Penerapan Algoritma *K-Medoids* pada Pengelompokan Wilayah Desa atau Kelurahan di Kabupaten Kutai Kartanegara (Studi Kasus : Data Hasil Pendataan Potensi Desa (PODES) Tahun 2018)

The Application of K-Medoids Algorithm to Clustering Village or Political District in Kutai Kartanegara Regency (Case Study: Village Potential Data Collection Results (PODES) in 2018)

Rizky Nur Ibrahim[1], Memi Nor Hayati[2], dan Fidia Deny Tisna Amijaya[3]
[1]Laboratorium Statistika Komputasi FMIPA Universitas Mulawarman
[2] Laboratorium Statistika Terapan, FMIPA Universitas Mulawarman
[3] Laboratorium Matematika Komputasi FMIPA Universitas Mulawarman
Email: rizky.n.ibrahim@gmail.com

Abstract

Kutai Kartanegara Regency (Kukar) was recorded as the largest contributor to the poor population in East Kalimantan (Kaltim) Province in 2017, so that appropriate strategies are needed to solve proverty problems. The development strategy is prioritized for the regions with the largest number of poor people. Identification is conducted based on facilities, infrastructures, access, social, population and economy is provided in the Village Potential data (PODES). K-Medoids is a grouping method that uses representative objects as a central point, which can be used to find out the characteristics of a region. This research is aimed to find out the optimal cluster formed by choosing the largest value of Silhouette Coefficient (SC) from the grouping of villages / political district in Kukar Regency using PODES data in 2018. Clusters that will be formed in this research are 2 clusters, 3 clusters, 4 clusters and 5 clusters. Based on the analysis, it can be seen that the value of SC 2 cluster is 0.430, the value of SC 3 cluster is 0.174, the value of SC 4 cluster is 0.175 and the value of SC 5 cluster is 0.196. So that the largest SC or optimal cluster values obtained in the grouping of 2 clusters with a SC value of 0.430. Cluster 1 consists of 186 villages / political dsitrict and cluster 2 consists of 46 villages / political district.

Keywords: K Medoids, Silhouette Coefficient, Village Potential

Pendahuluan

Data mining adalah suatu proses untuk mendapatkan informasi yang berguna dari gudang basis data berskala besar yang membantu dalam proses pengambilan keputusan. *Data mining* juga dapat digunakan untuk melakukan proses pengelompokan atau klasterisasi (*clustering*), dengan tujuan untuk mengetahui pola universal dari data-data yang ada (Prasetyo, 2012).

Klasterisasi (*clustering*) adalah proses pengelompokan himpunan data ke dalam beberapa kelompok atau klaster sedemikian sehingga objek-objek dalam suatu klaster memiliki kemiripan yang tinggi (homogen), namun sangat berbeda (memiliki ketidakmiripan) dengan objek-objek di klaster lainnya (Suyanto, 2017). Menurut Kauffman dan Rousseeuw (1990), salah satu algoritma dalam analisis klaster adalah algoritma *k-medoids*. Algoritma *k-medoids* dapat mengelompokan data yang mengandung *outlier* (pencilan). Algoritma *k-medoids* merupakan metode berbasis partisi yang menggunakan objek representatif yang disebut *medoids* sebagai titik pusat atau *centroid*.

Analisis klaster dapat digunakan untuk mengetahui karakteristik wilayah yang sama berdasarkan sarana, prasarana, akses, sosial, penduduk serta ekonomi. Data tentang keberadaan, ketersediaan dan perkembangan potensi yang dimiliki setiap wilayah administrasi pemerintah dikumpulkan dalam Potensi Desa (PODES) (Fathia, 2016). Menurut data (BPS, 2017), potensi ekonomi merupakan fokus perhatian dalam menyusun strategi pembangunan. Pembangunan ini diprioritaskan bagi daerah-daerah dengan jumlah penduduk miskin terbesar. Kabupaten Kutai Kartanegara (Kukar) tercatat sebagai penyumbang penduduk miskin terbesar di Provinsi Kalimantan Timur (Kaltim) dengan jumlah penduduk miskin 56.570 jiwa pada Tahun 2017.

Data mining

Data mining merupakan sebuah langkah dalam proses *Knowledge Discovery in Database* (KDD) yang terdiri dari penerapan analisis data dan penemuan algoritma yang menghasilkan enumerasi tertentu terhadap pola pada data. *Data mining* ditujukan untuk mengekstrak (mengambil intisari) pengetahuan dari sekumpulan data sehingga didapatkan struktur yang dapat dimengerti manusia serta meliputi basis data dan manajemen data, prapemrosesan data, pertimbangan model dan inferensi, ukuran ketertarikan, pertimbangan kompleksitas, pascapemrosesan terhadap struktur yang ditemukan, visualisasi, dan *online updating* (Suyanto, 2017).

Menurut Defiyanti (2017), karakteristik *data mining* sebagai berikut:

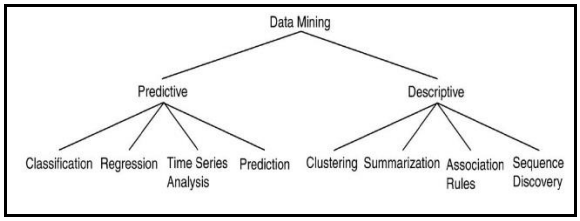
1. *Data mining* berhubungan dengan penemuan sesuatu yang tersembunyi dan pola data tertentu yang tidak diketahui sebelumnya.
2. *Data mining* biasa menggunakan data yang sangat besar. Biasanya data yang besar digunakan untuk membuat hasil lebih dipercaya.
3. *Data mining* berguna untuk membuat keputusan yang kritis, terutama dalam strategi.

Fungsi *Data Mining*

Menurut Dunham (2003), *data mining* melibatkan banyak algoritma yang berbeda untuk menyelesaikan tugas yang berbeda. Algoritma data mining dapat dikarakteristikkan terdiri dari tiga bagian:

1. Model: Tujuan dari algoritma ini adalah untuk menyesuaikan model dengan data.
2. Preferensi: Beberapa kriteria harus digunakan agar sesuai dengan satu model di atas yang lain.
3. Pencarian: Semua algoritma memerlukan beberapa teknik untuk mencari data.

Data mining secara umum dapat diklasifikasikan menjadi dua jenis berdasarkan pada spesifik tugas yang ingin dicapai.



Gambar 1 Model dan Tugas *Data Mining*

Operasi *Data Mining*

Data mining merupakan bagian dari proses *Knowledge Discovery in Database* (*KDD*) yang sering melibatkan aplikasi berulang dari metode *data mining* tertentu. *Data mining* menggunakan model yang cocok untuk menentukan pola dari data yang diamati. Model tersebut berperan menyimpulkan, apakah model tersebut mencerminkan pengetahuan berguna atau menarik. Secara detail *data mining* dalam proses *KDD* disajikan pada Gambar 2.

3. Standardisasi Data

Jika rentang nilai antar objek memiliki perbedaan skala yang cukup besar yang dapat menyebabkan bias dalam analisis klaster, maka data asli perlu dilakukan standardisasi. Standardisasi dapat menyingkirkan atau menghilangkan pengaruh dari unit pengukuran dan dapat memperkecil perbedaan antar kelompok atau klaster (Supranto, 2010). Standardisasi/normalisasi data dapat dilakukan dengan cara semua dimensi atau sub-variabel penyusun ditransformasi ke dalam data standar (nilai rata-rata sama dengan nol, variansi sama dengan satu) (Prasetyo, 2014).

$$\hat{x}_{ij} = \frac{x_{ij} - \overline{x}_j}{s_j} \qquad (1)$$

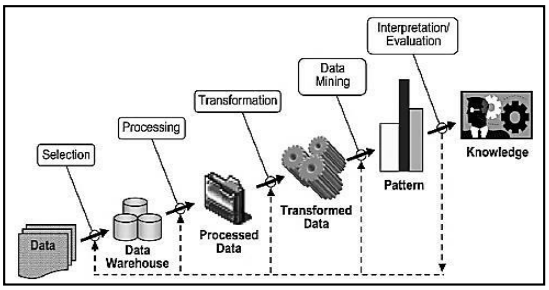
dengan:

x_{ij} : Data ke-*i* variabel ke-*j*

$\overline{x}_j$: Rata-rata pada variabel ke-*j*

S_j : Standar deviasi pada variabel ke-*j*

$\hat{x}_{ij}$: Standardisasi data ke-*i* variabel ke-*j*



Gambar 2 Tahapan Proses *Knowledge Discovery in Database*

4. Analisis Klaster

Analisis klaster (*cluster analysis*) adalah pekerjaan mengelompokkan data (objek) yang didasarkan hanya pada informasi yang ditemukan dalam data yang menggambarkan objek tersebut dan hubungan di antaranya. Tujuannya adalah agar objek-objek yang bergabung dalam sebuah kelompok merupakan objek-objek yang mirip (berhubungan) satu sama lain dan berbeda (tidak berhubungan) dengan objek dalam kelompok lain. Lebih besar kemiripannya (homogenitas) dalam kelompok dan lebih besar perbedaannya di antara kelompok lainnya (Prasetyo, 2012).

Asumsi dalam analisis kelompok yaitu sampel yang diambil harus mewakili populasi (representatif) dan tidak adanya variabel penelitian yang memiliki hubungan linier yang besar dengan variabel lainnya (nonmultikolinieritas). Menurut Gujarati & Porter (2010), jika pada variabel penelitian tersebut terdapat korelasi yang cukup tinggi yaitu di atas 0,8 maka dapat dikatakan adanya gejala multikolinieritas.

$$r_{x_j x_l} = \frac{n\left(\sum_{i=1}^{n} x_{ij} x_{il}\right) - \left(\sum_{i=1}^{n} x_{ij}\right)\cdot\left(\sum_{i=1}^{n} x_{il}\right)}{\sqrt{n\left(\sum_{i=1}^{n} x_{ij}^{2}\right) - \left(\sum_{i=1}^{n} x_{ij}\right)^{2}}\sqrt{n\left(\sum_{i=1}^{n} x_{il}^{2}\right) - \left(\sum_{i=1}^{n} x_{il}\right)^{2}}}, i = 1,2,...,n \qquad (2)$$

dimana

$r_{x_j x_l}$ = Nilai koefisien korelasi antara variabel *x* ke-*j* dan ke-*l*

n = Banyaknya data

5. Algoritma *K-Medoids*

Algoritma *k-medoids* merupakan metode berbasis partisi yang menggunakan objek representatif yang disebut *medoids* sebagai titik pusat atau *centroid*. Algoritma *k-medoids* melakukan partisi dengan cara meminimalkan jumlah ketidakmiripan antara setiap objek *i* dan objek representatif terdekat. Setiap objek yang tersisa dikelompokan dengan objek representatif yang paling mirip dan perhitungan jarak dihitung dari jarak antar masing-masing data (Suyanto, 2017).

Menurut Han & Kamber (2006), algoritma *k-means* tidak efektif terhadap data yang mengandung pencilan (*outlier*). Alternatif lain dari algoritma *k-means* adalah dengan cara mengambil objek representatif dalam sebuah klaster yang merupakan titik terpusat dalam sebuah klaster (*medoids*). Dengan demikian metode berbasis partisi masih dapat dilakukan berdasarkan prinsip meminimalkan jumlah ketidaksamaan antara masing-masing objek dengan objek representatif yang menjadi dasar dari algoritma *k-medoids*.

Adapun tahapan-tahapan dari algoritma *k-medoids* adalah sebagai berikut:

1. Memilih secara acak objek sebanyak *K* sebagai objek representatif o_m (*medoids*).
2. Menghitung jarak *euclidean* untuk setiap objek terhadap masing-masing *medoids* seperti dinyatakan oleh Persamaan (3) sebagai berikut:

$$d(x_{ij}, o_{mj}) = \sqrt{\left(x_{i1} - o_{m1}\right)^2 + \left(x_{i2} - o_{m2}\right)^2 + \ldots + \left(x_{iq} - o_{mq}\right)^2} \quad (3)$$

dengan $d(x_{ij}, o_{mj})$ adalah jarak dari data ke-*i* pada variabel ke-*j* terhadap *medoids* ke-*m* pada variabel ke-*j* dimana *m* = 1,2,...,*K* serta *j*=1,2,...,*q*.

3. Menetapkan setiap objek ke gugus yang sesuai dengan *medoids* terdekat dan menghitung fungsi objektif yang merupakan jumlah ketidakmiripan dari semua objek ke *medoids* terdekat berdasarkan jarak antara objek terhadap setiap *medoids* yang paling minimum.
4. Memilih secara acak objek yang tidak representatif o_h (*non medoids*).
5. Menghitung jarak *euclidean* untuk setiap objek terhadap masing-masing *non-medoids* seperti dinyatakan oleh Persamaan (4) sebagai berikut:

$$d(x_{ij}, o_{hj}) = \sqrt{\left(x_{i1} - o_{h1}\right)^2 + \left(x_{i2} - o_{h2}\right)^2 + \ldots + \left(x_{iq} - o_{hq}\right)^2} \quad (4)$$

dengan $d(x_{ij}, o_{hj})$ adalah jarak dari data ke-*i* pada variabel ke-*j* terhadap *non-medoids* ke-*h* pada variabel ke-*j* dimana *h* = 1,2,...,*n-K* serta *j*=1,2,...,*q*.

6. Menetapkan setiap objek ke gugus yang sesuai dengan *non-medoids* terdekat dan menghitung fungsi objektif yang merupakan jumlah *dissimilarity* dari semua objek ke *non-medoids* terdekat berdasarkan jarak antara objek terhadap setiap *medoids* yang paling minimum.
7. Menghitung selisih dari fungsi objektif dengan cara mengurangkan fungsi objektif *non-medoids* dengan fungsi objektif *medoids*.
8. Mengganti *medoids* o_m dengan *non-medoids* o_h apabila pertukaran semacam mengurangi fungsi objektif.
9. Mengulangi langkah (4-8) sampai tidak ada lagi perubahan objek representatif.
10. Analisis selesai jika sudah tidak terdapat perubahan objek representatif.

6. Validasi Data Hasil Klasterisasi

Salah satu metode evaluasi yang dapat digunakan untuk melihat kualitas dan kekuatan klaster adalah metode *silhouette coefficient*. Metode ini merupakan metode validasi klaster yang menggabungkan metode *cohesion* dan *separation*.

1. Menghitung rata-rata jarak dari suatu data ke-*i* dengan semua data yang berada pada satu klaster yang sama dengan menggunakan Persamaan (5).

$$a_i = \frac{1}{n_p - 1} \sum_{r=1}^{n_p - 1} d_{i,r} \;,\; r \neq i \;, \quad (5)$$

dengan p = 1, 2, ..., *K*.

2. Menghitung rata-rata jarak suatu data ke-*i* dengan semua data yang berada pada klaster yang berbeda dengan menggunakan Persamaan (7), kemudian ambil nilai terkecilnya berdasarkan Persamaan (6)

$$b_i = \min\left\{d_i(p)\right\}, r \neq i \;, \quad (6)$$

dengan rumus jarak suatu data ke-*i* dengan semua data pada klaster yang berbeda adalah

$$d_i(p) = \frac{1}{n_p} \sum_{r=1}^{n_p} d_{i,r} \;, \quad (7)$$

dengan p = 1, 2, ..., *K*.

3. Menghitung nilai *silhouette coefficient* untuk setiap data ke-*i*

$$SC_1(i) = \frac{b_i - a_i}{\max\{a_i, b_i\}}, i = 1, 2, ..., n, \quad (8)$$

Nilai *SC* dari sebuah klaster ($SC_2(p)$) diperoleh dengan menghitung rata-rata nilai $SC_1(i)$ semua data yang bergabung dalam klaster tersebut dengan menggunakan Persamaan (9).

$$SC_2(p) = \frac{1}{n_p} \sum_{x_i \in C_p}^{n_p} SC_1(i) \quad (9)$$

Setelah itu nilai *SC* global diperoleh dengan menghitung rata-rata nilai $SC_2(p)$ dari semua klaster dengan menggunakan Persamaan (10).

$$SC = \frac{\sum_{p=1}^{K} \left(n_p \times SC_2(p) \right)}{\sum_{p=1}^{K} n_p}, \quad (10)$$

dengan

a_i : Rata-rata jarak data ke-*i* dengan semua data pada klaster yang sama
b_i : Rata-rata jarak data ke-*i* dengan semua data pada klaster yang berbeda
$SC_1(i)$: Nilai *silhouette coefficient* pada data ke-i
$SC_1(p)$: Nilai *silhouette coefficient* pada klaster ke-*p*
SC : Nilai *silhouette coefficient global*
X_i : Data pengamatan ke-*i*
C_p : Klaster ke-*p*
n_p : Jumlah data dalam klaster ke-*p*
K : Banyaknya klaster

Nilai *silhouette coefficient* berdasarkan Kauffman dan Rousseuw (1990) dapat dilihat pada Tabel 1.

Tabel 1 Nilai *Silhouette Coefficient*

No	Rentang Nilai *SC*	Keterangan
1.	$0,7 < SC \leq 1$	*Strong Structure*
2.	$0,5 < SC \leq 0,7$	*Medium Structure*
3.	$0,25 < SC \leq 0,5$	*Weak Structure*
4.	$SC \leq 0,25$	*No Structure*

7. Desa Tertinggal

Penentuan desa tertinggal seringkali dilakukan dalam rangka menetapkan penyaluran bantuan pemerintah agar bantuan tersebut dapat disalurkan dengan tepat. Penetapan status desa tertinggal diharapkan menjadi identifikasi daerah kemiskinan. Daerah tertinggal umumnya adalah daerah yang kondisinya relatif kurang berkembang dibandingkan daerah lain dalam skala nasional. Hal tersebut yang dicerminkan oleh empat faktor yang diduga menjadi penyebab kemajuan atau ketertinggalan suatu desa menurut Bappenas (2006) dalam Agusta (2007), yaitu faktor alam atau lingkungan, faktor kelembagaan, faktor sarana/prasarana dan akses serta faktor sosial ekonomi penduduk.

Hasil Penelitian dan Pembahasan

Data yang digunakan dalam penelitian ini menggunakan data PODES di 232 desa/kelurahan yang ada di Kabupaten Kukar pada tahun 2018. Adapun variabel yang digunakan dalam penelitian ini sebanyak 15 variabel yaitu :

X_1 : Kepadatan Penduduk
X_2 : Ketersediaan sarana pendidikan/sekolah
X_3 : Ketersediaan tenaga kesehatan
X_4 : Ketersediaan sarana kesehatan
X_5 : Jumlah berlangganan telepon kabel
X_6 : Jumlah penginapan
X_7 : Jumlah bangunan pasar
X_8 : Jumlah supermarkettoserba
X_9 : Jumlah keberadaan bank
X_{10}: Jumlah pengguna fasilitas perkreditan
X_{11}: Jumlah Koperasi Non KUD lainnya
X_{12}: Jumlah keluarga pengguna listrik PLN
X_{13}: Jumlah keluarga tinggal di bantaran sungai
X_{14}: Jumlah keluarga tinggal di permukiman kumuh
X_{15}: Jumlah penderita gizi buruk

Hasil pengolahan statistika deskriptif data PODES di 232 desa/kelurahan yang ada di Kabupaten Kukar pada tahun 2018 dapat dilihat pada Tabel 2.

Tabel 2 Statistika Deskriptif

Variabel	Banyak Data	Mini mum	Maksi mum	Rata-rata
X_1	232	1	1688	119,86
X_2	232	0	32	6,66
X_3	232	0	29	3,04
X_4	232	0	80	5,07
X_5	232	0	700	5,138
X_6	232	0	9	0,34
X_7	232	0	3	0,39
X_8	232	0	475	36,73
X_9	232	0	6	0,33
X_{10}	232	0	5	0,33
X_{11}	232	0	9	0,48
X_{12}	232	6	7261	915
X_{13}	232	0	1125	62,16
X_{14}	232	0	347	4,13
X_{15}	232	0	4	0,16

Berdasarkan Tabel 2, pada kolom banyak data dapat dilihat bahwa seluruh variabel penelitian memiliki total pengamatan yang sama yaitu 232 data pengamatan. Sebagai contoh kepadatan kependudukan dapat dilihat pada variabel X_1. Kepadatan penduduk tertinggi berasal dari Kelurahan Melayu, Kecamatan Tenggarong yaitu sebesar 1.688 jiwa/km^2. Sedangkan kepadatan penduduk terendah berasal dari Kecamatan Tabang yaitu Desa Muara Belinau yaitu sebesar 1 jiwa/km^2. Rata-rata kepadatan penduduk di

Kabupaten Kukar adalah 119,86 jiwa. Demikian seterusnya untuk data yang lain.

Dari hasil pada analisis 2 *cluster*, 3 *cluster*, 4 *cluster* dan 5 *cluster* didapatkan nilai *SC* masing-masing dapat dilihat pada Tabel 3:

Tabel 3 Perbandingan Hasil Validasi Klaster Berdasarkan Nilai *SC Global*

Jumlah Klaster	Klaster	Jumlah Anggota	*SC*
2	1	186	0,430
	2	46	
3	1	143	0,174
	2	18	
	3	71	
4	1	126	0,175
	2	22	
	3	18	
	4	66	
5	1	118	0,196
	2	8	
	3	22	
	4	18	
	5	66	

Tabel 3 terlihat bahwa penerapan algoritma *k-medois* pada masing-masing klaster menghasilkan jumlah klaster optimum dengan di bentuk menjadi 2 klaster karena memiliki *SC* yang lebih besar dibandingkan dengan hasil *SC* pada pembentukan 3 klaster, 4 kalster dan 5 klaster, yang menunjukkan bahwa nilai *SC* yang didapatkan dapat digunakan sebagai pendukung keputusan untuk menilai jumlah klaster yang paling cocok digunakan.

Tabel 4 Nilai Rata-Rata Variabel untuk Masing-Masing Klaster

Variabel	Klaster ke-*i*	
	1	**2**
X_1 (Kepadatan penduduk)	77	295
X_2 (Sekolah)	5	12
X_3 (Tenaga kesehatan)	2	7
X_4 (Sarana kesehatan)	2	13
X_5 (Telepon kabel)	1	18
X_6 (Penginapan)	0	1
X_7 (Pasar semi/permanen)	0	1
X_8 (Supermerket/toserba)	26	81
X_9 (Bank)	0	1
X_{10} (Fasilitas perkreditan)	0	0
X_{11} (Koperasi)	0	0
X_{12} (Pengguna listrik PLN)	653	2002
X_{13} (Tinggal di bantaran tepi sungai)	58	76
X_{14} (Tinggal di permukiman kumuh)	1	15
X_{15} (Penderita gizi buruk)	0	0

Tabel 4 terlihat bahwa setelah dilakukan pengelompokkan wilayah desa/kelurahan di Kabupaten Kutai Kartanegara dengan menggunakan algoritma *k-medoids* dapat diketahui rata-rata nilai variabel pada klaster 1 lebih kecil dibandingkan dengan rata-rata nilai variabel pada klaster 2.

Kesimpulan

Berdasarkan hasil penelitian dan pembahasan, maka kesimpulan yang dapat diambil adalah sebagai berikut

1. Klaster optimal yang terbentuk pada pengelompokkan desa/kelurahan di Kabupaten Kukar berdasarkan indikator desa tertinggal dengan menggunakan metode *K-Medoids* adalah sebanyak 2 klaster yaitu klaster 1 dan klaster 2. Klaster 1 beranggotakan 186 desa/kelurahan dan klaster 2 beranggotakan 46 desa/kelurahan.
2. Nilai *Silhouette Coefficient* untuk validasi data hasil *clustering* wilayah desa/kelurahan di Kabupaten Kukar berdasarkan indikator desa tertinggal dengan menggunakan metode *K-Medoids* yaitu pada 2 klaster dengan nilai sebesar 0,430 yang menyatakan bahwa struktur klaster yang dihasilkan pada pengelompokkan ini adalah *weak structure*.

Daftar Pustaka

Agusta, I. (2007). Desa Tertinggal di Indonesia. *Jurnal Transdisiplin Sosiologi, Komunikasi dan Ekologi Manusia*, 1(2), 233-235.

Alwi. (2018). Analisis Klaster Untuk Pengelompokan Kabupaten/Kota di Provinsi Sulawesi Selatan Berdasarkan Indikator Kesejahteraan Rakyat. *Jurnal MSA, Vol.6, NO. 1 ED. JAN-JUNI 2018*

Defiyanti. (2017). Optimalisasi *K-Medoid* dalam Pengklasteran Mahasiswa Pelamar Beasiswa dengan *Cubic Clustering Criterion*. *Jurnal TEKNOSI, Vol.03, No. 01*.

Dunham, Margareth H. (2003). *Data Mining Introductory and Advanced Topics*.New Jersey: Prentice Hall

Fathia. (2016). Analisis Klaster Kecamatan Di Kabupaten Semarang Berdasarkan Potensi Desa Menggunakan Metode Ward dan Single Linkage. *Jurnal Gaussian, Volume 5, Nomor 4, ISSN:239-2541*.

Han, J., & Kamber, M. (2006). *Data Mining: Concept and Techniques*. San Fransisco: Morgan Kauffman Publisher.

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding Groups in Data*. New York: John Willey & Sons.

Prasetyo, E. (2012). *Data Mining: Konsep dan Aplikasi menggunakan Matlab*. Yogyakarta: Andi Offset.

Suyanto. (2017). *Data Mining untuk Klasifikasi dan Klasterisasi Data*. Bandung : Informatika.