

BAB III

K-MEANS CLUSTERING

3.1 Analisis Klaster

Analisis klaster merupakan salah satu teknik multivariat metode interdependensi (saling ketergantungan). Oleh karena itu, dalam analisis klaster tidak ada perbedaan antara variabel bebas (*independent variable*) dan variabel terikat (*dependent variable*).

Analisis klaster adalah teknik yang digunakan untuk menggabungkan observasi ke dalam kelompok atau klaster (Sharma, 1996:185), sedemikian sehingga:

1. Setiap kelompok atau klaster homogen mempunyai karakteristik tertentu. Hal ini berarti bahwa observasi dalam setiap kelompok sama dengan observasi lain dalam satu kelompok yang sama;
2. 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.

Analisis klaster digunakan untuk mengelompokkan data observasi yang hanya berdasarkan pada informasi yang ditemukan dalam data, di mana data tersebut harus menggambarkan observasi dan hubungannya. Oleh karena itu, tujuan dari analisis ini adalah observasi dalam satu kelompok mirip satu sama lain dan berbeda dari observasi dalam kelompok lain. Semakin besar kemiripan

(homogenitas) dalam kelompok dan semakin besar perbedaan (heterogenitas) antar kelompok maka klastering akan lebih baik atau lebih berbeda (Tan *et al*, 2006:490).

Dalam analisis klaster, pengelompokan observasi ke dalam klaster dilakukan dengan menggunakan teknik-teknik yang berawal dari kemiripan antar semua pasangan observasi. Kemiripan ini didasarkan pada beberapa ukuran jarak. Metode lain dalam pengelompokan dapat menggunakan pilihan awal sebagai pusat klaster atau perbandingan di dalam dan antar variabilitas klaster. Selain itu, pengelompokan juga dapat menggunakan variabel klaster yang kemiripannya didasarkan pada matriks korelasi (Rencher, 2002:451).

Pada prinsipnya analisis klaster merupakan proses untuk mereduksi sejumlah objek yang besar menjadi lebih sedikit yang disebut klaster. Analisis klaster digunakan oleh peneliti yang belum mengetahui anggota dari suatu kelompok. Analisis klaster disebut juga *Q-analysis*, *classification analysis*, pengenalan pola (*pattern recognition*), analisis segmentasi (*numerical taxonomy*).

Berdasarkan paparan tersebut, terdapat dua langkah utama dalam analisis klaster yaitu memilih ukuran kemiripan dan memilih algoritma dalam pembentukan klaster.

3.2 Metode Pengelompokan

Dalam analisis klaster, terdapat banyak metode untuk mengelompokkan observasi ke dalam klaster. Secara umum metode pengelompokan dalam analisis klaster dibedakan menjadi metode hirarki (*Hierarchical Clustering Method*) dan metode non hirarki (*Nonhierarchical Clustering Method*). Metode hirarki

digunakan apabila belum ada informasi jumlah kluster yang dipilih. Sedangkan metode non hirarki bertujuan untuk mengelompokkan n objek ke dalam k kluster ($k < n$), di mana nilai k telah ditentukan sebelumnya (Prayitno, 2007:5).

3.2.1 Metode Hirarki

Pada dasarnya metode ini dibedakan menjadi dua metode pengelompokan, yaitu:

1. Metode Aglomeratif

Proses pengelompokan dengan pendekatan metode aglomeratif (*Down to Top*) dimulai dengan n kluster sehingga masing-masing kluster memiliki tepat satu objek, kemudian tentukan dua kluster terdekat dan gabungkan kluster tersebut menjadi satu kluster baru. Proses penggabungan dua kluster diulangi sampai diperoleh satu kluster yang memuat semua himpunan data. Perlu diperhatikan bahwa setiap penggabungan dalam metode ini selalu diikuti dengan perbaikan matriks jarak. Hasil analisis kluster dari metode ini dapat disajikan dalam bentuk dendogram.

2. Metode Divisif

Proses pengelompokan dengan pendekatan metode divisif (*Top to Down*) dimulai dengan n objek yang dikelompokkan menjadi satu kluster, kemudian kluster tersebut dipartisi ke dalam dua kluster pada setiap langkah sampai diperoleh n kluster dengan setiap kluster memiliki satu objek.

3.2.2 Metode Non-Hirarki

Dalam metode ini data dibagi dalam k partisi, setiap partisi mewakili sebuah klaster. Secara umum proses metode non-hirarki sebagai berikut:

1. Pilih k centroid klaster awal atau *seed*, di mana k merupakan jumlah klaster yang diinginkan.
2. Tempatkan setiap observasi ke dalam klaster yang terdekat.
3. Tempatkan kembali setiap observasi ke dalam k klaster menurut aturan penghentian yang sudah ditentukan.
4. Proses berhenti jika tidak ada observasi yang berpindah lagi, jika belum ulangi langkah kedua.

Beberapa algoritma non- hirarki berbeda dalam aturan untuk memperoleh centroid klaster awal dan aturan yang digunakan untuk menempatkan kembali observasi. Beberapa aturan yang digunakan untuk memperoleh *seed* awal antara lain:

1. Pilih k observasi pertama dengan tidak ada data yang hilang sebagai centroid atau *seed* klaster awal.
2. Pilih observasi pertama dengan tidak ada data yang hilang sebagai *seed* klaster pertama, lalu *seed* klaster kedua dipilih dari observasi yang mempunyai jarak terjauh dari sebelumnya, dan seterusnya.
3. Pilih secara random k observasi dengan tidak ada data yang hilang sebagai pusat klaster atau *seed*.
4. Perbaiki *seed* yang dipilih dengan menggunakan aturan tertentu sehingga jarak *seed* tersebut sejauh mungkin.

5. Gunakan *heuristic* tentang identifikasi pusat kluster sehingga jarak pusat kluster tersebut sejauh mungkin.
6. Gunakan *seed* yang disediakan oleh peneliti.

Setelah *seed* diidentifikasi, kluster awal yang dibentuk akan menempatkan kembali $n - k$ observasi sisanya ke dalam *seed* yang terdekat dengan observasi tersebut.

Beberapa algoritma non hirarki juga berbeda terkait dengan prosedur yang digunakan dalam penempatan kembali observasi ke dalam k kluster. Adapun aturan penempatan kembali observasi sebagai berikut:

1. Hitung centroid setiap kluster dan tempatkan kembali observasi ke dalam kluster berdasarkan centroid terdekat. Centroid tidak *update* sementara menempatkan kembali setiap observasi ke dalam k kluster, centroid dihitung ulang setelah penempatan kembali semua observasi yang telah dibuat. Jika perubahan dalam centroid kluster lebih besar daripada kriteria konvergensi yang dipilih maka penempatan kembali setiap observasi terus dilakukan. Proses penempatan kembali dilanjutkan hingga perubahan centroid kurang dari kriteria konvergensi yang dipilih.
2. Hitung centroid setiap kluster dan tempatkan kembali observasi ke dalam kluster berdasarkan centroid terdekat. Untuk penempatan kembali setiap observasi, hitung ulang centroid kluster di mana observasi ditempatkan dan kluster dari mana observasi ditempatkan.

Sekali lagi penempatan kembali dilanjutkan hingga perubahan centroid kluster kurang dari kriteria konvergensi yang dipilih.

3. Tempatkan kembali observasi sedemikian sehingga beberapa fungsi objektif diminimumkan. Metode ini biasa disebut sebagai metode mendaki bukit (*hill climbing methods*). Beberapa fungsi objektif yang diminimumkan antara lain:
 - a. Trace matriks $SSCP_w$
 - b. Determinan matriks $SSCP_w$
 - c. Trace matriks $(SSCP_w)^{-1}SSCP_b$
 - d. Nilai eigen terbesar dari matriks $(SSCP_w)^{-1}SSCP_b$

Pada dasarnya, algoritma non-hirarki dibedakan atas teknik *partitioning*, *overlapping* dan *hybrid*. Sebelum membahas *partitioning* sebagai dasar metode *K-Means*, secara singkat akan dibahas *overlapping* dan *hybrid*.

Overlapping terjadi apabila data tumpang tindih sehingga suatu objek dapat termasuk ke dalam beberapa kluster. Dalam teknik ini data mempunyai nilai keanggotaan (*membership*). Sedangkan *hybrid* merupakan teknik penggabungan antara metode hirarki dan non-hirarki.

Dalam pendekatan *partitioning*, observasi dibagi ke dalam k kluster tanpa menggunakan matriks jarak di antara semua pasangan titik seperti pada pendekatan hirarki. Metode ini disebut juga metode optimisasi (*optimization methods*). Secara umum teknik ini dimulai dengan penentuan k titik dalam ruang berdimensi p (R^p) untuk menentukan estimasi awal pusat kluster (Everitt, 1974:25).

Misalkan C_j merepresentasikan j kluster dengan jumlah anggota $m_j \geq 2$ dan misalkan $C_k \neq \emptyset, C_k \neq C_j$, dan $C_k \subset C_j$, kemudian misalkan $C_p = C_j - C_k \neq \emptyset$, diberikan oleh

$$\bar{x}_q = \frac{1}{m_q} \sum_{i \in C_q} x_i, q = j, k, p \quad (3.1)$$

persamaan (3.1) didefinisikan sebagai centroid dan

$$e_q = \sum_{i \in C_q} \|x_i - \bar{x}_q\|^2, q = j, k, p \quad (3.2)$$

persamaan (3.2) merupakan jumlah kuadrat jarak dari titik ke centroid. Persamaan (3.2) dapat disederhanakan menjadi:

$$\begin{aligned} e_q &= \sum_{i \in C_q} (x_i - \bar{x}_q)' (x_i - \bar{x}_q) \\ &= \sum_{i \in C_q} \{ \|x_i\|^2 - 2x_i' \bar{x}_q + \|\bar{x}_q\|^2 \} \\ &= \sum_{i \in C_q} \|x_i\|^2 - 2(x_1' \bar{x}_q + x_2' \bar{x}_q + \dots + x_{m_q}' \bar{x}_q) + \sum_{i \in C_q} \|\bar{x}_q\|^2 \\ &= \sum_{i \in C_q} \|x_i\|^2 - 2(x_1' + x_2' + \dots + x_{m_q}') \bar{x}_q + m_q \|\bar{x}_q\|^2 \\ &= \sum_{i \in C_q} \|x_i\|^2 - 2m_q \bar{x}_q' \bar{x}_q + m_q \|\bar{x}_q\|^2 \\ &= \sum_{i \in C_q} \|x_i\|^2 - 2m_q \|\bar{x}_q\|^2 + m_q \|\bar{x}_q\|^2 \\ &= \sum_{i \in C_q} \|x_i\|^2 - m_q \|\bar{x}_q\|^2 \end{aligned} \quad (3.3)$$

Centroid x_p dan jumlah kuadrat jarak e_p dapat direpresentasi sebagai fungsi dari

\bar{x}_j , \bar{x}_k , e_j , dan e_k , sehingga

$$\bar{x}_p = \frac{m_j \bar{x}_j - m_k \bar{x}_k}{m_j - m_k} \quad (3.4)$$

Bukti:

$$\begin{aligned} \bar{x}_p &= \frac{1}{m_p} \sum_{i \in C_p} x_i \\ &= \frac{1}{m_j - m_k} \left(\sum_{i \in C_j} x_i - \sum_{i \in C_k} x_i \right) \\ &= \frac{1}{m_j - m_k} (m_j \bar{x}_j - m_k \bar{x}_k) \\ &= \frac{m_j \bar{x}_j - m_k \bar{x}_k}{m_j - m_k} \end{aligned}$$

dan

$$e_p = e_j - e_k - \frac{m_j m_k}{m_j - m_k} \|\bar{x}_j - \bar{x}_k\|^2 \quad (3.5)$$

Bukti:

$$\begin{aligned} e_p &= \sum_{i \in C_p} \|x_i\|^2 - m_p \|\bar{x}_p\|^2 \\ &= \sum_{i \in C_j} \|x_i\|^2 - \sum_{i \in C_k} \|x_i\|^2 - \left(m_j - m_k \left\| \frac{m_j \bar{x}_j - m_k \bar{x}_k}{m_j - m_k} \right\|^2 \right) \\ &= \sum_{i \in C_j} \|x_i\|^2 - \sum_{i \in C_k} \|x_i\|^2 - \left(\frac{1}{m_j - m_k} \|m_j \bar{x}_j - m_k \bar{x}_k\|^2 \right) \\ &= (e_j + m_j \|\bar{x}_j\|^2) - (e_k + m_k \|\bar{x}_k\|^2) - \frac{1}{m_j - m_k} \|m_j \bar{x}_j - m_k \bar{x}_k\|^2 \end{aligned}$$

$$\begin{aligned}
&= e_j - e_k + \left(m_j - \frac{m_j^2}{m_j - m_k} \right) \|\bar{x}_j\|^2 - \left(m_k + \frac{m_k^2}{m_j - m_k} \right) \|\bar{x}_k\|^2 \\
&\quad + \frac{2m_j m_k}{m_j - m_k} \bar{x}_j' \bar{x}_k \\
&= e_j - e_k - \frac{m_j m_k}{m_j - m_k} \left(\|\bar{x}_j\|^2 + \|\bar{x}_k\|^2 - 2\bar{x}_j' \bar{x}_k \right) \\
&= e_j - e_k - \frac{m_j m_k}{m_j - m_k} \|\bar{x}_j - \bar{x}_k\|^2
\end{aligned}$$

Untuk kasus khusus, di mana $C_k = \{k\}$, persamaan (3.4) dan (3.5) menjadi

$$\bar{x}_p = \frac{m_j \bar{x}_j - \bar{x}_k}{m_j - 1} \quad (3.6)$$

$$e_p = e_j - \frac{m_j m_k}{m_j - 1} \|\bar{x}_j - \bar{x}_k\|^2 \quad (3.7)$$

3.3 Metode *K-Means*

Metode *K-Means* diperkenalkan oleh James B MacQueen pada tahun 1967 dalam *proceedings of the 5th berkeley symposium on Mathematical Statistics and Probability* (Johnson, 1998:555). Dasar pengelompokan dalam metode ini adalah menempatkan objek berdasarkan rata-rata (mean) kluster terdekat. Oleh karena itu, metode ini bertujuan untuk meminimumkan *error* akibat partisi n objek ke dalam k kluster. *Error* partisi disebut juga sebagai fungsi objektif.

Misalkan $X = \{x_i\}, i = 1, 2, \dots, n$ merupakan titik-titik dalam ruang berdimensi n (R^n) dan titik tersebut dikelompokkan ke dalam i kluster, $C_i, i = 1, 2, \dots, K$. Misalkan c_i centroid dari kluster C_i sehingga jumlah kuadrat antara c_i dan titik di dalam kluster yaitu x_i , didefinisikan sebagai:

$$J(c_k) = \sum_{x_i \in C_i} (c_i - x_i)^2 \quad (3.8)$$

Prinsip dasar metode *K-Means* adalah meminimumkan jumlah kuadrat *error* dari seluruh i klaster, yaitu:

$$SSE = \sum_{i=1}^K \sum_{x_i \in C_i} (c_i - x_i)^2 \quad (3.9)$$

3.3.1 Komponen *K-Means*

Algoritma *K-Means* memerlukan 3 komponen yaitu:

1. Jumlah Klaster K

Seperti yang telah dijelaskan sebelumnya, *K-Means* merupakan bagian dari metode non-hirarki sehingga dalam metode ini jumlah K harus ditentukan terlebih dahulu. Jumlah klaster K dapat ditentukan melalui pendekatan metode hirarki. Namun perlu diperhatikan bahwa tidak terdapat aturan khusus dalam menentukan jumlah-klaster K , terkadang jumlah klaster yang diinginkan tergantung pada subjektif seseorang.

2. Klaster Awal

Klaster awal yang dipilih berkaitan dengan penentuan pusat klaster awal (centroid awal). Dalam hal ini, terdapat beberapa pendapat dalam memilih klaster awal untuk metode *K-Means* sebagai berikut:

- a. Berdasarkan Hartigan (1975), pemilihan klaster awal dapat ditentukan berdasarkan interval dari jumlah setiap observasi.
- b. Berdasarkan Rencher (2002), pemilihan klaster awal dapat ditentukan melalui pendekatan salah satu metode hirarki.
- c. Berdasarkan Teknomo (2007), pemilihan klaster awal dapat secara acak dari semua observasi.

Oleh karena adanya pemilihan kluster awal yang berbeda ini maka kemungkinan besar solusi kluster yang dihasil akan berbeda pula.

3. Ukuran Jarak

Dalam hal ini, ukuran jarak digunakan untuk menempatkan observasi ke dalam kluster berdasarkan centroid terdekat. Ukuran jarak yang digunakan dalam metode *K-Means* adalah jarak Euclid.

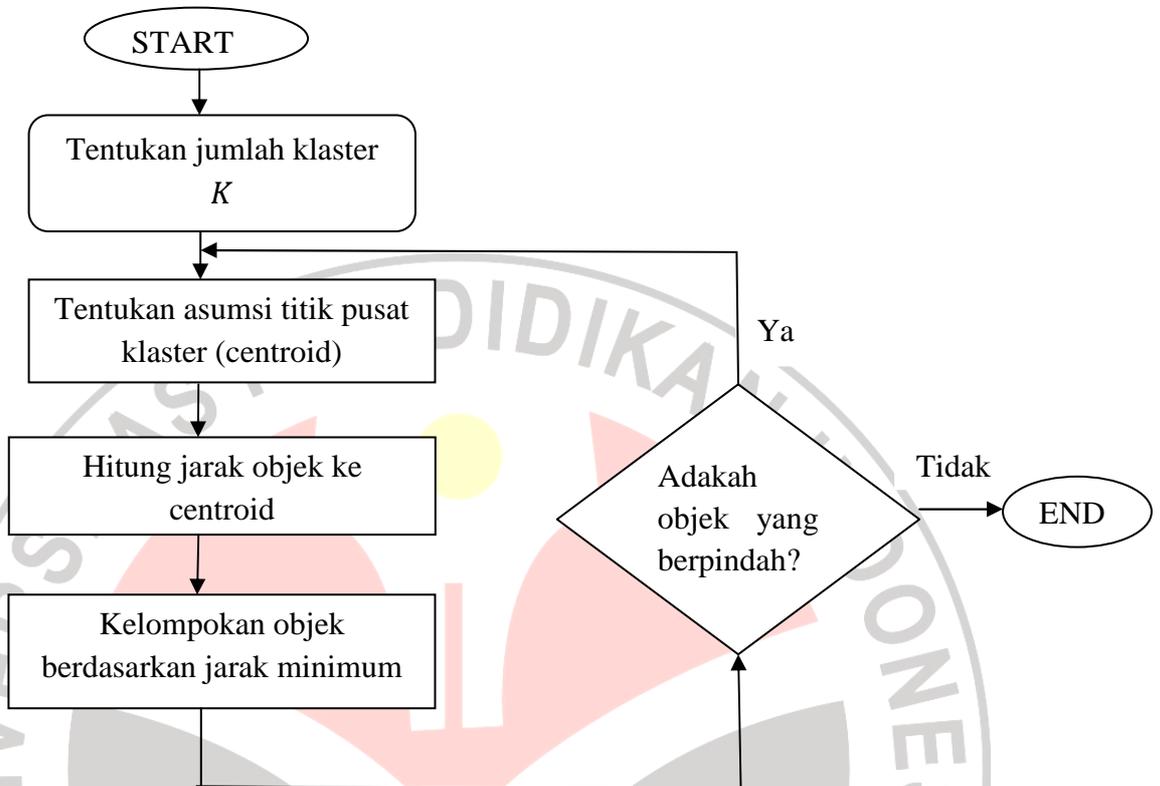
3.3.2 Algoritma *K-Means*

Adapun algoritma *K-means* dalam pembentukan kluster sebagai berikut:

1. Misalkan diberikan matriks data $X = \{x_{ij}\}$ berukuran $n \times p$ dengan $i=1,2,\dots,n, j=1,2,\dots,p$ dan asumsikan jumlah kluster awal K
2. Tentukan centroid.
3. Hitung jarak setiap objek ke setiap centroid dengan menggunakan jarak Euclid atau dapat ditulis sebagai berikut:

$$d(x_i, c_i) = \sqrt{(x_i - c_i)^2}$$
4. Setiap objek disusun ke centroid terdekat dan kumpulan objek tersebut akan membentuk kluster.
5. Tentukan centroid baru dari kluster yang baru terbentuk, di mana centroid baru itu diperoleh dari rata-rata setiap objek yang terletak pada kluster yang sama.
6. Ulangi langkah 3, jika centroid awal dan baru tidak sama.

Secara umum, algoritma di atas dapat disusun dalam gambar 3.1 berikut:

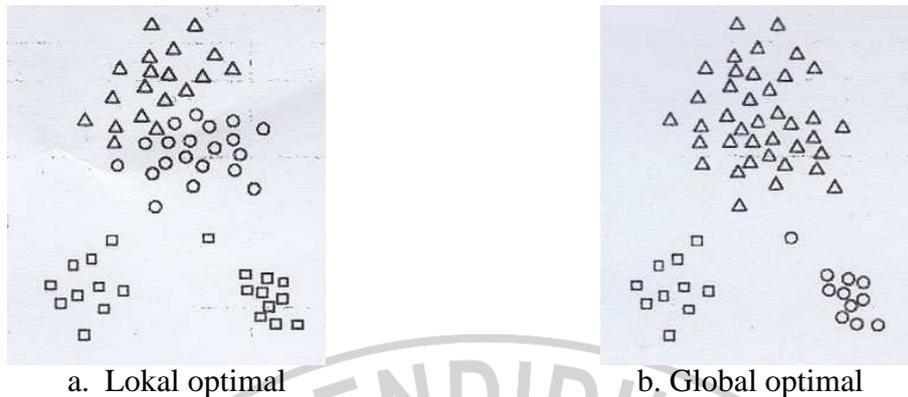


Gambar 3.1 Algoritma *K-Means*

3.3.3 Batas Minimum Lokal dan Optimum Global

Metode hirarki tidak efisien digunakan dalam mengelompokkan observasi dalam jumlah besar. Untuk mengatasi masalah tersebut maka dianjurkan agar pengelompokan menggunakan metode *K-Means*. Metode ini dapat mengelompokkan data berukuran diatas 200 sampel.

Disamping kelebihan tersebut, *K-Means* memiliki kelemahan terutama solusi kluster yang hanya mencapai lokal optimum daripada global optimum seperti pada gambar 3.2. Hal ini sangat dipengaruhi oleh pemilihan centroid secara acak.



Gambar 3.2 Solusi dari $K = 3$

Misalkan n objek dipartisi ke dalam k kluster sedemikian sehingga jarak Euclid setiap pasangan objek dalam satu kluster lebih kecil dari ρ dan jarak Euclid setiap pasangan objek dalam kluster yang berbeda lebih besar dari ρ . Maka partisi tersebut akan menghasilkan solusi kluster yang hanya mencapai lokal optimum dan sebaliknya.

3.3.4 K -Means Sebagai Masalah Optimisasi

Seperti telah dijelaskan sebelumnya bahwa metode K -Means sebagai salah satu metode optimisasi yang bertujuan meminimumkan *error* partisi (fungsi objektif). Dalam hal ini akan ditentukan *update* centroid kluster terbaik yang dapat meminimumkan fungsi objektif tersebut. Untuk meminimumkan (3.9) akan dicari centroid ke k (c_k) dan turunan parsialnya disamakan dengan nol, diperoleh

$$\begin{aligned} \frac{\partial}{\partial c_k} SSE &= \frac{\partial}{\partial c_k} \sum_{i=1}^K \sum_{x_i \in C_i} (c_i - x_i)^2 \\ &= \sum_{i=1}^K \sum_{x_i \in C_i} \frac{\partial}{\partial c_k} (c_i - x_i)^2 \end{aligned}$$

$$\begin{aligned}
&= \sum_{x_i \in C_k} \frac{\partial}{\partial c_k} (c_k - x_k)^2 \\
&= 2 \sum_{x_i \in C_k} (c_k - x_k) = 0 \\
&= 2 \sum_{x_i \in C_k} c_k - 2 \sum_{x_i \in C_k} x_k = 0 \\
&= \sum_{x_i \in C_k} c_k = \sum_{x_i \in C_k} x_k \\
&= m_k c_k = \sum_{x_i \in C_k} x_k \\
&= \frac{1}{m_k} \sum_{x_i \in C_k} x_k
\end{aligned}$$

di mana m_k adalah jumlah observasi dalam kluster ke k . Dengan demikian, centroid terbaik sebenarnya rata-rata dari kluster itu sendiri.

3.4 Validasi Kluster

Validasi merupakan proses untuk menilai hasil algoritma kluster. Oleh karena itu, proses ini bertujuan untuk menjamin bahwa solusi kluster yang dihasilkan dalam analisis kluster dapat menggambarkan populasi sebenarnya.

Gordon (Yatskiv dan Gusarova, 2005:75) mengatakan bahwa terdapat 3 pendekatan utama dalam melakukan validasi kluster yaitu:

1. *External test*, dalam uji ini data dibagi menjadi dua bagian. Solusi kluster dari data hasil analisis kluster dibandingkan dengan solusi

klaster dari data yang tidak diikutsertakan dalam analisis klaster tersebut.

2. *Internal test*, dalam uji ini solusi klaster digunakan untuk melihat kualitas klaster dengan cara membandingkan solusi klaster hasil metode hirarki dan metode iteratif (non-hirarki).
3. *Relative test*, dalam uji ini beberapa solusi klaster yang berbeda dari data dibandingkan menggunakan algoritma yang sama dengan parameter yang berbeda.

Pada dasarnya, validasi memberikan informasi tentang ketepatan jumlah klaster yang telah dipilih. Dalam hal ini, jumlah klaster yang terbentuk dikatakan baik apabila solusi klaster yang dihasilkan tidak jauh berbeda berdasarkan pendekatan yang digunakan. Validasi dengan pendekatan *internal test* lebih sering digunakan dalam praktik analisis klaster disebabkan pendekatan ini lebih sederhana dan mudah.

Dalam tugas akhir ini, validasi lebih ditekankan pada pendekatan *relative test* dengan statistik uji sebagai berikut:

1. *Root Mean Square Standard Devition (RMSTD)*

RMSTD adalah simpangan baku gabungan dari semua variabel yang membentuk klaster, dan didefinisikan sebagai:

$$RMSTD = \sqrt{\frac{\sum_{i=1}^{n_c} \sum_{j=1}^{r_i} (x_{ij} - \bar{x}_i)^2}{\sum_{i=1}^{n_c} (r_i - 1)}} \quad (3.10)$$

nilai RMSTD berada pada interval $(0, \infty)$, dan nilai RMSTD yang kecil mengindikasikan adanya homogenitas yang tinggi dalam klaster.

2. Root Square (RS)

RS adalah rasio dari SS_b dan SS_t , dan didefinisikan sebagai:

$$RS = \frac{SS_b}{SS_t} = \frac{SS_t - SS_w}{SS_t} = \frac{\{\sum_{j=1}^n (X_j - \bar{X})^2\} - \{\sum_{i=1}^{n_c} \sum_{j=1}^{r_i} (X_{ij} - \bar{X})^2\}}{\sum_{j=1}^n (X_j - \bar{X})^2} \quad (3.11)$$

Nilai RS berada pada interval (0,1), dengan 0 berarti bahwa tidak ada perbedaan antar kluster dan 1 berarti bahwa terdapat perbedaan yang maksimum antar kluster. Dengan demikian haruslah nilai RS besar karena mengindikasikan heterogenitas antar kluster.

3. Davies Bouldin Indeks (IDB)

IDB bertujuan untuk memaksimalkan jarak antara kluster (*inter cluster*) yang satu dengan kluster yang lain (*separation value*) dan meminimumkan jarak antara titik (*intra cluster*) dalam sebuah kluster (*compactness value*). IDB didefinisikan sebagai:

$$DB = \frac{1}{n_c} \sum_{i=1}^{n_c} R_i \quad (3.12)$$

di mana $R_i = \sum_{j=1 \dots n_c, i \neq j} \max \frac{s_i + s_j}{d_{ij}}$,

perhatikan bahwa jarak *inter cluster* (d_{ij}), didefinisikan sebagai:

$$d_{ij} = \sqrt{\sum_{k=1}^m (c_{ik} - c_{jk})^2} \quad (3.13)$$

dengan c_i dan c_j adalah centroid kluster i dan j . Sedangkan jarak *intra cluster* (s_i), didefinisikan sebagai:

$$s_i = \frac{1}{r_i} \sqrt{\sum_{x \in c_i} (x - c_i)^2} \quad (3.14)$$

Nilai IDB berada pada interval $(0,1)$, nilai minimum dari IDB akan menunjukkan jumlah kluster optimal.

