Cluster analysis is a multivariate analysis technique that aims to cluster the observational data or variables into clusters in such a way that each cluster is homogeneous according to the factors used for clustering. This study used the Centroid linkage algorithm that was useful for forming groups based on the distance between centroids and the K-Medoids algorithm that was based on the use of the most centered object (medoid) to group districts/cities and obtained comparison results based on the education indicator data in South Sulawesi. The implementation of the Centroid Linkage Algorithm and K-Medoids on the education indicator data in South Sulawesi in 2018, showed that the grouping of districts/cities in South Sulawesi produced 2 clusters with cluster 1 of 21 districts/cities, and cluster 2 of 3. To determine the best method, it was seen from the value of the Standard Deviation ratio in the cluster (S_W) and Standard Deviation between Clusters (S_B) showed the same standard deviation ratio (S) in the Centroid Linkage algorithm and K-Medoids that was equal to 104,967.

Keywords: *Centroid Linkage, Cluster Analysis, K-Medoids, Standard Deviation.*

DAFTAR ISI

HALAMAN JUDUL.....	i
LEMBAR PERNYATAAN KEOTENTIKAN.....	ii
LEMBAR PERSETUJUAN PEMBIMBING	iii
HALAMAN PENGESAHAN.....	iv
KATA PENGANTAR	iv
PERNYATAAN PERSETUJUAN PUBLIKASI	vii
ABSTRAK	ix
ABSTRACT	x
DAFTAR ISI.....	xi
DAFTAR GRAFIK.....	xiii
DAFTAR TABEL.....	xiv
DAFTAR LAMPIRAN	xv
BAB I PENDAHULUAN.....	1
1.1 Latar Belakang	1
1.2 Rumusan Masalah.....	3
1.3 Batasan Masalah	3
1.4 Tujuan Penelitian	3
1.5 Manfaat Penelitian	3
BAB II TINJAUAN PUSTAKA.....	4
2.1 Analisis Multivariat	4
2.2 Matriks Varians-Kovarians dan Matriks Korelasi.....	4
2.3 Analisis Cluster.....	6
2.4 Ukuran Jarak Euclidean	7
2.5 Centroid Linkage	8
2.6 K-Medoids	9
2.7 Multikolinearitas	10
2.8 Principal Component Analysis (PCA).....	11
2.9 Penentuan Jumlah Cluster Optimum	12
2.10 Penentuan Metode Terbaik dengan Simpangan Baku	12

2.11	Interpretasi Cluster	13
2.12	Indikator Pendidikan	14
BAB III METODE PENELITIAN.....		17
3.1	Sumber Data.....	17
3.2	Deskripsi Variabel	17
3.3	Metode Analisis	18
BAB IV HASIL DAN PEMBAHASAN		21
4.1	Statistik Deskriptif	21
4.2	Uji Multikolinearitas	22
4.3	Principal Component Analysis	23
4.4	Menentukan Ukuran Jarak Euclidean	27
4.5	Menentukan Jumlah Cluster Optimum dengan Indeks Validitas Silhouette.....	29
4.6	Proses Pengclusteran dengan Algoritma Centroid Linkage dan K-Medoids	30
4.6.1	Algoritma Centroid Linkage	30
4.6.2	Algoritma K-Medoids	32
4.7	Menentukan Jumlah Anggota Cluster.....	33
4.8	Menentukan Metode Terbaik dengan Simpangan Baku	35
4.9	Interpretasi Cluster	37
BAB V PENUTUP.....		39
5.1	Kesimpulan	39
5.2	Saran	40
DAFTAR PUSTAKA		41
LAMPIRAN		44

DAFTAR GRAFIK

Grafik 4. 1 Boxplot Statistik Deskriptif	22
Grafik 4. 2 Plot Silhouette.....	29
Grafik 4. 3 Dendogram Analisis Cluster Hirarki Algoritma Centroid Linkage... 32	
Grafik 4. 4 Plot Analisis Cluster Non Hirarki Algoritma K-Medoids	33

DAFTAR TABEL

Tabel 2. 1 Koefisien Korelasi.....	10
Tabel 4. 1 Matriks Korelasi antar Variabel	23
Tabel 4. 2 Proporsi varians dan proporsi varians kumulatif.....	24
Tabel 4. 3 Koefisien Principal Component (PC).....	25
Tabel 4. 4 Nilai atau Principal Component Analysis Score	27
Tabel 4. 5 Jarak Euclidean Kabupaten/Kota	28
Tabel 4. 6 Cluster Membership	33
Tabel 4. 7 Cluster beserta anggotanya dengan Algoritma Centroid Linkage dan K-Medoids.....	34
Tabel 4. 8 Rata-Rata Variabel (Centroid) Setiap Cluster	37

DAFTAR LAMPIRAN

Lampiran 1. Data Indikator Pendidikan Menurut Ijazah dan Jenis Kelamin Tahun 2018.....	45
Lampiran 2. Matriks Varians Kovarians	46
Lampiran 3. Koefisien Principal Component (PC)	47
Lampiran 4. Proporsi varians dan proporsi varians kumulatif	48
Lampiran 5. Nilai atau Principal Component Analysis Score	49
Lampiran 6. Jarak Euclidean Kabupaten/Kota.....	50
Lampiran 7. Cluster Membership.....	53
Lampiran 8. Rata-Rata Variabel (Centroid) Cluster 1	54
Lampiran 9. Rata-Rata Variabel (Centroid) Cluster 2	56
Lampiran 10. Rata-Rata Variabel pada Setiap Kabupaten/Kota.....	57
Lampiran 11. Agglomeration Schedule Algoritma Centroid Linkage	58
Lampiran 12. Output Algoritma K-Medoids.....	59
Lampiran 13. Output Plot Indeks Validitas Silhouette	59
Lampiran 14. Perhitungan manual algoritma k-medoids	60

BAB I

PENDAHULUAN

1.1 Latar Belakang

Pendidikan adalah adalah proses pengubahan sikap dan tata laku seseorang atau kelompok orang dalam usaha mendewasakan manusia melalui upaya pengajaran dan pelatihan. Pendidikan berkualitas merupakan prinsip dasar pembangunan untuk menciptakan sumber daya manusia yang berdaya saing tinggi. Disamping itu, akses untuk memperoleh kesempatan belajar yang sama dan merata dalam setiap jenjang pendidikan diharapkan dapat tercapai guna menjamin kualitas pendidikan yang inklusif dan merata. Pendidikan juga menjadi salah satu tujuan dalam *Sustainable Development Goals* (SDG's) (Badan Pusat Statistik, 2019).

Pentingnya peran pendidikan terhadap kemajuan bangsa, pengukuran dan penghitungan indikator-indikator pendidikan perlu dilakukan untuk melihat sejauh mana pemerataan pendidikan. Agar dapat mengetahui pemerataan pendidikan atau karakteristik indikator pendidikan menurut jenis kelamin dan tingkat pendidikan maka dilakukan pengelompokan menggunakan analisis *cluster*. Analisis *cluster* merupakan suatu metode untuk mengelompokkan objek atau unit penelitian ke dalam beberapa kelompok dan setiap unit penelitian dalam suatu kelompok akan mempunyai ciri yang relatif sama sedangkan antar kelompok unit penelitian memiliki sifat yang berbeda.

Analisis *cluster* dibagi menjadi dua metode yaitu *hierarchical method* dan *non-hierarchical method*. Dalam *hierarchical method* jumlah kelompok yang akan diperoleh belum diketahui, sedangkan dalam *non-hierarchical method* diasumsikan ada k kelompok terlebih dahulu. *Hierarchical method* terdiri dari *Single Linkage*, *Complete Linkage*, *Centroid Linkage*, *Average Linkage* dan *Ward's Method*. Sedangkan metode yang termasuk *non-hierarchical method* adalah metode *K-Means* dan *K-Medoids* (Rachmatin, 2014).

Pada penelitian ini, akan digunakan pendekatan *hierarchical method* yaitu *Centroid Linkage*. Metode ini dikenal lebih memiliki beban komputasi yang relatif lebih ringan karena hanya cukup menghitung titik tengah antar *cluster*. Metode ini juga baik untuk data yang mengandung outlier. Selain itu, untuk mendapatkan keterbandingan dengan metode yang lain akan digunakan *non-hierarchical method* yaitu metode *K-Medoids*. Metode *k-medoids* ini menggunakan objek pada kumpulan objek untuk mewakili sebuah *cluster*. Kelebihan dari metode ini mampu mengatasi kelemahan dari metode *k-means* yang *sensitive* terhadap *outlier* dan hasil proses *clustering* tidak bergantung pada urutan masuk pada suatu himpunan data (Setiyawati, 2017).

Penelitian *centroid linkage* sebelumnya dilakukan oleh Rini Silvi (2018) dengan judul Analisis Cluster dengan Data Outlier Menggunakan *Centroid Linkage* dan *K-Means Clustering* untuk Pengelompokan Indikator HIV/AIDS di Indonesia. Perbandingan rasio S_W/S_B , *Centroid Linkage* lebih homogen dibandingkan *K-means*. Adapun penelitian yang menggunakan metode *k-medoids clustering* dilakukan oleh Dini Marlina (2018) dengan judul Implementasi Algoritma *K-Medoids* dan *K-Means* untuk Pengelompokan Wilayah Sebaran Cacat pada Anak. Validitas yang digunakan pada penelitian ini adalah validitas *Silhouette Coefficient* dan dihasilkan *K-Medoids* lebih baik dalam melakukan pengelompokan pada data sebaran Anak Cacat dibandingkan dengan algoritma *K-Means*. Dalam penelitian Elok Fitriani (2016) yang berjudul “Pengelompokan Kabupaten/Kota di Jawa Timur berdasarkan Indikator Pendidikan Tahun 2013 menggunakan Analisis *Hierarchical Cluster*” dan terbentuk dua *cluster* yakni *cluster 1* dengan 29 Kabupaten memiliki tingkat pendidikan lebih tinggi dibandingkan *cluster 2* dengan 9 kabupaten.

Berdasarkan uraian di atas, penulis tertarik mengimplementasikan analisis *cluster* khususnya Algoritma *Centroid Linkage* dan *K-Medoids* dalam mengelompokan Kabupaten/Kota di Sulawesi Selatan berdasarkan Indikator Pendidikan.

1.2 Rumusan Masalah

Berdasarkan uraian pada latar belakang, masalah yang akan dikaji dalam penelitian ini adalah :

1. Bagaimana pengelompokan Kabupaten/Kota di Sulawesi Selatan menggunakan algoritma *Centroid Linkage* dan *K-Medoids* berdasarkan data indikator pendidikan?
2. Bagaimana hasil perbandingan antara algoritma *Centroid Linkage* dan *K-Medoids* untuk pengelompokan Kabupaten/Kota di Sulawesi Selatan berdasarkan data indikator pendidikan?

1.3 Batasan Masalah

Penelitian ini difokuskan pada penentuan algoritma terbaik menggunakan nilai simpangan baku yang mempunyai rasio terkecil dan penggunaan jarak *euclidean* serta data indikator pendidikan menurut ijazah dan jenis kelamin tahun 2018.

1.4 Tujuan Penelitian

Tujuan penelitian ini adalah sebagai berikut :

1. Untuk memperoleh hasil pengelompokan Kabupaten/Kota di Sulawesi Selatan menggunakan algoritma *Centroid Linkage* dan *K-Medoids* berdasarkan data indikator Pendidikan.
2. Untuk memperoleh hasil perbandingan antara algoritma *Centroid Linkage* dan *K-Medoids* untuk pengelompokan Kabupaten/Kota di Sulawesi Selatan berdasarkan data indikator Pendidikan.

1.5 Manfaat Penelitian

Penelitian ini diharapkan memberikan sumbangsi terhadap perkembangan ilmu pengetahuan dalam rangka memperluas dan memperdalam wawasan khususnya pada bidang kajian Analisis *Cluster*. Selain itu, penelitian ini sebagai referensi mengenai proses pengelompokan menggunakan Algoritma *Centroid Linkage* dan *K-Medoids*. Serta sumber informasi indikator pendidikan sebagai faktor penting dalam mendukung pembangunan negara.

BAB II

TINJAUAN PUSTAKA

2.1 Analisis Multivariat

Analisis multivariat adalah analisis statistik yang digunakan untuk menganalisis data yang terdiri dari beberapa variabel dan variabel-variabel tersebut saling berkorelasi satu sama lain. Secara umum analisis multivariat dibagi menjadi dua, yaitu analisis dependensi dan analisis interdependensi. Ciri dari analisis dependensi adalah adanya satu atau beberapa variabel yang berfungsi sebagai variabel tergantung dan variabel bebas, seperti, analisis regresi linear berganda, analisis diskriminan, analisis logit, dan analisis korelasi kanonik. Ciri dari analisis interdependensi adalah semua variabelnya bersifat independen. Berikut ini yang termasuk dalam analisis interdependensi adalah analisis faktor, analisis *cluster* dan *multidimensional scaling* (Sarwono, J, 2007).

Menurut Fadhli (2011), data dalam analisis multivariat dapat dinyatakan dalam bentuk matriks yang terdapat N objek dan M variabel. Misalkan pada suatu objek terdapat M variabel yaitu $X = (X_1, X_2, \dots, X_M)$ dari banyaknya N objek, maka data tersebut dapat disajikan dalam bentuk matriks seperti di bawah ini:

$$X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1M} \\ x_{21} & x_{22} & \dots & x_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \dots & x_{NM} \end{bmatrix}$$

2.2 Matriks Varians-Kovarians dan Matriks Korelasi

Matriks varians-kovarians dari sampel disimbolkan dengan Σ . Vektor mean sampel dan matriks varians-kovarians sampel dapat diperoleh dengan cara sebagai berikut:

$$\Sigma = \frac{1}{N-1} \sum_{i=1}^N \left(\begin{bmatrix} (x_{i1} - \bar{x}_1) \\ (x_{i2} - \bar{x}_2) \\ \vdots \\ (x_{iM} - \bar{x}_M) \end{bmatrix} [x_{i1} - \bar{x}_1 \quad x_{i2} - \bar{x}_2 \quad \dots \quad x_{iM} - \bar{x}_M] \right)$$

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

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

$$\Sigma = \begin{bmatrix} S_1^2 & S_{12} & \dots & S_{1M} \\ S_{12} & S_2^2 & \dots & S_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ S_{1M} & S_{2M} & \dots & S_M^2 \end{bmatrix} \quad (2.1)$$

Ukuran keeratan hubungan linear antara variabel random X_i dan X_j adalah koefisien korelasi sampel r_{ij} yang dinyatakan dalam varians S_i^2 dan kovarians S_{ij} pada persamaan berikut ini:

$$r_{ij} = \frac{S_{ij}}{\sqrt{S_i^2} \sqrt{S_j^2}} \quad (2.2)$$

Matriks koefisien korelasi adalah matriks simetri berukuran $M \times M$ yang dapat dinyatakan dalam persamaan 2.2 berikut ini:

$$R = \begin{bmatrix} \frac{S_1^2}{\sqrt{S_1^2} \sqrt{S_1^2}} & \frac{S_{12}}{\sqrt{S_1^2} \sqrt{S_2^2}} & \dots & \frac{S_{1M}}{\sqrt{S_1^2} \sqrt{S_M^2}} \\ \frac{S_{12}}{\sqrt{S_1^2} \sqrt{S_2^2}} & \frac{S_2^2}{\sqrt{S_2^2} \sqrt{S_2^2}} & \dots & \frac{S_{2M}}{\sqrt{S_2^2} \sqrt{S_M^2}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{S_{1M}}{\sqrt{S_1^2} \sqrt{S_M^2}} & \frac{S_{2M}}{\sqrt{S_2^2} \sqrt{S_M^2}} & \dots & \frac{S_M^2}{\sqrt{S_M^2} \sqrt{S_M^2}} \end{bmatrix}$$

$$= \begin{bmatrix} 1 & r_{12} & \dots & r_{1M} \\ r_{12} & 1 & \dots & r_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ r_{1M} & r_{2M} & \dots & 1 \end{bmatrix} \quad (2.3)$$

Apabila semua variabel yang digunakan memiliki unit satuan yang sama maka matriks - yang digunakan adalah matriks varians kovarians. Σ adalah varians kovarians dari variabel asal (Wardani, 2017).

2.3 Analisis Cluster

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. Data mengenai ukuran kesamaan tersebut dapat dianalisis dengan analisis *cluster* sehingga dapat ditentukan objek yang masuk dalam *cluster* (Gudono, 2011).

Analisis *cluster* dibagi menjadi dua metode yaitu metode hirarki (*hierarchical method*) dan metode non hirarki (*non-hierarchical method*). Metode hirarki digunakan bila jumlah kelompok ditentukan berdasarkan hasil analisis. Sedangkan metode non hirarki bertujuan untuk mengelompokkan N observasi ke dalam K *cluster* ($K < N$), nilai K telah ditentukan sebelumnya. Metode hirarki merupakan metode pengelompokan yang terstruktur dan bertahap berdasarkan pada kemiripan sifat antar objek. Kemiripan sifat tersebut dapat ditentukan dari kedekatan jarak (Sarwono, J, 2007).

Hierarchical method dibagi menjadi dua, yaitu metode *agglomerative* (pemusatan) dan metode *divisive* (penyebaran). Metode-metode yang termasuk dalam metode *agglomerative* adalah metode Pautan Tunggal (*Single Linkage*), metode Pautan Lengkap (*Complete Linkage*), metode Antar Pusat (*Centroid Linkage*), metode Pautan Rata-rata (*Average Linkage*) dan metode Ward (*Ward's Method*). Metode yang termasuk *non-hierarchical method* adalah metode *K-Means* dan *K-Medoids* (Rachmatin, 2014). Sedangkan metode *divisive* merupakan proses peng*cluster*an yang didasarkan pada persamaan nilai rata-rata antar objek.

Ciri-ciri *cluster* yang baik yaitu mempunyai (Santoso, S, 2005):

1. Homogenitas (*within-cluster*), yaitu kesamaan yang tinggi antar anggota dalam satu *cluster*. Sebagai contoh, *cluster* konsumen rumah yang peduli lingkungan tentu terdiri dari orang-orang yang mengutamakan kebersihan dan kenyamanan lingkungan rumahnya, lain halnya dengan mereka yang

mengutamakan harga rumah murah tentu tidak dapat digabungkan sebagai anggota *cluster* tersebut.

2. Heterogenitas (*between-cluster*), yaitu perbedaan tinggi antar *cluster* yang satu dengan *cluster* yang lain. Dalam contoh di atas, *cluster* konsumen rumah yang peduli lingkungan tentu memiliki pendapat yang jelas berbeda dengan anggota *cluster* yang mementingkan harga murah.

2.4 Ukuran Jarak *Euclidean*

Tujuan analisis *cluster* adalah mengelompokkan objek yang mirip kedalam *cluster* yang sama. Oleh karena itu memerlukan beberapa ukuran untuk mengetahui seberapa mirip atau berbeda objek-objek tersebut. Pendekatan yang biasa digunakan adalah mengukur kemiripan yang dinyatakan dalam jarak (*distance*) antara pasangan objek. Pada analisis *cluster* terdapat beberapa ukuran jarak untuk mengukur kesamaan antar objek, diantaranya adalah jarak Manhattan (*City Block*), jarak Mahalanobis dan jarak *Euclidean*. Jarak Manhattan mempunyai kelebihan yaitu dapat mendeteksi keberadaan outlier dengan baik. Sedangkan kelebihan dari jarak Mahalanobis adalah memperhitungkan korelasi antar variabel yang mungkin ada sehingga setiap variabel diberikan pembobot yang sama.

Pada penelitian ini akan digunakan ukuran jarak *Euclidean*. Jarak *Euclidean* adalah jarak yang paling umum dan paling sering digunakan dalam analisis *cluster*. Jarak antara dua buah *cluster* diukur sebagai jarak *Euclidean* antara kedua rata-rata (*centroid*) *cluster*. Jarak ini harus memenuhi asumsi bahwa semua variabel yang diamati tidak berkorelasi dan antar variabel memiliki satuan yang sama (Mongi, C. E, 2015). Jarak *Euclidean* antara *cluster* ke-*i* dan ke-*j* dari *M* variabel didefinisikan:

$$d_{ij} = \sqrt{\sum_{p=1}^M (x_{ip} - x_{jp})^2} \quad (2.4)$$

dengan,

d_{ij} : kuadrat jarak *Euclidean* antar objek ke-*i* dengan objek ke-*j*.

M : jumlah variabel *cluster*

x_{ip} : nilai atau data dari objek ke- i pada variabel ke- p

x_{jp} : nilai atau data dari objek ke- j pada variabel ke- p

2.5 Centroid Linkage

Centroid linkage merupakan salah satu metode *hierarchical clustering* yang berguna untuk membentuk kelompok berdasarkan jarak antar *centroid*-nya. *Centroid* adalah rata-rata dari semua anggota dalam *cluster* tersebut. Metode ini dikenal lebih memiliki beban komputasi yang relatif lebih ringan karena hanya cukup menghitung titik tengah antar *cluster* (Kurniawan, 2009).

Metode ini dibangun dengan memperhatikan pengecilan nilai standar deviasi *cluster* sekecil-kecilnya. Metode ini menggabungkan dua *cluster* melalui jarak terdekat di antara titik pusat antar *cluster*. Metode ini sangat ampuh untuk memperkecil *variance within cluster* karena melibatkan titik pusat pada saat penggabungan antar *cluster*. Penggabungan *cluster* pada *centroid linkage* didasarkan pada lokasi titik pusat yang terbentuk pada tahap sebelumnya. Metode ini juga baik untuk data yang mengandung outlier (Karlita, 2006).

Langkah-langkah *Centroid Linkage* (Rachmatin, 2014):

- a. Asumsikan setiap data merupakan *cluster*.
- b. Menghitung jarak antar *cluster* dengan *Euclidean Distance* dengan persamaan:

$$d_{ij} = \sqrt{\sum_{p=1}^M (x_{ip} - x_{jp})^2} \quad (2.5)$$

$i = 1, 2, 3, \dots, N; j = 1, 2, 3, \dots, N$ dan M adalah banyaknya variabel.

- c. Memilih jarak terkecil antar *cluster* lalu menggabungkan kedua objek yang memiliki jarak terkecil tersebut. Misalkan *cluster* U dan *cluster* V memiliki jarak terdekat, maka U dan V bergabung dalam satu *cluster*.
- d. Menghitung *centroid* dari U dan V dengan rumus :

$$X_{(UV)} = \frac{(N_U \times \bar{x}_U) + (N_V \times \bar{x}_V)}{N_U + N_V} \quad (2.6)$$

dengan :

N_U, N_V = Banyak objek pada *cluster* ke U dan V

\bar{x}_U, \bar{x}_V = Rata-rata objek pada *cluster* ke U dan V

- e. Bentuk matriks data baru dengan data dari *cluster* gabungan U dan V yang diperoleh dari langkah keempat.
- f. Ulangi langkah kedua, demikian seterusnya sampai semua data bergabung dengan jumlah *cluster* yang diinginkan.

2.6 *K-Medoids*

K-Medoids juga dikenal sebagai *Partitioning Around Medoids* (PAM) adalah varian dari metode *K-Means*. Hal ini didasarkan pada penggunaan objek yang paling terpusat (*medoid*) bukan dari objek rata-rata (*mean*) yang dimiliki oleh setiap *cluster*, dengan tujuan mengurangi sensitivitas dari partisi sehubungan dengan nilai ekstrim yang ada dalam himpunan data (Vercillis, 2009). *K-Medoids* hadir untuk mengatasi kelemahan *K-Means* yang sensitif terhadap *outlier* karena suatu objek dengan suatu nilai yang besar mungkin secara substansial menyimpang dari distribusi data (Han,J., Kamber, M, 2006).

Algoritma *K-Medoids* menggunakan metode partisi *clustering* untuk mengelompokkan sekumpulan N objek menjadi sejumlah K *cluster*. Algoritma ini menggunakan objek pada kumpulan objek untuk mewakili sebuah *cluster*. Objek yang terpilih untuk mewakili sebuah *cluster* disebut dengan *medoid*. *Cluster* dibangun dengan menghitung kedekatan yang dimiliki antara *medoid* dengan objek *non-medoid* (Setiyawati, 2017).

Langkah-langkah algoritma *K-Medoids* (Dini Marlina, 2018):

- a. Inisialisasi pusat *cluster* sebanyak K (jumlah *cluster*).
- b. Mengalokasikan setiap data (objek) ke *cluster* terdekat menggunakan persamaan ukuran jarak *Euclidean Distance* dengan persamaan:

$$d_{ij} = \sqrt{\sum_{p=1}^M (x_{ip} - x_{jp})^2} \quad (2.7)$$

$i = 1,2,3, \dots, N; j = 1,2,3, \dots, N$ dan M adalah banyaknya variabel.

- c. Memilih secara acak objek pada masing-masing *cluster* sebagai kandidat *medoid* baru.
- d. Menghitung jarak dengan persamaan (2.6) setiap objek yang berada pada masing-masing *cluster* dengan kandidat *medoid* baru.

- e. Menghitung total simpangan (s) dengan menghitung nilai :

$$s = \text{total } distance \text{ baru} - \text{total } distance \text{ lama}$$

Jika $s \geq 0$ maka proses dihentikan, tetapi sebaliknya jika $s < 0$, maka tukar objek dengan data *cluster* untuk membentuk sekumpulan k objek baru sebagai *medoid*.

- f. Mengulangi langkah 3 sampai 5 hingga tidak terjadi perubahan *medoid*, sehingga didapatkan *cluster* beserta anggota *cluster* masing-masing.

2.7 Multikolinearitas

Multikolinearitas adalah suatu peristiwa terjadinya korelasi yang kuat antara dua atau lebih variabel kelompok. Multikolinearitas merupakan masalah yang perlu diperhatikan dalam analisis multivariat pada umumnya, karena pengaruhnya yang sangat besar dalam menghasilkan solusi, sehingga mengganggu proses analisis (Hair, 1998). Sebaiknya data tidak mengandung yang saling berhubungan karena akan menyebabkan hasil yang bias pada variabel yang saling berhubungan. Manakala beberapa variabel saling berkorelasi, korelasi tersebut akan menyebabkan pembobotan yang tidak berimbang sehingga akan mempengaruhi hasil analisis (Puspitasari, M, 2016).

Untuk mendeteksi ada atau tidaknya multikolinearitas dapat dilakukan dengan koefisien korelasi sederhana (Korelasi *Pearson*) antar variabel bebas, jika terdapat nilai yang mencapai atau melebihi 0,8 maka terjadi multikolinearitas (Diah Safitri, 2012). Koefisien korelasi digunakan untuk mengetahui keeratan hubungan. Terdapat lima pedoman yang menunjukkan kuat lemahnya hubungan korelasi variabel, ditunjukkan dengan tabel koefisien korelasi berikut:

Tabel 2. 1 Koefisien Korelasi

Koefisien Korelasi	Kuat Lemahnya Korelasi
$\pm 0,81$ hingga $\pm 1,00$	Sangat Kuat
$\pm 0,61$ hingga $\pm 0,80$	Kuat
$\pm 0,41$ hingga $\pm 0,60$	Sedang
$\pm 0,21$ hingga $\pm 0,40$	Lemah
$\pm 0,00$ hingga $\pm 0,20$	Tidak Ada Hubungan

Sumber : Hair, 1998

2.8 *Principal Component Analysis (PCA)*

Principal Component Analysis (PCA) atau analisis komponen utama merupakan salah satu analisis multivariat yang digunakan untuk mereduksi dimensi data dari yang berukuran besar dan saling berkorelasi menjadi dimensi yang lebih kecil dan tidak saling berkorelasi. *Principal Component (PC)* merupakan suatu kombinasi linear dari variabel-variabel asal. Pembentukan PC berdasarkan dua cara yaitu matriks kovarians atau matriks korelasi (Johnson dan Wichern, 2007).

Tahapan menentukan PC berdasarkan matriks korelasi sebagai berikut:

1. Membuat matriks varian kovarians Σ .
2. Pereduksian PC dimulai dengan mencari nilai eigen $\lambda_1, \lambda_2, \lambda_3, \dots, \lambda_M$ yang diperoleh dari persamaan:

$$|\Sigma - \lambda I| = 0 \quad (2.8)$$

Nilai eigen selalu diurutkan dari yang terbesar sampai terkecil. Nilai eigen menunjukkan besarnya total varian yang dijelaskan oleh PC yang terbentuk. Pasangan nilai eigen dan vektor eigen yang saling ortonormal adalah $(\lambda_1, \gamma_1), (\lambda_2, \gamma_2), \dots, (\lambda_M, \gamma_M)$ dengan $\lambda_1 \geq \lambda_2 \geq \dots \lambda_M \geq 0$ maka PC ke- p didefinisikan sebagai berikut:

$$PC_p = \gamma_{1p}x_1 + \gamma_{2p}x_2 + \dots + \gamma_{Mp}x_M \quad (2.9)$$
$$p = 1, 2, \dots, M$$

Nilai *eigen* dapat menjelaskan besarnya kontribusi keragaman masing-masing komponen utama dalam menjelaskan keragaman data asal. Apabila komponen utama, tahap selanjutnya adalah menghitung skor komponen utama dari setiap objek yang akan digunakan untuk dianalisis lebih lanjut.

Menurut Johnson dan Wichern (2007), banyak komponen utama menjelaskan keragaman data dengan baik dilihat dari proporsi keragaman komponen utama. Banyak komponen utama dikatakan sangat baik apabila proporsi keragaman sebesar 80%, dihitung menggunakan persamaan :

$$Total\ Varian = \frac{\lambda_p}{\sum_{p=1}^M \lambda_p} \times 100\% \quad (2.10)$$

2.9 Penentuan Jumlah *Cluster* Optimum

Penentuan jumlah *cluster* optimum dilakukan dengan pendekatan indeks validitas *silhouette*. Indeks validitas *silhouette* merupakan suatu ukuran statistik yang digunakan untuk menyeleksi permasalahan penentuan jumlah *cluster* optimal yang dapat merepresentasikan grafis singkat seberapa baik setiap objek terletak dalam *cluster*.

Asumsikan data sudah dikelompokkan ke dalam *cluster*. Untuk setiap objek i , misalkan $a(i)$ adalah rata-rata jarak objek i ke semua objek dalam *cluster* yang sama dan $b(i)$ adalah rata-rata jarak minimum objek i ke semua objek pada suatu *cluster* serta i bukan anggota *cluster*. Dari penjelasan yang telah dipaparkan indeks validitas *silhouette* dapat ditulis dengan persamaan sebagai berikut:

$$S(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \quad (2.11)$$

Rata-rata $S(i)$ dari seluruh objek dalam suatu *cluster* menunjukkan seberapa dekat kemiripan objek dalam suatu *cluster* yang juga menunjukkan seberapa tepat objek telah dikelompokkan. Semakin dekat $S(i)$ kepada 1, maka semakin baik pengelompokan objek. Sebaliknya, semakin dekat $S(i)$ kepada -1, maka semakin buruk pengelompokan objek. Jumlah *cluster* K yang optimal merupakan estimasi dari harga K yang paling memaksimalkan nilai rata-rata $S(i)$ dan jika terdapat satu *cluster* yang anggotanya terdiri dari satu objek maka nilai rata-rata $S(i)$ adalah 0 (Nicolaus, 2016).

2.10 Penentuan Metode Terbaik dengan Simpangan Baku

Untuk mengetahui metode mana yang mempunyai kinerja terbaik, dapat digunakan rata-rata simpangan baku dalam *cluster* (S_W) dan simpangan baku antar *cluster* (S_B) (Bunkers, dkk.1996 dalam Laeli, 2014). Rumus rata-rata simpangan baku dalam *cluster* (σ_W):

$$S_W = \frac{1}{K} \sum_{k=1}^K S_k \quad (2.12)$$

dengan,

K = Banyaknya *cluster* yang terbentuk

S_k = Simpangan baku *cluster* ke- k

Rumus simpangan baku *cluster* ke- k (S_k) :

$$S_k = \sqrt{\frac{1}{N_k - 1} \sum_{i=1}^{N_k} (x_i - \bar{x}_k)^2} \quad (2.13)$$

dengan,

N = Jumlah anggota dari setiap *cluster*

\bar{x}_k = Rata-rata *cluster* ke- k

x_i = Anggota *cluster*, dari $i = 1, 2, \dots, N_k$

Rumus simpangan baku antar *cluster* (S_B) :

$$S_B = \left[\frac{1}{K - 1} \sum_{i=1}^K (\bar{x}_k - \bar{x})^2 \right]^{\frac{1}{2}} \quad (2.14)$$

dengan,

\bar{x}_k = Rata-rata *cluster* ke- k

\bar{x} = Rataan keseluruhan *cluster*

Rumus Rasio simpangan baku (S):

$$S = \frac{S_W}{S_B} \times 100\% \quad (2.15)$$

dengan,

S_W = Simpangan baku dalam *cluster*

S_B = Simpangan baku antar *cluster*

Metode yang mempunyai rasio terkecil merupakan metode terbaik. *Cluster* yang baik adalah *cluster* yang mempunyai homogenitas (kesamaan) yang tinggi antar anggota dalam satu *cluster* (*within cluster*) dan heterogenitas yang tinggi antar *cluster* yang satu dengan *cluster* yang lain (*between cluster*) (Santoso, S, 2005).

2.11 Interpretasi *Cluster*

Proses ini dimulai dengan suatu ukuran yang sering digunakan yaitu *centroid cluster*. Membuat profil dan interpretasi *cluster* tidak hanya untuk memperoleh suatu gambaran saja, melainkan untuk menyediakan rata-rata untuk menilai korespondensi pada *cluster* yang terbentuk serta profil *cluster* memberikan arahan bagi signifikansi praktis. Selain itu juga dilakukan untuk

menjelaskan karakteristik dari setiap kelompok berdasarkan dengan tujuan untuk memberi label pada masing-masing kelompok tersebut.

Adapun cara menghitung mean (*centroid*) yaitu (Moh Rizal Rizki, 2013):

$$C_{pk} = \frac{\sum_{i=1}^{N_k} x_{ik}}{N_k} \quad (2.16)$$

dengan,

C_{pk} : Nilai rata-rata (*centroid*) *cluster*

x_{ik} : Nilai atau data dari objek i pada variabel k dalam *cluster* yang dicari nilai *centroid*

N_k : Jumlah objek dalam *cluster* yang dicari nilai *centroid*

2.12 Indikator Pendidikan

Pendidikan merupakan hak asasi setiap warga negara dan untuk itu setiap warga negara Indonesia berhak memperoleh pendidikan yang bermutu sesuai dengan minat dan bakat yang dimilikinya tanpa memandang status sosial, status ekonomi, suku, etnis, agama, dan gender. Adanya pemerataan akses dan peningkatan mutu pendidikan membuat warga negara Indonesia memiliki kecakapan hidup (*life skills*) sehingga mendorong tegaknya pembangunan manusia seutuhnya serta masyarakat madani dan modern yang dijiwai nilai-nilai Pancasila. Indonesia adalah salah satu contoh negara berkembang dengan laju pertumbuhan penduduk yang tinggi namun tingkat pertumbuhan ekonomi masih rendah (Saripudin, 2005).

Pendidikan merupakan salah satu aspek terpenting dalam pembangunan bangsa. Untuk itu setiap bangsa harus menempatkan pembangunan pendidikan sebagai prioritas utama dalam program pembangunan nasional. Dengan demikian, Sumber Daya Manusia (SDM) yang bermutu merupakan produk pendidikan dan kunci keberhasilan suatu negara. Kesempatan memperoleh pendidikan kepada seluruh penduduk baik laki-laki maupun perempuan agar pembangunan dapat dilakukan oleh penduduk dengan kualitas pendidikan yang baik tanpa membedakan jenis kelamin (Badan Pusat Statistik, 2019).

Indikator pendidikan merupakan alat ukur yang digunakan untuk melihat seberapa baik kualitas pendidikan yang sudah ada. Indikator pendidikan menjadi sangat penting mengingat adanya permasalahan yang mendasar dalam pendidikan

seperti angka putus sekolah yang masih cukup tinggi, adanya kesenjangan mendapatkan kesempatan pendidikan, serta kualitas pendidikan yang belum bisa memenuhi kebutuhan lapangan kerja yang semakin kompetitif. Indikator pendidikan yang digunakan dalam penelitian ini adalah penduduk berumur lebih dari 15 tahun menurut tingkat pendidikan dan jenis kelamin (Badan Pusat Statistik, 2019).

Gambaran mengenai peningkatan sumber daya manusia dapat dilihat dari kualitas tingkat pendidikan penduduk usia 10 tahun keatas. Level pendidikan penduduk diketahui dari tingkat pendidikan yang ditamatkan dengan diidentifikasi melalui ijazah/STTB tertinggi yang dimiliki. Indikator ini dapat pula digunakan untuk melihat perkembangan kualitas sumber daya manusia dengan mengetahui level tertinggi pendidikan antar waktu dan wilayah. Semakin tinggi tingkat pendidikan yang ditamatkan maka menggambarkan semakin baik pula kualitas pendidikan manusianya (Badan Pusat Statistik, 2019).

Tingkat Pendidikan Tertinggi adalah jenjang pendidikan tertinggi yang pernah diduduki bagi yang sudah tidak bersekolah atau yang sedang diduduki bagi yang masih bersekolah. Penduduk yang tamat SD diperhitungkan lama sekolah selama 6 tahun, tamat SMP lama sekolah 9 tahun, tamat SMA lama sekolah 12 tahun, tanpa memperhitungkan pernah tinggal kelas atau tidak. Diploma/Perguruan Tinggi termasuk di dalamnya Diploma I/II, Diploma III/Akademi dan Diploma IV/S1/S2/S3.

1. SD/MI, berlaku untuk anggota rumah tangga yang pernah atau sekarang masih duduk di Sekolah Dasar (SD) atau Madrasah Ibtidaiyah (MI).
2. SMP/MTs, berlaku untuk anggota rumah tangga yang pernah atau sekarang masih duduk di Sekolah Menengah Pertama (SMP) yang sederajat Madrasah Tsanawiyah (MTs).
3. SMA/SMK/MA, berlaku untuk anggota rumah tangga yang pernah atau sekarang masih duduk di sekolah menengah atas (SMA) atau yang sederajat Sekolah Menengan Kejuruan dan Madrasah Aliyah.

4. Perguruan Tinggi pada tingkat Diploma (D1/D2/D3), yaitu anggota rumah tangga yang pernah atau sekarang masih mengikuti jenjang pendidikan setelah SMA atau sederajat, biasanya bersifat kejuruan atau keterampilan praktis dengan masa pendidikan sekitar satu sampai empat tahun.
5. Perguruan Tinggi pada tingkat Sarjana (D4/S1), berlaku untuk anggota rumah tangga yang pernah atau sekarang masih mengikuti jenjang pendidikan sarjana penuh. Diploma IV (D4) digolongkan sebagai S1.
6. Perguruan Tinggi pada tingkat Pasca Sarjana (S2/S3), yaitu berlaku untuk anggota rumah tangga yang pernah atau sekarang masih mengikuti jenjang pendidikan pasca sarjana (Badan Pusat Statistik, 2019).