BAB II
TINJAUAN PUSTAKA

2.1. Jenis Data

Data merupakan kumpulan angka, fakta, fenomena atau keadaan sebagai hasil pengamatan, pengukuran dan pencacahan sehingga dapat digunakan sebagai dasar untuk menarik kesimpulan (Said, 2013).

Berdasarkan skala ukur, jenis data dapat dikelompokkan menjadi empat (Siegel, 1992).

1. Data Nominal
   Disebut juga data kategorik karena mengandung unsur penamaan dan bersifat kualitatif sehingga harus dirubah menjadi bentuk numerik. Peubah jenis kelamin dikategorikan sebagai laki-laki (1) dan perempuan (2).
2. Data Ordinal
   Juga bersifat kategorik dan memiliki tingkatan atau urutan di mana urutan skor menunjukkan arah tingkatan, seperti kepuasan konsumen terhadap suatu barang (sangat puas = 1, puas = 2, kurang puas = 3 dan tidak puas = 4).
3. Data interval
   Mempunyai karakteristik mengandung unsur penamaan, memiliki urutan bertingkat, bersifat interval yang bermakna dan nilai nol tidak mutlak, pada saat peubah bernilai nol bukan berarti tidak ada karakteristik yang diukur pada saat tersebut, seperti suhu ruangan (^oF, ^oC, ^oR).
4. Data Rasional
   Merupakan tingkatan data paling tinggi, sifat data nominal, ordinal, interval dan memiliki nilai nol mutlak. Pada data rasional semua operasi matematika dapat dilakukan. Seperti tinggi badan, berat badan, umur, dll.
   Jenis data di atas dapat digolongkan menjadi data metrik (skala interval dan rasional) dan nonmetrik (nominal dan ordinal).

2.2. Analisis Kelompok

Analisis kelompok adalah metode pengelompokan dengan cara membagi atau memecah obyek-obyek pada kelompok besar menjadi kelompok-kelompok yang lebih kecil di mana obyek dalam kelompok memiliki karakteristik yang relatif sama (Lattin dkk, 2006).

Konsep dasar analisis kelompok adalah pengukuran jarak (*distance*) dan kesamaan (*similarity*). Jarak menyatakan ukuran pisah

antar obyek sedangkan *similarity* adalah ukuran kedekatan. Konsep ini penting karena penggabungan obyek pada analisis kelompok didasarkan pada kedekatan. Jarak (*distance type measure*) digunakan pada data metrik, sedangkan kesesuaian (*matching type measure*) pada data nonmetrik.

Terdapat dua metode pengelompokan yaitu metode pengelompokan hirarki (*agglomerative* dan *divisive*) digunakan apabila belum ditentukan banyak kelompok dan metode pengelompokan non-hirarki (*K-means)* telah ditentukan sebanyak *K* kelompok yang akan dibentuk (Badriyah, 2006). Analisis kelompok dilandaskan pada data yang berstruktur X_{pi}.

Tabel 2.1. Struktur Data

Obyek *(i,j)*	*p*			
	1	*2*	…	*q*
1	X_{11}	X_{21}		X_{q1}
2	X_{12}	X_{22}		X_{q2}
:				*:*
:				*:*
n	X_{1n}	X_{2n}		X_{qn}
Total	$\sum_{i=1}^{n} X_{1i}$	$\sum_{i=1}^{n} X_{2i}$	…	$\sum_{i=1}^{n} X_{nq}$

di mana:

X_{pi} = nilai pengamatan obyek ke-*i* peubah ke-*p*
(i,j) = *1,2,3,…,n* ; $i \neq j$
n = banyaknya obyek
p = *1,2,3,…,q*
q = banyaknya peubah

Pada analisis kelompok, asumsi yang harus dipenuhi adalah nonmultikolinieritas yaitu peubah tidak saling berkorelasi.

2.3. Korelasi Antar Peubah

Yitnosumarto (1985) mengemukakan bahwa korelasi adalah keeratan hubungan linier antar dua peubah:

$$r_{pp'} = \rho_{pp'} = \frac{\text{cov}(X_p, X_{p'})}{\sqrt{JKX_p JKX_{p'}}} = \frac{\sum_{p=1}^{q}(X_{pi} - \overline{x}_{p.})(X_{p'i} - \overline{x}_{p'.})}{\sqrt{\sum_{p=1}^{q}(X_{pi} - \overline{x}_{p.})^2 (X_{p'i} - \overline{x}_{p'.})^2}} \quad (2.1)$$

di mana :

X_{pi} = nilai pengamatan obyek ke-*i* peubah ke-*p*

$X_{p'i}$ = nilai pengamatan obyek ke-*i* peubah ke-*p'*

$\overline{x}_{p.}$ = rata-rata peubah ke-*p*

$\overline{x}_{p'.}$ = rata-rata peubah ke-*p'*

$p, p' = 1,2,...,q \; ; p \neq p'$

Nilai $-1 \leq r \leq 1$ memiliki karakteristik:

1. Hanya merupakan suatu ukuran hubungan linier.
2. Simetris, sehingga $r_{(p,p')} = r_{(p',p)}$ seperti tampak pada matriks korelasi berordo *q*.

$$\underset{(qxq)}{\boldsymbol{R}} = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1q} \\ r_{21} & r_{22} & \cdots & r_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ r_{q1} & r_{q2} & \cdots & r_{qq} \end{bmatrix}$$

3. Peubah yang saling bebas akan memiliki nilai r = 0.

Analisis komponen utama digunakan untuk mengatasi korelasi antar peubah.

2.4. Analisis Komponen Utama

Analisis komponen utama adalah sebuah teknik untuk membangun peubah baru yang merupakan kombinasi linier dari peubah asli. Banyaknya peubah baru maksimum sama dengan banyaknya peubah asli dan tidak saling berkorelasi.

Misal vektor peubah asal yaitu $\boldsymbol{X}' = (X_1, X_2, \ldots, X_q)$ dengan vektor rata-rata $\boldsymbol{\mu}$ dan matriks ragam peragam $\boldsymbol{\Sigma} = \left[\sigma_{pp'}^{\;2}\right]$ ditransformasi menjadi vektor peubah baru $\boldsymbol{Y}' = (Y_1, Y_2, \ldots, Y_q)$. Kombinasi linier pada analisis komponen utama adalah :

$$Y_1 = \boldsymbol{a}'_1 \boldsymbol{X} = a_{11}X_1 + a_{12}X_2 + ... + a_{1q}X_q$$
$$Y_2 = \boldsymbol{a}'_2 \boldsymbol{X} = a_{21}X_1 + a_{22}X_2 + ... + a_{2q}X_q$$
$$\cdots$$
$$Y_q = \boldsymbol{a}'_q \boldsymbol{X} = a_{q1}X_1 + a_{q2}X_2 + ... + a_{qq}X_q$$

$\boldsymbol{a}_1, \boldsymbol{a}_2, ..., \boldsymbol{a}_q$ adalah vektor koefisien untuk setiap komponen utama yang bersesuaian. Ragam dan peragam peubah **Y** adalah:

$$Var(Y_p) = \boldsymbol{a}'_p \sum \boldsymbol{a}_p \qquad p = 1,2,...,q$$
$$Cov(Y_p, Y_{p'}) = \boldsymbol{a}'_p \sum \boldsymbol{a}_{p'} \qquad p,p' = 1,2,...,q \; ; p \neq p'$$

$\sum$ = matriks ragam peragam peubah asal. Menghitung komponen utama $Y_p = \boldsymbol{a}'_p \boldsymbol{X}$, sama dengan menghitung vektor koefisien $\boldsymbol{a}_p$ sedemikian sehingga $\boldsymbol{a}'_p \sum \boldsymbol{a}_p$ adalah ragam komponen utama yang akan mencapai nilai maksimum dengan syarat $\boldsymbol{a}'_p \boldsymbol{a}_p = 1$. Berdasarkan fungsi Lagrange

$$L = \boldsymbol{a}'_p \sum \boldsymbol{a}_p - \lambda_p (\boldsymbol{a}'_p \boldsymbol{a}_p - 1) \qquad (2.2)$$

Apabila L diturunkan terhadap $\boldsymbol{a}_p$ kemudian disamakan dengan nol, menghasilkan :

$$\frac{\partial L}{\partial \boldsymbol{a}} = 2\boldsymbol{a}_p \sum - 2\lambda_p \boldsymbol{a}_p = 0$$
$$2(\sum - \lambda_p \boldsymbol{I})\boldsymbol{a}_p = 0$$
$$(\sum - \lambda_p \boldsymbol{I})\boldsymbol{a}_p = 0 \qquad (2.3)$$

Persamaan (2.3) akan menghasilkan jawaban nontrivial apabila matriks $(\sum - \lambda_p \boldsymbol{I})$ bersifat singular, melalui persamaan ciri.

$$\left|(\sum - \lambda_p \boldsymbol{I})\right| = 0 \qquad (2.4)$$

Persamaan (2.4) akan menghasilkan akar-akar ciri $\lambda_1, \lambda_2, ..., \lambda_q$ di mana $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_q > 0$ akan menentukan vektor ciri $\boldsymbol{a}_p$.

Akar ciri yang akan digunakan pada komponen utama pertama didasarkan pada:

$$\sum \boldsymbol{a}_p - \lambda_p \boldsymbol{I} \boldsymbol{a}_p = 0$$

$$(\sum -\lambda_p \boldsymbol{I})\boldsymbol{a}_p = 0$$

$$\sum \boldsymbol{a}_p = \lambda_p \boldsymbol{I}\boldsymbol{a}_p \qquad (2.5)$$

Jika kedua sisi persamaan (2.5) dikalikan dengan $\boldsymbol{a}_p^{'}$ akan diperoleh persamaan :

$$\boldsymbol{a}_p^{'} \sum \boldsymbol{a}_p = \lambda_p \qquad (2.6)$$

Dengan demikian ragam setiap komponen utama bersesuaian dengan nilai setiap akar ciri. Persamaan (2.6) menunjukkan bahwa ragam komponen utama maksimum adalah akar ciri terbesar dari matriks $\sum$.

Terdapat dua tipe masukan pada analisis komponen utama yaitu matriks ragam peragam dan matriks korelasi (Johnson dan Winchern, 2006).

1. Matriks Ragam Peragam

   Digunakan apabila peubah memiliki satuan sama . Pandang $\sum$ sebagai matriks ragam peragam untuk vektor acak $\boldsymbol{X}'$ = $(X_1, X_2, \ldots, X_q)$ dengan pasangan akar ciri dan vektor ciri yang saling ortonormal $(\lambda_1, a_1), (\lambda_2, a_2), \ldots, (\lambda_q, a_q)$ di mana $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_q \geq 0$ maka komponen utama ke-*p* didefinisikan sebagai :

   $Y_p = \boldsymbol{a}_p^{'}\boldsymbol{X} = a_{p1}X_1 + a_{p2}X_2 + \ldots + a_{pq}X_q \qquad p = 1,2,\ldots,q$

   Skor komponen yang dihasilkan adalah

$$SK_{pi}vc = \begin{bmatrix} a_{1i} & a_{2i} & \ldots & a_{qi} \end{bmatrix} \begin{bmatrix} X_{p1} - \bar{X}_1 \\ X_{p2} - \bar{X}_2 \\ \vdots \\ X_{pq} - \bar{X}_q \end{bmatrix}$$

   atau

$$= \boldsymbol{a}_i^{'}(\boldsymbol{X}_p - \bar{\boldsymbol{X}}) \qquad (2.7)$$

   di mana :

   $SK_{pi}vc$ = skor komponen obyek ke-*i* peubah ke-*p*

   $\boldsymbol{a}_i^{'}$ = vektor koefisien komponen utama obyek ke-*i*

   $\boldsymbol{X}_p$ = vektor data peubah ke-*p*

   $\bar{\boldsymbol{X}}$ = vektor rata-rata peubah asal

2. Matriks Korelasi

Jika peubah memiliki satuan berbeda maka matriks korelasi digunakan sehingga matriks masukan semua peubah ditransformasi menjadi peubah normal baku yang memiliki satuan sama.

$$Z_{pi} = \frac{X_{pi} - \bar{x}_{p.}}{S_p} \tag{2.8}$$

di mana :

X_{pi} = nilai pengamatan obyek ke-*i* peubah ke-*p*

$\bar{x}_{p.}$ = rata-rata peubah ke-*p*

S_p = simpangan baku peubah ke-*p*

Komponen utama ke-*p* didefinisikan sebagai berikut :

$$Y_p = \boldsymbol{a}_{\boldsymbol{p}}^{'}\boldsymbol{Z} = a_{p1}Z_1 + a_{p2}Z_2 + ... + a_{pq}Z_q \qquad p = 1,2,...,q$$

maka skor komponen utama obyek ke-*i* adalah:

$$SK_{pi}cr = \boldsymbol{a}_i^{'}\mathbf{D}^{1/2}(\boldsymbol{X}_p - \bar{\boldsymbol{X}}) \tag{2.9}$$

di mana :

$SK_{pi}cr$ = skor komponen obyek ke-*i* peubah ke-*p*

$\boldsymbol{a}_i^{'}$ = vektor koefisien komponen utama obyek ke-*i*

$\boldsymbol{X}_p$ = vektor data peubah ke-*p*

$\bar{\boldsymbol{X}}$ = vektor rata-rata peubah asal

$\mathbf{D}^{1/2}$ = diag$\left[(S_{pp})^{-1/2}\right]$, S_{pp} adalah diagonal utama ke-*p* matriks **S,** *p = 1,2,...,q*

Banyaknya komponen utama untuk dapat menjelaskan keragaman data dengan baik menurut Johnson dan Winchern (2006) dilihat dari proporsi keragaman kumulatif komponen utama minimum sebesar 80-90%, dihitung menggunakan persamaan :

$$P = \frac{\lambda_q}{\lambda_1 + \lambda_2 + ... + \lambda_p} \times 100\% \tag{2.10}$$

2.5. Pengelompokan *Bottom-Up*

Pengelompokan *agglomerative* adalah suatu metode pengelompokan hirarki yang bersifat *bottom-up*, dimulai dari meletakkan setiap obyek sebagai sebuah kelompok tersendiri (*atomic cluster*) kemudian menggabungkan *atomic cluster-atomic cluster* tersebut menjadi kelompok yang lebih besar sampai semua obyek bergabung menjadi satu kelompok tunggal.

Kelebihan metode *bottom-up* adalah mampu mengelompokan banyaknya obyek berukuran kecil (Chipman dan Tibshirani, 2006). Kekurangan metode ini adalah (Dillon dan Goldstein, 1984) :

a. Kurang efisien mengelompokan obyek berukuran besar.
b. Sensitif terhadap *outlier*.

Penggabungan 2 kelompok pada tahap awal dengan metode *agglomerative* memerlukan ukuran ketidakmiripan (fungsi jarak) antar obyek (Hair dkk, 2012). Semakin besar nilai ketidakmiripan antara dua obyek menunjukkan semakin besar pula perbedaan antara kedua obyek itu. Ukuran ketidakmiripan antara obyek ke-*i* dengan obyek ke-*j* (d_{ij}) memenuhi persyaratan berikut :

1. $d_{ij} \geq 0$, untuk setiap i dan j
2. $d_{ij} = 0$, untuk $i = j$
3. $d_{ij} = d_{ji}$, sehingga banyaknya jarak yang terbentuk adalah $\frac{n^2 - n}{2}$

Salah satu fungsi ketidakmiripan obyek i dan j adalah jarak Euclidean :

$$d_{ij} = \sqrt{\sum_{p=1}^{q} \left(X_{pi} - X_{pj} \right)^2} \tag{2.11}$$

di mana :

d_{ij} = jarak antar obyek ke-*i* dengan obyek ke-*j*
(i,j) = $1,2,3,\ldots,n$; $i \neq j$
X_{pi} = nilai pengamatan obyek ke-*i* peubah ke-*p*
X_{pj} = nilai pengamatan obyek ke-*j* peubah ke-*p*

Menurut Manly (1988), penggunaan jarak Euclidean memenuhi 3 syarat yaitu peubah tidak saling berkorelasi, memiliki satuan pengukuran sama dan normal baku.

Nilai ukuran ketidakmiripan antara obyek ke-*i* dengan obyek ke-*j* (d_{ij}) disajikan dalam bentuk matriks awal $\boldsymbol{D}_0$ berordo n.

$$\boldsymbol{D}_0 = (d_{ij})_{nxn} = \begin{bmatrix} 0 & d_{12} & \cdots & d_{1n} \\ d_{21} & 0 & \cdots & d_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \cdots & 0 \end{bmatrix}$$

Beberapa metode pengelompokan secara *agglomerative* adalah *single linkage* yaitu pengelompokkan berdasar jarak minimum, *complete linkage* berdasar jarak maksimum dan *average linkage* berdasar jarak rata-rata antar obyek.

Metode *average linkage* dimulai dengan menghitung jarak antar dua obyek (d_{ij}) dan menggabungkan obyek-obyek yang saling bedekatan, misalkan kelompok U dan V pada tahap awal untuk mendapatkan kelompok (UV). Kemudian menghitung jarak antara kelompok (UV) dengan kelompok lain, misal W adalah :

$$d_{(UV)W} = \frac{1}{n_{(UV)} n_W} \sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij}^{*} \qquad (2.12)$$

di mana:

$n_{(UV)}$ = banyaknya anggota dalam kelompok (UV)

n_W = banyaknya anggota dalam kelompok W

d_{ij}^{*} = jarak obyek *i* pada kelompok (UV) dan obyek *j* pada kelompok W

Berikut adalah algoritma pengelompokan dengan metode *agglomerative average linkage*:

a. Setiap obyek dipandang sebagai kelompok tersendiri (*atomic cluster).*

b. Menghitung $\frac{n^2 - n}{2}$ jarak antar *atomic cluster* (d_{ij}) menggunakan fungsi jarak Euclidean sebagai unsur-unsur matriks awal $\boldsymbol{D}_0 = (d_{ij})_{nxn}$

c. Pasangan kelompok yang memiliki jarak terdekat adalah statistik peringkat pertama $d_{(UV)}$

d. Menghitung kembali $\frac{(n-1)^2 - (n-1)}{2}$ jarak kelompok pertama dengan obyek lain ($d_{(uv)w}$) menggunakan metode pengelompokan *average linkage* menghasilkan matriks jarak

pertama $\boldsymbol{D}_1 = (d_{ij})_{(n-1)x(n-1)}$. Hal ini dikarenakan 2 obyek pada tahap awal membentuk satu kelompok tunggal, sehingga banyak obyek berkurang 1. Penggabungan berakhir sampai *n-1* obyek sehingga terbentuk $\boldsymbol{D}_{n-2}$ berordo *n(n-2).*

Untuk memahami bagaimana pengelompokan menggunakan metode *average linkage*, pandang hasil pengamatan X_1 dan X_2 pada 4 obyek.

Tabel 2.2. Data Hipotetik

Obyek	Peubah	
	X_1	X_2
1	7	6
2	3	2
3	8	11
4	6	9

Pada tahap awal, sebanyak $\frac{n^2-n}{2} = \frac{4^2-4}{2} = 6$ jarak Euclidean antar obyek :

$$d_{12} = \sqrt{\sum_{p=1}^{q}(X_{p1}-X_{p2})^2} = \sqrt{(7-3)^2+(6-2)^2} = 5.657$$

$$d_{13} = \sqrt{\sum_{p=1}^{q}(X_{p1}-X_{p3})^2} = \sqrt{(7-8)^2+(6-11)^2} = 5.099$$

$$d_{14} = \sqrt{\sum_{p=1}^{q}(X_{p1}-X_{p4})^2} = \sqrt{(7-6)^2+(6-9)^2} = 3.162$$

$$d_{23} = \sqrt{\sum_{p=1}^{q}(X_{p2}-X_{p3})^2} = \sqrt{(3-8)^2+(2-11)^2} = 10.296$$

$$d_{24} = \sqrt{\sum_{p=1}^{q}(X_{p2}-X_{p4})^2} = \sqrt{(7-6)^2+(2-9)^2} = 7.612$$

$$d_{34} = \sqrt{\sum_{p=1}^{q}(X_{p3}-X_{p4})^2} = \sqrt{(8-6)^2+(11-9)^2} = 2.828$$

Matriks simetri jarak awal antar obyek adalah :

$$\boldsymbol{D}_0 = (d_{ij})_{4x4} = \begin{array}{c} \textit{Obyek} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{bmatrix} 0 & 5.657 & 5.099 & 3.162 \\ 5.657 & 0 & 10.296 & 7.616 \\ 5.099 & 10.296 & 0 & 2.828 \\ 3.162 & 7.616 & 2.828 & 0 \end{bmatrix} \end{array}$$

Setiap obyek diperlakukan sebagai kelompok kemudian dua obyek terdekat digabung. Statistik peringkat yang terbentuk :

d_{34}	<	d_{14}	<	d_{13}	<	d_{12}	<	d_{24}	<	d_{23}
2.828		3.162		5.099		5.657		7.616		10.296

Jarak terdekat adalah min(d_{ij}) = d_{34} = 2.828, kemudian obyek 3 dan 4 digabung membentuk kelompok (34). Menghitung jarak antara kelompok (34) dengan obyek 1 dan 2.

$$d_{(34)1} = \frac{d_{13}+d_{14}}{2} = \frac{5.099+3.162}{2} = 4.131$$

$$d_{(34)2} = \frac{d_{23}+d_{24}}{2} = \frac{10.296+7.616}{2} = 8.956$$

Karena obyek 3 dan 4 membentuk satu kelompok tunggal, maka matriks jarak berordo 3 sebanyak $\frac{n^2-n}{2} = \frac{3^2-3}{2} = 3$ jarak Euclidean antar obyek

$$\boldsymbol{D}_1 = (d_{ij})_{3x3} = \begin{array}{c} \textit{kelompok}\ (34) \\ (34) \\ 1 \\ 2 \end{array} \begin{array}{c} \begin{array}{ccc} 1 & 2 \end{array} \\ \begin{bmatrix} 0 & 4.131 & 8.956 \\ 4.131 & 0 & 5.657 \\ 8.956 & 5.657 & 0 \end{bmatrix} \end{array}$$

Jarak terdekat yaitu $d_{(34)1}$= 4.131

$d_{(34)1}$	<	d_{12}	<	$d_{(34)2}$
4.131		5.657		8.956

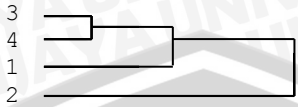
Kelompok (34) dan (1) digabung menjadi kelompok tunggal (341), kemudian hitung jarak kelompok (341) dan (2) :

$$d_{(341)2} = \frac{d_{342}+d_{12}}{2} = \frac{8.956+5.657}{2} = 7.307$$

Matriks jarak berordo 2 karena kelompok (34) dan (1) telah menjadi kelompok tunggal (341) pada langkah sebelumnya.

$$\boldsymbol{D}_2 = (d_{ij})_{2x2} = \begin{array}{c} \textit{Kelompok}\ (341) \\ (341) \\ 2 \end{array} \begin{array}{c} \begin{array}{cc} & 2 \end{array} \\ \begin{bmatrix} 0 & 7.307 \\ 7.307 & 0 \end{bmatrix} \end{array}$$

Langkah terakhir yaitu kelompok (341) dan (2) digabung pada kelompok tunggal (3412). Dendrogram hasil pengelompokan dengan metode *average linkage* pada Gambar 2.1 menunjukkan bahwa jarak kelompok (34) lebih dekat daripada kelompok (341) dan kelompok (3412).



Gambar 2.1. Hasil pengelompokan dengan metode *average linkage*

2.6. Pengelompokan *Top-Down*

K-means merupakan suatu metode pengelompokan non hirarki yang bersifat *top-down* yaitu membagi *n* obyek ke dalam *K* kelompok berdasarkan algoritma:

a. Mempartisi obyek sebanyak *K* kelompok (ditentukan oleh peneliti).
b. Menghitung pusat kelompok menggunakan persamaan :

$$C_{(k)p} = \frac{1}{n_p} \sum_{i=1}^{n} X_{pi} \qquad (2.13)$$

di mana:

$c_{(k)p}$ = pusat kelompok ke-*k* peubah ke-*p*
k = *1,2, ..., K ; K* = banyaknya kelompok
n_p = banyaknya obyek pada peubah ke-*p*
X_{pi} = nilai pengamatan obyek ke-i peubah ke-*p*
i = *1,2, ...,n* ; *n* = banyaknya obyek
p =*1,2, ...,q* ; *q* = banyaknya peubah

c. Menghitung jarak setiap obyek ke pusat kelompok menggunakan fungsi jarak Euclidean.
d. Menentukan obyek yang memiliki jarak terdekat dengan pusat kelompok. Jika obyek berpindah dari posisi awal (langkah a) maka pusat kelompok harus dihitung kembali.
e. Mengulangi langkah (b) – (d) sampai tidak ada lagi obyek yang berpindah posisi.

Kelebihan metode *K-means* adalah mampu mengelompokkan banyaknya obyek berukuran besar dan tidak membutuhkan operasi matematika rumit. Kekurangan metode ini adalah (Arai dan Barakbah, 2007) :

a. Banyaknya kelompok harus ditentukan terlebih dahulu.
b. Pusat kelompok awal mempengaruhi hasil pengelompokan.
c. Sulit mencapai global optimum.
d. Tidak dapat menunjukkan kedekatan antar obyek karena jarak dihitung dari pusat kelompok.

Untuk mempermudah pemahaman algoritma *K-means*, pandang data pada Tabel 2.3, langkah pertama adalah mempartisi obyek menjadi

K kelompok (misal K = 2), sehingga terbentuk kelompok (12) dan (34) dan pusat kelompok dihitung seperti disajikan pada Tabel 2.3.

Tabel 2.3. Data hipotetik koordinat pusat

Kelompok	Koordinat pusat	
	$\bar{x}_1$	$\bar{x}_2$
(12)	$\frac{7+3}{2}=5$	$\frac{6+2}{2}=4$
(34)	$\frac{8+6}{2}=7$	$\frac{11+9}{2}=10$

Jarak setiap obyek dari pusat kelompok dihitung dan ditentukan obyek yang berjarak lebih dekat dengan pusat kelompok. Dengan menggunakan jarak Euclidean, jarak antar obyek adalah:

$$d(1,(12)) = \sqrt{(7-5)^2+(6-4)^2} = 2.828$$

$$d(1,(34)) = \sqrt{(7-7)^2+(6-10)^2} = 4$$

Jarak obyek 1 paling dekat dengan pusat kelompok (12), sehingga proses dilanjutkan.

$$d(2,(12)) = \sqrt{(3-5)^2+(2-4)^2} = 2.828$$

$$d(2,(34)) = \sqrt{(3-7)^2+(2-10)^2} = 8.944$$

Proses dilanjutkan kembali karena jarak obyek 2 paling dekat dengan pusat kelompok (12),

$$d(3,(12)) = \sqrt{(8-7)^2+(11-10)^2} = 7.616$$

$$d(3,(34)) = \sqrt{(8-7)^2+(11-10)^2} = 1.414$$

Jarak obyek 3 paling dekat dengan pusat kelompok (34), proses terus dilanjutkan.

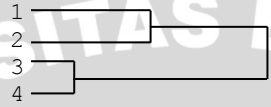
$$d(4,(12)) = \sqrt{(6-5)^2+(9-4)^2} = 5.099$$

$$d(4,(34)) = \sqrt{(6-7)^2+(9-10)^2} = 1.414$$

Jarak obyek 4 paling dekat dengan pusat kelompok (34), maka tidak ada obyek yang berpindah dari posisi awal sehingga proses dihentikan. Berikut adalah ringkasan jarak Euclidean setiap obyek terhadap pusat kelompok :

Kelompok	Jarak Euclidean Setiap Obyek Terhadap Pusat Kelompok			
	1	2	3	4
(12)	2.828	2.828	7.616	5.099
(34)	4	8.944	1.414	1.414

Hasil pengelompokan dengan metode *K-means* (*K* = 2) yaitu kelompok (12) dan (34) tersaji pada dendrogram (Gambar 2.2).



Gambar 2.2. Hasil pengelompokan dengan metode *K-means*

2.7. Pengelompokan *Hybrid* Melalui *Mutual Cluster*

Pada tahun 2006, Chipman dan Tibshirani memperkenalkan metode pengelompokan baru bernama *hybrid clustering* yang mengkombinasikan kelebihan metode *bottom-up (agglomerative)* dan *top-down (k-means)*. Algoritma *bottom-up* baik dalam mengidentifikasi banyaknya obyek berukuran kecil karena metode ini dimulai dari *atomic cluster* (kelompok tersendiri) sampai bergabung menjadi kelompok besar sedangkan algoritma *top-down* baik dalam mengidentifikasi banyaknya obyek berukuran besar karena pengelompokan dimulai dari kelompok tunggal yang memiliki banyak anggota sampai menjadi kelompok beranggotakan sedikit obyek. Pengelompokan *hybrid* melalui *mutual cluster* dilakukan mengikuti 2 langkah yaitu menentukan *mutual cluster* secara *bottom-up* dan *top-down* dari hasil *mutual cluster* (Chipman dan Tibshirani, 2006).

Mutual cluster menggunakan jarak terjauh antar obyek dalam himpunan *S*, ekuivalen dengan diameter *(S)* yang lebih kecil dari jarak terdekat antara obyek dalam himpunan *S* dengan obyek lain yang tidak termasuk dalam himpunan *S*. Teorema ini dapat dirumuskan pada persamaan berikut:

$$d(x,y) > diameter\ (S) \equiv \max_{w \in S, z \in S} d\ (w,z) \qquad (2.14)$$

di mana:

S = sebuah himpunan bagian dari data

(w,z)= pasangan obyek dalam *S* yang memiliki jarak terjauh

x = obyek *S*

y = obyek lain yang tidak termasuk dalam *S*

Hal ini menunjukkan bahwa jarak maksimum antar obyek dalam sebuah *mutual cluster* lebih kecil dibandingkan jarak minimum antar obyek di luar *mutual cluster* sehingga obyek yang berada dalam sebuah *mutual cluster* tidak pernah dipisahkan (Chipman dan Tibshirani, 2006).

Pandang Tabel 2.2, pengelompokan obyek secara *bottom-up* menggunakan metode *average linkage* dengan jarak Euclidean menghasilkan matriks awal :

$$\boldsymbol{D}_0 = (d_{ij})_{4x4} = \begin{array}{c} \textit{Obyek} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{bmatrix} 0 & 5.657 & 5.099 & 3.162 \\ 5.657 & 0 & 10.296 & 7.616 \\ 5.099 & 10.296 & 0 & 2.828 \\ 3.162 & 7.616 & 2.828 & 0 \end{bmatrix} \end{array}$$

Obyek 3 dan 4 memiliki jarak terdekat min(d_{ij}) = d_{34} = 2.828 sehingga membentuk kelompok (34). Kelompok ini menjadi *mutual cluster* pertama. Jarak antara kelompok (34) dengan obyek 1 dan 2 adalah :

$$d_{(34)1} = \frac{d_{13}+d_{14}}{2} = \frac{5.099+3.162}{2} = 4.131$$

$$d_{(34)2} = \frac{d_{23}+d_{24}}{2} = \frac{10.296+7.616}{2} = 8.956$$

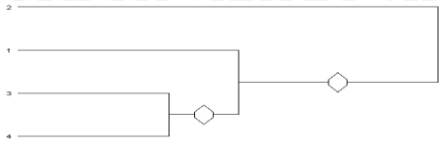
dan menghasilkan matriks jarak baru :

$$\boldsymbol{D}_1 = (d_{ij})_{3x3} = \begin{array}{c} \textit{kelompok}\ (34) \\ (34) \\ 1 \\ 2 \end{array} \begin{array}{c} \begin{array}{ccc} & 1 & 2 \end{array} \\ \begin{bmatrix} 0 & 4.131 & 8.956 \\ 4.131 & 0 & 5.657 \\ 8.956 & 5.657 & 0 \end{bmatrix} \end{array}$$

Jarak terdekat adalah d_{341}= 4.131. Kelompok (34) dan obyek 1 digabung membentuk kelompok (341) sebagai *mutual cluster* kedua karena jarak obyek maksimum kelompok (341) = d_{13}(5.099) < d_{12}(5.657) di mana d_{12} merupakan jarak terkecil antara obyek kelompok (341) dengan kelompok (2). Kemudian kelompok (341) dan (2) yang berjarak $d_{(341)2} = \frac{d_{342}+d_{12}}{2} = \frac{8.956+5.657}{2}$ = 7.307 digabung menjadi kelompok tunggal.

$$\boldsymbol{D}_2 = (d_{ij})_{2x2} = \begin{array}{c} \textit{Kelompok}\ (341) \\ (341) \\ 2 \end{array} \begin{array}{c} \begin{array}{cc} & 2 \end{array} \\ \begin{bmatrix} 0 & 7.307 \\ 7.307 & 0 \end{bmatrix} \end{array}$$

Hasil pengelompokan secara *bottom-up* diilustrasikan pada dendrogram seperti terlihat pada Gambar 2.3 di mana *mutual cluster* ditunjukkan dengan tanda belah ketupat.



Gambar 2.3. Hasil pengelompokan secara *bottom-up*

Langkah kedua adalah melakukan pengelompokan *hybrid* dari hasil *mutual cluster* berdasar algoritma *K-means* dengan *K*=2. Obyek-obyek dibagi menjadi 2 kelompok di mana *mutual cluster* yang telah terbentuk harus tetap ada dan dipertahankan. Dengan demikian terbentuk kelompok (341) dan (2).

Hasil pengelompokan dengan metode *hybrid* melalui *mutual cluster* diilustrasikan dendrogram pada Gambar 2.4.

Gambar 2.4. Hasil pengelompokan dengan metode *hybrid* melalui *mutual cluster*

2.8. Pemilihan Metode Pengelompokan Terbaik

Pemilihan metode pengelompokan terbaik didasarkan pada kepadatan kelompok (*cluster density)* yang ditentukan oleh ragam dalam kelompok (*within cluster,* V_w) dan ragam antar kelompok (*between cluster,* V_b). Pengelompokan yang baik akan memiliki nilai V_w minimum dan V_b maksimum (Man dkk, 2009). Nilai V_w dan V_b dihitung dengan persamaan:

$$V_w = \frac{1}{n-K}\sum_{k=1}^{K}(n_k - 1)V_k \qquad (2.15)$$

di mana $V_k = \dfrac{\sum_{i=1}^{n_k}(X_{ki} - \bar{x}_k)^2}{n_k - 1}$ (2.16)

$$V_b = \frac{1}{K-1}\sum_{k=1}^{K} n_k(\bar{x}_k - \bar{x})^2 \qquad (2.17)$$

dengan :

V_k = ragam kelompok ke-*k*

X_{ki} = nilai pengamatan obyek ke-i kelompok ke-k

$\bar{x}_k$ = rata-rata kelompok ke-k

$\bar{x}$ = rata-rata $\bar{x}_k$

n = banyaknya semua obyek

n_k = banyaknya obyek kelompok ke-k

k = $1,2,\ldots,K$; K = banyaknya kelompok

Metode pengelompokan terbaik menghasilkan rasio minimum *Vw* terhadap *Vb* (Arai dan Barakbah, 2007).

$$\min\left(\frac{V_w}{V_b}\right) \tag{2.18}$$

2.9. Indikator Pendidikan

Konsep dan definisi yang digunakan dalam indikator pendidikan adalah (BPS Jatim, 2011) :

a. Angka Partisipasi Sekolah (APS) adalah proporsi banyaknya penduduk yang bersekolah menurut kelompok usia sekolah tertentu.

$$\text{APS} = \frac{\textit{Banyaknya penduduk usia sekolah tertentu yang sedang sekolah}}{\textit{Banyaknya penduduk usia sekolah tertentu}}\ x\ 100\%$$

b. Rata-rata Lama Sekolah adalah waktu yang dihabiskan setiap penduduk berusia minimum 15 tahun untuk menempuh pendidikan formal yang pernah dijalani.

c. Angka Melek Huruf (AMH) adalah persentase penduduk yang dapat membaca dan menulis huruf (Latin dan atau huruf lain).

$$\text{AMH} = \frac{\textit{Banyaknya penduduk usia tertentu yang dapat membaca dan menulis}}{\textit{Banyaknya penduduk usia tertentu}}\ x\ 100\%$$