BAB II

LANDASAN TEORI

Pada bab ini akan dibahas teori-teori mengenai analisis klaster (kelompok), analisis korelasi, analisis komponen utama, lalu akan dijelaskan juga mengenai jarak *Square Euclidean*, pengelompokan menggunakan metode *bottom-up* dan *top-down*, serta metode *Hybrid Mutual Clustering* menggunakan jarak *Square Euclidean*. Sebagai awalan, akan dijelaskan mengenai analisis kelompok.

2.1 Analisis Kelompok

Analisis kelompok merupakan suatu analisis multivariat yang digunakan untuk mengelompokan objek pengamatan menjadi beberapa kelompok berdasarkan ukuran kemiripan antar objek, sehingga objek-objek yang berada dalam satu kelompok memiliki kemiripan yang lebih homogen dibandingkan objek dari kelompok yang berbeda (Johnson & Wichern, 2002). Seperti diketahui, analisis kelompok akan membagi sejumlah data pada satu atau beberapa *cluster* tertentu. Sebuah *cluster* yang baik adalah *cluster* yang mempunyai ciri sebagai berikut:

1. Homogenitas (kesamaan) yang tinggi antara anggota dalam satu *cluster* atau biasa disebut *within cluster*. Sebagai contoh, konsumen restoran yang mengutamakan interior ruangan tentu terdiri dari orang-orang yang mengutamakan kebersihan dan kenyamanan restoran. Mereka yang mengutamakan harga makanan yang murah tidak dapat digabungkan menjadi 'anggota' *cluster* tersebut.

2. Heterogenitas (perbedaan) yang tinggi antara *cluster* satu dengan *cluster* yang lain atau biasa disebut *between cluster*. Dalam contoh sebelumnya, anggota dari *cluster* konsumen restoran yang mengutamakan interior ruangan tentu mempunyai pendapat yang jelas berbeda dengan anggota *cluster* konsumen yang mengutamakan harga makanan yang murah.

Dari dua hal di atas dapat disimpulkan bahwa ciri sebuah *cluster* yang baik adalah *cluster* yang mempunyai anggota yang semirip mungkin satu sama lain, tetapi sangat tidak mirip dengan anggota *cluster* yang lain. Kata 'mirip' diartikan sebagai tingkat kesamaan karakteristiknya.

Proses *clustering* pada dasarnya mencari dan mengelompokkan data yang mirip satu dengan yang lain, maka kriteria mirip (*similarity*) adalah dasar dari metode *clustering*. Proses pengolahan data sehingga sekumpulan data mentah dapat dikelompokkan menjadi satu atau beberapa *cluster* adalah sebagai berikut:

1. Menetapkan ukuran jarak antar data. Mengukur kesamaan antar objek (*similarity*). Sesuai prinsip dasar *cluster* yang mengelompokkan objek yang mempunyai kemiripan, maka proses pertama adalah mengukur seberapa jauh ada kesamaan antar objek. Ada 2 metode yang digunakan, yaitu mengukur korelasi antar sepasang objek variabel dan mengukur jarak antara dua objek.

2. Melakukan proses standardisasi data jika diperlukan. Setelah cara mengukur jarak ditetapkan, yang juga perlu diperhatikan adalah apakah satuan data mempunyai perbedaan yang besar. Sebagai contoh, jika variabel penghasilan mempunyai satuan juta, sedangkan usia seseorang hanya mempunyai satuan puluhan, maka perbedaan yang mencolok ini akan membuat perhitungan jarak menjadi tidak valid.

3. Melakukan proses *clustering*. Setelah data yang dianggap mempunyai satuan yang sangat berbeda sudah diseragamkan, langkah selanjutnya adalah membuat *cluster*. Proses inti dari *clustering* adalah pengelompokan data yang bisa dilakukan dengan 2 metode, yaitu metode hirarki dan metode non hirarki.

Lalu terdapat pula asumsi pada analisis kelompok diantaranya adalah sampel yang diambil benar-benar bisa mewakili populasi yang ada dan kemungkinan adanya korelasi antar objek. Jika terdapat korelasi, maka dianjurkan untuk melakukan analisis komponen utama yang akan dijelaskan pada subbab berikutnya.

2.2 Analisis Korelasi

Analisis korelasi mencoba mengukur keeratan hubungan antara dua peubah melalui sebuah bilangan yang disebut koefisien korelasi. Ukuran hubungan linear antara dua peubah diduga dengan koefisien korelasi dirumuskan sebagai berikut (Walpole, 1995):

$$r = \frac{n\sum_{i=1}^{n} X_{1i}X_{2i} - \left(\sum_{i=1}^{n} X_{1i}\right)\left(\sum_{i=1}^{n} X_{2i}\right)}{\sqrt{\left[n\sum_{i=1}^{n} X_{1i}^2 - \left(\sum_{i=1}^{n} X_{1i}\right)^2\right]\left[n\sum_{i=1}^{n} X_{2i}^2 - \left(\sum_{i=1}^{n} X_{2i}\right)^2\right]}}. \tag{2.1}$$

Keterangan:

r = koefisien korelasi

X_{1i} = amatan ke-i pada kelompok pertama

X_{2i} = amatan ke-i pada kelompok kedua

n = jumlah amatan.

Analisis kelompok tidak dapat dilakukan jika terdapat korelasi antar peubah, sehingga dilakukan analisis komponen utama dengan tujuan membentuk

peubah-peubah baru yang tidak saling berkorelasi, yang akan dibahas pada subbab berikut.

2.3 Analisis Komponen Utama

Analisis Komponen Utama (*Principal Component Analysis*) adalah analisis multivariat yang mentransformasi variabel-variabel asal yang saling berkorelasi menjadi variabel-variabel baru yang tidak saling berkorelasi dengan mereduksi sejumlah variabel tersebut sehingga mempunyai dimensi yang lebih kecil namun dapat menerangkan sebagian besar keragaman variabel aslinya.

Banyaknya komponen utama yang terbentuk sama dengan banyaknya variabel asli. Pereduksian (penyederhanaan) dimensi dilakukan dengan kriteria persentase keragaman data yang diterangkan oleh beberapa komponen utama pertama. Apabila beberapa komponen utama pertama telah menerangkan lebih dari 75 % keragaman data asli, maka analisis cukup dilakukan sampai dengan komponen utama tersebut.

Bila komponen utama diturunkan dari populasi multivariat normal dengan random vektor $X = \left(\overline{X1}, \overline{X2}, ..., \overline{Xp}\right)$ dan vektor rata-rata $\mu = (\mu_1, \mu_2, ..., \mu_p)$ dan matriks kovarians $\mathbf{\Sigma}$ dengan akar ciri *(eigenvalue)* yaitu $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$ didapat kombinasi linier komponen utama yaitu sebagai berikut (Johnson & Wichern, 2002).

$$Y_1 = e_1'X = e_{11}'X_1 + e_{21}'X_2 + ... + e_{p1}'X_p$$

$$Y_2 = e_2'X = e_{12}'X_1 + e_{22}'X_2 + ... + e_{p2}'X_p$$

$$\vdots$$

$$Y_p = e_p'X = e_{1p}'X_1 + e_{2p}'X_2 + ... + e_{pp}'X_p$$

Maka $Var(Y_i) = ei'\Sigma ei$ dan $Cov(Y_i, Y_k) = ei'\Sigma ei$ dimana $i, k = 1, 2, ..., p$.

Syarat untuk membentuk komponen utama yang merupakan kombinasi linear dari variabel X agar mempunyai varian maksimum adalah dengan memilih vektor ciri *(eigen vector)* yaitu $e = (e_1, e_2, ..., e_p)$ sedemikian hingga $Var(Y_i) = ei'\Sigma ei$ maksimum dan $ei'ei = 1$. Pembentukan komponen utama dijelaskan seperti berikut:

1. Komponen utama pertama adalah kombinasi linear $e_1'X$ yang memaksimumkan $Var(e_1'X)$ dengan syarat $e_1'e_1 = 1$.

2. Komponen utama kedua adalah kombinasi linear $e_2'X$ yang memaksimumkan $Var(e_2'X)$ dengan syarat $e_2'e_2 = 1$.

3. Komponen utama ke-i adalah kombinasi linear $e_i'X$ yang memaksimumkan $Var(e_i'X)$ dengan syarat $e_i'e_k = 1$ dan $Cov(e_i'e_k) = 0$ untuk $k < 1$.

Antar komponen utama tersebut tidak berkorelasi dan mempunyai variasi yang sama dengan akar ciri dari $\mathbf{\Sigma}$ merupakan varian dari komponen utama $\mathbf{Y}$, sehingga matriks ragam peragam dari $\mathbf{Y}$ adalah:

$$\sum = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_p \end{bmatrix}.$$

Total keragaman variabel asal akan sama dengan total keragaman yang diterangkan oleh komponen utama yaitu:

$$\sum_{j=1}^{p} \text{var}(X_i) = tr(\mathbf{\Sigma}) = \lambda_{\mathbf{1}} + \lambda_{\mathbf{2}} + \cdots + \lambda_{\mathbf{p}} = \sum_{\mathbf{j=1}}^{\mathbf{p}} \text{var}(\mathbf{Y_i}).$$

Penyusutan dimensi dari variabel asal dilakukan dengan mengambil sejumlah kecil komponen yang mampu menerangkan bagian terbesar keragaman

data. Apabila komponen utama yang diambil sebanyak q komponen, dimana $q < p$, maka proporsi dari keragaman total yang bisa diterangkan oleh komponen utama ke-i adalah:

$$\frac{\lambda_i}{\lambda_1 + \lambda_2 + \cdots + \lambda_p}; i = 1, 2, \ldots, p.$$

Penurunan komponen utama dari matriks korelasi dilakukan apabila data sudah terlebih dahulu ditransformasikan ke dalam bentuk baku $\mathbf{Z}$. Transformasi ini dilakukan terhadap data yang satuan pengamatannya tidak sama. Bila variabel yang diamati ukurannya pada skala dengan perbedaan yang sangat lebar atau satuan ukurannya tidak sama, maka variabel tersebut perlu dibakukan *(standardized)*. Variabel baku ($\mathbf{Z}$) didapat dari transformasi terhadap variabel asal dalam matriks berikut:

$$Z = \left(V^{1/2}\right)^{-1}(X - \mu).$$

$\mathbf{V}^{1/2}$ adalah matriks simpangan baku dengan unsur diagonal utama adalah ${a_{ii}}^{1/2}$ sedangkan unsur lainnya adalah nol. Nilai harapan $E(\mathbf{Z}) = 0$ dan keragamannya adalah:

$$Cov(Z) = \left(V^{1/2}\right)^{-1}\Sigma\left(V^{1/2}\right)^{-1} = \rho.$$

Dengan demikian komponen utama dari $\mathbf{Z}$ dapat ditentukan dari vektor ciri yang didapat melalui matriks korelasi variabel asal ρ. Untuk mencari akar ciri dan menentukan vektor pembobotnya sama seperti pada matriks $\mathbf{\Sigma}$. Sementara *trace* matriks korelasi ρ akan sama dengan jumlah p variabel yang dipakai.

Salah satu tujuan dari analisis komponen utama adalah mereduksi dimensi data asal yang semula terdapat p variabel bebas menjadi q komponen utama (dimana $q < p$). Kriteria pemilihan q yaitu:

1. Proporsi kumulatif keragaman data asal yang dijelaskan oleh q komponen utama minimal 80%, dan proporsi total variansi populasi bernilai cukup besar (Johnson, 2002).

2. Dengan menggunakan *scree plot* yaitu plot antara i dengan λ_i, berdasarkan *scree plot* ditentukan dengan melihat letak terjadinya belokan dengan menghapus komponen utama yang menghasilkan beberapa nilai eigen kecil membentuk pola garis lurus (Johnson, 2002).

Contoh 2.3.1. Berikut akan diberikan data tentang skor jawaban benar dari 20 butir soal yang diujikan kepada 8 peserta tes sebagai berikut:

Tabel 2.1: Hasil nilai Tes A dan Tes B

Peserta	Tes A	Tes B
1	2	9
2	16	20
3	8	0
4	18	11
5	10	13
6	4	4
7	10	17
8	12	16

Analisis yang akan dilakukan adalah analisis komponen utama untuk mengetahui apakah peserta tes harus menghadapi kedua tes tersebut atau cukup salah satunya saja. Langkah pertama yang dilakukan adalah melihat nilai korelasi yang bertujuan untuk mencari keeratan hubungan antar 2 peubah. Untuk mencari nilai korelasi menggunakan persamaan pada (2.1), maka hasilnya adalah sebagai berikut:

$$r = \frac{8 \cdot 1044 - (80 \cdot 90)}{\sqrt{[(8 \cdot 1008) - 80^2][(8 \cdot 1332) - 90^2]}}$$
$$r = 0,56$$

Dari hasil perhitungan koefisien korelasi di atas, menunjukkan nilai koefisiennya kecil dan kurang signifikan, sehingga perlu dilakukan analisis komponen utama, dengan langkah sebagai berikut

Tabel 2.2: Hasil nilai Tes A dan Tes B

Peserta	Tes A	Tes B
1	2	9
2	16	20
3	8	0
4	18	11
5	10	13
6	4	4
7	10	17
8	12	16

Didapat matriks kovarians

$$\Sigma = \begin{bmatrix} 29,7142 & 20,5714 \\ 20,5714 & 45,6429 \end{bmatrix}$$

Dari matriks kovarians didapat nilai eigen

$$\begin{aligned} \lambda_1 &= 59,7379 \\ \lambda_2 &= 15,6193 \end{aligned}$$

Kemudian didapat vektor eigen berdasarkan λ_1 dan λ_2 sebagai berikut

$$\begin{aligned} e_1^{'} &= \begin{bmatrix} 0,5652 & 0,8249 \end{bmatrix} \\ e_2^{'} &= \begin{bmatrix} 0,8249 & -0,5652 \end{bmatrix} \end{aligned}$$

Dari vektor eigen di atas, maka didapat persamaan komponen utamanya

$$\begin{aligned} Y_1 &= 0,56X_1 + 0,82X_2 \\ Y_2 &= 0,82X_1 - 0,56X_2 \end{aligned}$$

Proporsi total variansi dari matriks kovarians yang dijelaskan komponen utama adalah

$$\frac{\lambda_1}{\lambda_1+\lambda_2} = 0,8$$
$$\frac{\lambda_2}{\lambda_1+\lambda_2} = 0,2$$

Terlihat dari data di atas $\lambda_1 = 59,7379$ persen sedangkan proporsi total variansi pertama yaitu 80 persen, sehingga dapat disimpulkan cukup diperlukan satu tes saja karena kontribusinya sudah cukup besar dalam menggambarkan keragaman data.

Setelah penjelasan teori tentang analisis klaster, analisis korelasi, dan analisis komponen utama, berikutnya kita akan mulai memasuki metode yang akan digunakan pada skripsi ini yaitu metode *Hybrid Mutual Cluster* yang akan dijelaskan pada subbab selanjutnta

2.4 Mutual Cluster

Mutual cluster adalah suatu pengelompokan yang menggunakan jarak terbesar antara pasangan dalam kelompok yang lebih kecil dari jarak terpendek ke setiap titik di luar kelompok. Hal ini berarti bahwa jarak maksimal antar obyek dalam sebuah mutual cluster lebih kecil dibandingkan jarak minimal beberapa obyek di luar mutual cluster. Data yang terkandung dalam sebuah mutual cluster tidak pernah dipisahkan (Chipman dan Tibshirani, 2006). Metode tersebut memiliki beberapa implikasi dalam sebuah *mutual cluster*. Implikasi yang paling jelas adalah untuk mendukung gagasan bahwa dalam sebuah *mutual cluster* berisi informasi pengelompokan yang kuat, tidak peduli pendekatan *linkage* mana yang digunakan. Hal ini dapat membantu dalam interpretasi metode *bottom-up*. Informasi tambahan tesebut dapat membantu dalam interpretasi dari *mutual cluster*, atau dalam menentukan keputusan untuk pembagian kelompok. *Hybrid* ini juga mempertahankan metode *top-down*

yang akurat membagi data menjadi pengelompokan yang baik. Tahapan awal akan dilakukan pengelompokan secara *bottom-up*. Jarak obyek satu dengan obyek yang lain dihitung yang selanjutnya akan dicari jarak terdekat (minimal). Kelompok inilah yang menjadi *mutual cluster* pertama. Setelah itu, jarak antara kelompok yang terbentuk dengan obyek yang lain dihitung kembali. Lalu jarak terdekat (minimal) juga dicari kembali. Langkah tersebut terus dilakukan sampai semua obyek bergabung menjadi satu kelompok besar. Penentuan *mutual cluster* harus memiliki jarak minimal antar obyek di luar *mutual cluster*. Pada tahapan ini *mutual cluster* yang telah terbentuk harus dipertahankan. Oleh karena itu, obyek-obyek tersebut akan dibagi 2 kelompok. Selanjutnya, koordinat dari pusat kelompok *(means)* masing-masing kelompok ditentukan. Kemudian jarak masing-masing obyek dari koordinat pusat dihitung dan kembali menentukan obyek ke kelompok terdekat. Jika obyek dipindahkan dari posisi awal, pusat kelompok harus diperbaharui sebelum diproses lebih lanjut.

2.5 Metode Pengelompokkan Objek

2.5.1 Metode Pengelompokan Hirarki

1. **Single Linkage**

   Input pada algoritma *single linkage* dapat berupa jarak atau kesamaan antara pasangan-pasangan objek. Grup dibentuk dari kesatuan individu dengan menggabungkan tetangga terdekatnya, dimana kata "tetangga terdekat" mengandung arti jarak terkecil atau kesamaan terbesar (terbanyak).

   Sebagai langkah awal kita harus menemukan jarak terkecil pada $D =$

$\{d_{ik}\}$ dan menggabungkan objek-objek yang saling berkorespondensi, katakanlah U dan V, untuk mendapatkan kelompok (UV). Jarak antara (UV) dan kelompok yang lainnya, katakanlah W, dihitung dengan cara

$$d_{(UV)W} = min\{d_{UW}, d_{VW}\}$$

Di sini, nilai d_{UW} dan d_{VW} adalah jarak antara tetangga terdekat dari kelompok U dan W serta kelompok V dan W, begitupun sebaliknya.

2. Complete Linkage

   Prosedur pengelompokan *complete linkage* hampir sama dengan *single linkage*, dengan satu pengecualian. Pada setiap tingkat, jarak (kesamaan) antar kelompok ditentukan dengan jarak (kesamaan) antara dua elemen, satu dari setiap kelompok, yakni yang paling jauh. Dengan demikian *complete linkage* menjamin bahwa dalam seluruh item pada kelompok terdapat jarak maksimum (atau kesamaan minimum).

   Algoritma aglomeratif umum dimulai dengan menemukan entri (elemen) dalam $D = \{d_{ik}\}$ dan menggabungkan objek yang berkorespondensi, misalkan U dan V, untuk membentuk kelompok (UV). Pada langkah ketiga, dalam algoritma umum, jarak (UV) dan kelompok lainnya, misalkan W ditentukan sebagai berikut:

   $$d_{(uv)w} = max\{d_{uw}, d_{vw}\}$$

   di mana d_{uw} dan d_{vw} merupakan jarak terjauh antara anggota kelompok U dan W serta kelompok V dan W, begitupun sebaliknya.

3. Average Linkage

   Perhitungan *Average Linkage* didasarkan pada rata-rata jarak dari seluruh objek pada suatu cluster dengan seluruh objek pada kelompok lain.

Algoritma yang digunakan dalam *Average Linkage* hampir sama dengan algoritma *agglomerative hierarchical clustering*. Dimulai dengan mencari jarak dari matriks, yaitu $D = \{d_{ik}\}$.

Untuk mencari objek terdekat, sebagai contoh U dan V, objek ini digabung ke dalam bentuk kluster UV. Untuk tahap ketiga, jarak antara UV dan kluster W adalah:

$$d_{(uv)w} = \frac{\sum_i \sum_j d_(ik)}{N_{(uv)} N_W}$$

di mana d_{ik} adalah jarak antara objek I pada kluster (UV) dan objek k pada kluster $W, N_{(UV)}$ dan N_W adalah jumlah item-item pada kluster (UV) dan W.

2.5.2 Metode Pengelompokkan non-Hirarki

Metode Pengelompokkan hirarki digunakan apabila belum ada informasi jumlah kelompok. Sedangkan metode pengelompokkan nonhirarki bertujuan pengelompokkan n objek ke dalam k kelompok ($k < n$). Salah satu pengelompokkan pada non hirarki adalah dengan menggunakan metode K-Means.

Metode ini merupakan metode pengelompokkan yang bertujuan pengelompokkan objek sedemikian sehingga jarak tiap-tiap objek ke pusat kelompok di dalam satu kelompok adalah kelompok minimum. Algoritma K-Means adalah:

1. Tentukan jumlah *cluster*.
2. Cari data yang lebih dekat dengan pusat *cluster*. Hitung jarak dari masing-masing objek dari pusat *cluster*. Tentukan kembali pusat *cluster*.
3. Ulangi langkah 2 sampai tidak ada yang berpindah posisi.

2.6 Jarak *Square Euclidean*

Jarak *Square Euclidean* merupakan jarak yang dikembangkan dari jarak *Euclidean*. Pada jarak *Euclidean*, jarak tersebut mempunyai tiga asumsi yang diantaranya adalah antar peubah tidak saling berkorelasi, memiliki satuan pengukuran yang sama, dan pengukuran pembakuan mempunyai rata-rata nol dan standar deviasi satu. Jarak *Euclidean* merupakan jarak antar objek, misalkan dua objek ke-i dan ke-j yang berada pada p dimensi dimana formulanya sebagai berikut (Johnson & Wichern, 2002):

$$d_{(X_i,X_j)} = \sqrt{\sum_{q=1}^{p} (X_{iq} - X_{jq})^2},$$

Keterangan:

$d_{(X_i,X_j)}$ = jarak antar objek pada X_i dengan objek pada X_j; $X_i \neq X_j$

q = banyak peubah;$1, 2, ..., p$

X_{iq} = nilai dari obyek X_i pada variabel ke-q

X_{jq} = nilai dari obyek X_j pada variabel ke-q

sedangkan pada jarak *Square Euclidean* yang merupakan pengembangan dari jarak *Euclidean* dapat diartikan sebagai suatu ukuran kesamaan jumlah kuadrat perbedaan tanpa akar kuadrat. Jarak *Square Euclidean* antara dua unit/observasi yang berdimensi p dengan koordinat $X_iq = (X_{i1}, X_{i2}, ..., X_{ip})$ dan $X_jq = (X_{j1}, X_{j2}, ..., X_{jp})$, dengan $X_i \neq X_j$. Formula jarak *Square Euclidean* adalah sebagai berikut (Hair dkk., 2010):

$$d_{(X_i,X_j)} = \sum_{q=1}^{p} (X_{iq} - X_{jq})^2.$$

Keterangan:

$d_{(X_i,X_j)}$ = kuadrat jarak antar objek pada X_i dengan objek pada X_j; $i \neq j$

q = banyak peubah;$1, 2, ..., p$

X_{iq} = nilai dari obyek X_i pada variabel ke-q

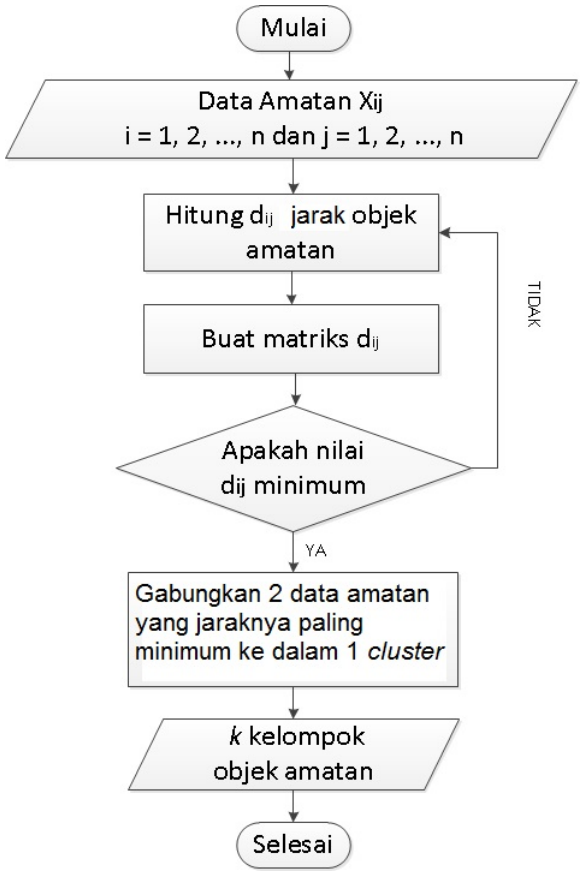
X_{jq} = nilai dari obyek X_j pada variabel ke-q

Jarak *Square Euclidean* digunakan dalam pengelompokan metode *bottom-up* dan *top-down*. Berikut akan dijelaskan tentang metode *bottom-up* dan *top-down*.

2.7 Pengelompokan Metode *Bottom-Up*

Pengelompokan dengan menggunakan metode *bottom-up* adalah suatu metode hierarki dimana n buah kelompok digabungkan menjadi satu kelompok tunggal. Metode *bottom-up* ini meletakkan setiap objek data sebagai sebuah kelompok tersendiri (*atomic cluster*) yang selanjutnya kelompok-kelompok tersebut bergabung menjadi kelompok besar sampai akhirnya semua objek menyatu dalam sebuah kelompok tunggal. Jarak antar objek diperlukan pada tahap awal dalam penggabungan 2 kelompok dengan metode *agglomerative* (Hair dkk., 2010).

Algoritma *bottom-up* dimulai dari menginput data amatan X_{ij} yang kemudian dihitung jarak objek amatan (d_{ij}). Selanjutnya buat matriks d_{ij} yang kemudian dilihat apakah nilai d_{ij} minumum, ulangi hitung nilai jarak objek amatan apabila nilai d_{ij} belum minimum. Setelah dapat nilai d_{ij} minimum, gabungkan objek amatan yang sama menjadi satu kelompok. Output berupa k kelompok objek amatan. Berikut adalah diagram alir pengelompokan metode *bottom-up (agglomerative)*:



Gambar 2.1: Diagram alir pengelompokan metode *bottom-up (agglomerative)*.

Contoh 2.7.1. Sebagai contoh dapat dilihat pada kasus berikut ini. Akan diberikan pengelompokan data hasil nilai renang mahasiswa sebagai berikut:

Tabel 2.3: Matriks Jarak Hasil Nilai Renang Mahasiswa

	M1	M2	M3	M4
M1	0	7	2	5
M2	7	0	5	2
M3	2	5	0	3
M4	5	2	3	0

Dari tabel di atas, berikut ini merupakan penyelesaian pengelompokan nilai dengan menggunakan metode klasterisasi hirarki *bottom-up* dengan *agglomerative*.

1. Tahap 1

Tabel 2.4: Matriks Jarak Hasil Nilai Renang Mahasiswa

	M1	M2	M3	M4
M1	0	7	2	5
M2	7	0	5	2
M3	2	5	0	3
M4	5	2	3	0

Dilihat dari matriks jarak diatas, jarak terkecil berada di M1 dan M3, yang kemudian akan dilakukan penggabungan M1 dan M3

2. Tahap 2

Tabel 2.5: Matriks Setelah Penggabungan M1 dan M3

	M1/M3	M2	M4
M1/M3	0	5	3
M2	5	0	2
M4	3	2	0

Setelah M1 dan M3 digabungkan menjadi satu, jarak terkecil sekarang berada di M2 dan M4. Maka selanjutnya akan digabungkan M2 dan M4.

3. Tahap 3

Tabel 2.6: Matriks Setelah Penggabungan M1/M3 dan M2/M4

	M1/M3	M2/M4
M1/M3	0	3
M2/M4	3	0

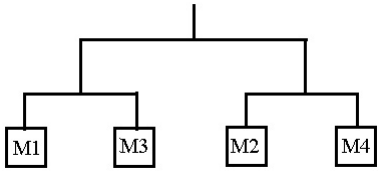
Setelah itu, jarak terkecil sekarang berada di M1/M3 dan M2/M4, gabungkan keduanya.

4. Tahap 4

Tabel 2.7: Matriks Setelah Penggabungan M1/M3/M2/M4

	M1/M2/M3/M4
M1/M2/M3/M4	0

Dari penyelesaian di atas maka dapat digambarkan pengelompokan tersebut menggunakan dendogram sebagai berikut:



Gambar 2.2: Dendogram Data Hasil Nilai Mahasiswa.

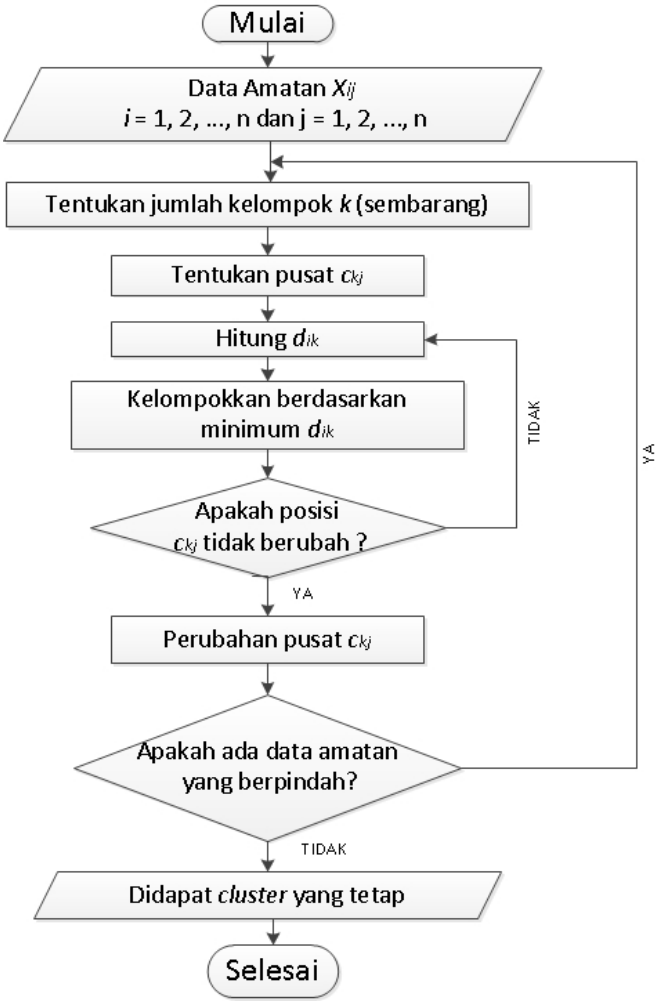
Dapat dilihat dari dendogram di atas disimpulkan mahasiswa 1 dan mahasiswa 3 berada dalam 1 *cluster*, dan mahasiswa 2 dan mahasiswa 4 berada dalam 1 *cluster*. Dalam pengelompokan metode *Hybrid Mutual Clustering*, selain metode *bottom-up*, dilakukan juga pengelompokan dengan metode *top-down*, sehingga pada subbab selanjutnya akan dijelaskan secara rinci mengenai pengelompokan metode *top-down*.

2.8 Pengelompokan Metode *Top-Down*

Pengelompokan dengan metode *top-down* adalah membagi n objek ke dalam k kelompok yang bertujuan untuk mengelompokkan objek sehingga jarak antar objek ke pusat kelompok di dalam satu kelompok minimum.

Proses pertama dalam mengelompokkan dengan menggunakan metode *top-down* yang bersifat non-hierarki *(k-means)* adalah terdapat data amatan X_i dan X_j. Kemudian, partisikan obyek ke dalam k kelompok. Langkah selanjutnya, hitung pusat kelompok dimana pusat kelompok itu sendiri merupakan

rata-rata dari keseluruhan obyek yang berasa dalam kelompok tersebut. Setelah itu, hitung jarak setiap obyek ke pusat kelompok dengan menggunakan jarak *Square Euclidean*. Jika terdapat obyek yang berpindah dari posisi awal, maka pusat kelompok dihitung kembali dan periksa kembali posisi obyek. Ulangi langkah-langkah tersebut sampai tidak ada obyek yang berpindah posisi. Perhitungan berhenti ketika obyek sudah tidak berpindah posisi dan membentuk kelompok *Hybrid Mutual Clustering*. Berikut adalah contoh pengelompokan dengan menggunakan metode *top-down (k-means)*:



Gambar 2.3: Diagram alir pengelompokan metode *top-down (k-means)*.

Contoh 2.8.1. Berikut ini merupakan data padi pada tahun 2013 di provinsi Jawa Timur, data yang digunakan hanya data produksi dan luas lahan (sumber: https://syafrudinmtop.blogspot.co.id/):

Tabel 2.8: Data Produksi dan Luas Lahan di Provinsi Jawa Timur Tahun 2013

No.	Kota /Kab	Luas Lahan	Produksi
1	Ponorogo	66,693	402,047
2	Trenggalek	31,136	182,848
3	Tulungagung	49,23	259,581
4	Blitar	50,577	289,494
5	Kediri	51,083	281,392
6	Malang	65,597	464,498
7	Lumajang	72,552	387,168
8	Jember	162,619	964,001
9	Banyuwangi	113,609	706,419
10	Bondowoso	61,33	329,557
11	Situbondo	48,902	290,954
12	Probolinggo	59,130	311,258

Selanjutnya akan dilakukan pengelompokan menggunakan *k-means cluster*, dengan pengambilan *cluster* awal 3 dan iterasi sebanyak 4 kali. Hasil tabel adalah sebagai berikut:

Tabel 2.9: Data Hasil Akhir pengelompokan Menggunakan *K-means Cluster*

NO	Kota /Kab	C1	C2	C3	Jarak Terpendek
1	Ponorogo	570,083	212,882	307,967	212,882
2	Trenggalek	792,141	25,362	530,027	25,362
3	Tulungagung	713,488	70,419	451,452	70,419
4	Blitar	683,749	100,168	421,663	100,168
5	Kediri	691,661	92,051	429,602	92,051
6	Malang	508,838	275,226	246,639	246,639
7	Lumajang	583,822	198,439	321,880	198,439
8	Jember	0	781,903	262,203	0
9	Banyuwangi	262,203	520,208	0	0
10	Bondowoso	642,479	140,220	380,471	140,220
11	Situbondo	682,586	101,723	420,474	101,723
12	Probolinggo	660,896	121,856	398,899	121,856

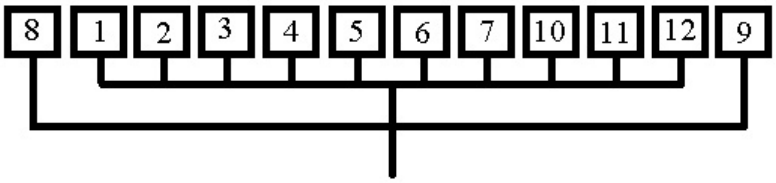
Maka diperoleh hasil akhir adalah pusat *cluster* 1 dengan menggunakan data ke-8, pusat *cluster* 2 dengan menggunakan data ke-7, pusat *cluster* 3 dengan menggunakan data ke-2. Dari hasil tabel di atas, pengelompokannya menjadi sebagai berikut:

Tabel 2.10: Data Hasil Pengelompokan Kota dengan Menggunakan *K-means Cluster*

No.	C1	C2	C3
1		1	
2		1	
3		1	
4		1	
5		1	
6		1	
7		1	
8	1		
9			1
10		1	
11		1	
12		1	

Pada tabel di atas, jika kota tersebut bernilai 1 pada setiap *cluster*, maka kota tersebut menjadi satu kelompok untuk tiap *cluster* tersebut. Dari penyelesaian di atas, maka dapat disimpulkan bahwa dalam *cluster* 1 terdapat satu kota, yaitu Kota Jember. Pada *cluster* 2 terdapat sepuluh kota untuk dijadikan satu kelompok, yaitu Kota Ponorogo, Kota Trenggelek, Kota Tulungagung, Kota Blitar, Kota Kediri, Kota Malang, Kota Lumajang, Kota Bondowoso, Kota Situbondo, dan Kota Probolinggo. Sedangkan pada *cluster* 3 terdapat satu kota, yaitu Kota Banyuwangi. Kota-kota tersebut dijadikan kelompok berdasarkan jarak terpendek terhadap pusat *cluster* awal.

Selanjutnya, penyelesaian di atas menghasilkan dendogram sebagai berikut:



Gambar 2.4: Dendogram Data Hasil Pengelompokkan Kota.