

## BAB II TINJAUAN PUSTAKA

### 2.1 Gambaram Umum Provinsi Jawa Tengah



Gambar 2.1 Peta Provinsi Jawa Tengah

Jawa Tengah merupakan salah satu provinsi di Pulau Jawa, letaknya diapit oleh dua provinsi besar, yaitu Jawa Barat dan Jawa Timur. Letaknya antara  $5^{\circ}40'$  dan  $8^{\circ}30'$  Lintang Selatan dan antara  $108^{\circ}30'$  dan  $111^{\circ}30'$  Bujur Timur (termasuk Pulau Karimunjawa). Jarak terjauh dari Barat ke Timur adalah 263 km dan dari Utara ke Selatan 226 km (tidak termasuk Pulau Karimunjawa) Provinsi Jawa Tengah terbagi menjadi 29 kabupaten dan 6 kota. Luas wilayah Jawa Tengah tercatat sebesar 3,25 juta hektar atau sekitar 25,04 persen dari luas Pulau Jawa (1,70 persen dari luas Indonesia).

Jumlah penduduk Provinsi Jawa Tengah berdasarkan proyeksi penduduk tahun 2016 sebanyak 34.019,10 ribu jiwa yang terdiri atas 16.871,19 ribu jiwa penduduk laki-laki dan 17.147,90 ribu jiwa penduduk perempuan. Dibandingkan dengan proyeksi jumlah penduduk tahun 2015, penduduk Jawa Tengah mengalami pertumbuhan sebesar 0,73 persen. Sementara itu besarnya angka rasio jenis kelamin tahun 2016 penduduk laki-laki terhadap penduduk perempuan sebesar 98,39. Kepadatan penduduk di Jawa Tengah tahun 2016 mencapai 1.045 jiwa/km<sup>2</sup>. Kepadatan Penduduk di 35 kabupaten/kota cukup beragam dengan kepadatan

penduduk tertinggi terletak di kota Surakarta dengan kepadatan sebesar 11.678 jiwa/km<sup>2</sup> dan terendah di Kabupaten Blora sebesar 477 jiwa/Km<sup>2</sup> (BPS Jateng).

## 2.2 Kemiskinan

Masalah kemiskinan merupakan salah satu persoalan mendasar yang menjadi pusat perhatian pemerintah di negara manapun. Sehingga berbagai upaya atau strategi penanggulangan dilakukan. Salah satu aspek penting untuk mendukung strategi penanggulangan kemiskinan adalah tersedianya data kemiskinan yang akurat dan tepat sasaran. Pengukuran kemiskinan yang dapat dipercaya dapat menjadi instrumen tangguh bagi pengambil kebijakan dalam memfokuskan perhatian pada kondisi hidup orang miskin. Data kemiskinan yang baik dapat digunakan untuk mengevaluasi kebijakan pemerintah terhadap kemiskinan, membandingkan kemiskinan antar waktu dan daerah, serta menentukan target penduduk miskin dengan tujuan untuk memperbaiki kondisi mereka (BPS Jateng, 2017).

Untuk mengukur kemiskinan, BPS menggunakan konsep kemampuan memenuhi kebutuhan dasar (*basic needs approach*). Dengan pendekatan ini, kemiskinan dipandang sebagai ketidakmampuan dari sisi ekonomi untuk memenuhi kebutuhan dasar makanan dan bukan makanan yang diukur dari sisi pengeluaran. Jadi Penduduk Miskin adalah penduduk yang memiliki rata-rata pengeluaran perkapita perbulan dibawah garis kemiskinan.

Di Indonesia angka kemiskinan masih cukup tinggi sebagaimana data yang dirilis BPS bahwa persentase penduduk miskin sekitar 10,7%. Selain itu ada juga penduduk yang hidup tidak jauh dari garis kemiskinan. Sekitar 68 juta penduduk Indonesia hidup tidak jauh dari batas Rp 11.000. Dengan sedikit sakit, bencana atau kehilangan pekerjaan, mereka bisa langsung kembali miskin (worldbank, 2014). Sebagian besar penduduk miskin tersebut berada di Pulau Jawa sebagaimana terlihat pada tabel 2.1 di bawah ini.

Tabel 2.1 Persentase dan Jumlah Penduduk di Indonesia per Pulau

| Pulau                  | Penduduk Miskin |               |
|------------------------|-----------------|---------------|
|                        | Persentase      | Jumlah (Juta) |
| Maluku dan Papua       | 21,98           | 1,65          |
| Bali dan Nusa Tenggara | 14,72           | 2,11          |
| Sulawesi               | 10,97           | 2             |
| Kalimantan             | 6,45            | 0,97          |
| Jawa                   | 10,09           | 14,83         |
| Sumatera               | 11,03           | 6,21          |

Sumber: Badan Pusat Statistik

Dari Tabel 2.1 terlihat bahwa Pulau Jawa merupakan penyumbang jumlah kemiskinan terbanyak di Indonesia. Namun dari sisi persentase berada diposisi keempat dengan persentase 10,09. Kemiskinan di Pulau Jawa jika ditinjau per Provinsi berdasarkan data BPS sebagaimana pada Tabel 1.1 dapat diketahui bahwa Provinsi Jawa Tengah merupakan provinsi dengan persentase penduduk tertinggi di Pulau Jawa.

### 2.3 Standarisasi Data

Jika diantara variabel-variabel penelitian terdapat perbedaan ukuran satuan data yang besar maka perlu dilakukan proses standarisasi data. perbedaan ukuran data yang besar dapat menyebabkan ketidakvalidan analisis klaster. Standarisasi sebagaimana yang umum digunakan adalah dengan bentuk z skor sebagai berikut:

$$z = \frac{x - \bar{x}}{s} \quad (1)$$

dengan:  $x$  = nilai data

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

$s$  = standar deviasi

### 2.4 Data Mining

Data Mining merupakan metode analisis data yang memadukan berbagai bidang keilmuan seperti statistik, teknik, mesin learning dan lainnya. Seiring berkembangnya teknologi khususnya dibidang komputasi teknik data mining juga

ikut dikembangkan. Tan dalam Prasetyo (2012) mendefinisikan data mining sebagai proses untuk mendapatkan informasi yang berguna dari gudang basis data yang besar. Han et al. (2012) mendefinisikan data mining sebagai proses menemukan pola dan pengetahuan menarik dari sejumlah besar data. Sumber data dapat mencakup database, gudang data, web, repositori informasi atau data lain yang dialirkan ke sistem secara dinamis. Hand et al. (2001) mendefinisikan sebagai sebuah analisa dari observasi data dalam jumlah yang besar untuk menemukan hubungan yang tidak diketahui sebelumnya dan metode baru untuk meringkas data agar mudah dipahami serta kegunaannya bagi pemilik data. Dari definisi yang telah dikemukakan tersebut, dapat dipahami bahwa data mining bertujuan untuk menemukan informasi penting yang ada dalam data. Data mining secara umum dibagi dua yakni *Supervised Learning* dan *Unsupervised Learning*. *Supervised Learning* merupakan pengklasifikasian data berdasarkan suatu label tertentu, sementara *Unsupervised Learning* data dikelompokkan tanpa mengikuti label tertentu. *Supervised Learning* memiliki beberapa metode seperti *Artificial Neural Network* (ANN), *Naive Bayes*, SVM dan lainnya. *Unsupervised Learning* berupa analisis kluster seperti *Ward*, *Single Linkage*, *K-Means*, *Fuzzy C-Means* dan lainnya.

## 2.5 Analisis Kluster (*Cluster Analysis*)

Dalam data mining, analisis kluster (*cluster analysis*) disebut juga *Unsupervised Learning* karena mengelompokkan data tanpa label tertentu. Tan (2006) dalam Prasetyo (2012) mendefinisikan analisis kluster sebagai pekerjaan mengelompokkan data (objek) yang didasarkan hanya pada informasi yang ditemukan dalam data yang menggambarkan objek tersebut dan hubungan diantaranya. Analisis kelompok merupakan metode analisis untuk mengelompokkan objek-objek pengamatan menjadi beberapa kelompok, sehingga akan diperoleh kelompok dimana objek-objek dalam satu kelompok memiliki banyak persamaan sedangkan dengan anggota kelompok lain memiliki banyak perbedaan (Johnson dan Winchern, 2007). Jadi, tujuan pengelompokan adalah agar objek-objek yang bergabung dalam sebuah kelompok merupakan

objek-objek yang mirip (atau berhubungan) satu sama lain dan berbeda (tidak berhubungan) dengan objek dalam kelompok yang lain. Lebih besar kemiripannya (homogenitas) dalam kelompok dan lebih besar perbedaannya diantara kelompok yang lain (Prasetyo, 2012).

Teknik pengelompokan banyak diterapkan dalam berbagai bidang seperti kedokteran, kesehatan, psikologi, hukum, statistik, astronomi, klimatologi dan sebagainya. Dalam bidang kedokteran, teknik pengelompokan dapat digunakan untuk mengelompokkan jenis-jenis penyakit berbahaya berdasarkan karakteristik/sifat-sifat penyakit pasien. Dalam bidang klimatologi dapat digunakan untuk mengetahui pola angin dan kondisi cuaca di udara sehingga bisa diketahui wilayah-wilayah yang rentan terhadap cuaca buruk dan sebagainya (Prasetyo, 2012). Secara umum analisis kluster dibagi menjadi dua, *Hierarchical Clustering* dan *Non-Hierarchical Clustering* atau *Partitioning*.

Menurut Simamora (2005) dalam melakukan analisis kluster ada beberapa tahapan, yaitu:

1. Merumuskan masalah dengan menjelaskan variabel-variabel yang menjadi dasar analisis kluster.
2. Menentukan ukuran jarak yang dipakai.
3. Menentukan metode atau prosedur pengklasteran yang digunakan.
4. Tentukan jumlah kluster
5. Interpretasikan kluster-kluster yang dibentuk. Kluster-kluster yang dihasilkan harus diinterpretasikan berdasarkan variabel-variabel yang dipakai untuk mengkluster.

#### 2.5.1 Ukuran Jarak Kedekatan

Salah satu hal yang paling penting dalam mengidentifikasi kelompok pengamatan yang mungkin ada dalam data adalah pengetahuan tentang bagaimana individu 'dekat' satu sama lain, atau seberapa jauh jaraknya (Everitt, 2011). Ukuran jarak yang sering digunakan adalah ukuran jarak Euclidean. Jarak Euclidean merupakan jarak antar objek, misalkan dua objek ke- $i$  dan ke- $j$  yang berada pada  $p$

dimensi berikut (Johnson dan Winchen, 2007). Persamaanya dapat dituliskan sebagai berikut:

$$d_{ij} = \sqrt{\sum_{k=1}^p (X_{ik} - X_{jk})^2} \quad (2)$$

Dimana:

$d_{ij}$  = jarak antara objek ke- $i$  dan objek ke- $j$

$p$  = Jumlah variabel kluster

$X_{ik}$  = data dari subjek ke- $i$  pada variabel ke- $k$

$X_{jk}$  = data dari subjek ke- $j$  pada variabel ke- $k$

### 2.5.2 Asumsi Analisis Kluster

Menurut Hair et al. (2006) setidaknya ada dua asumsi dalam analisis kluster yaitu:

#### 1. Kecukupan Sampel (sampel yang representatif)

Sampel yang mewakili atau sampel yang representatif adalah sampel yang diambil dapat mempresentasikan atau mewakili populasi yang ada. Pengujian sampel yang mewakili dapat dilakukan dengan uji Kaiser-Mayer-Olkin (KMO). Uji KMO banyak digunakan untuk melihat syarat kecukupan suatu sampel. Uji KMO ini mengukur kecukupan *sampling* secara menyeluruh dan mengukr kecukupan *sampling* untuk setiap indikator. Jika nilai KMO berkisar 0,5 sampai 1 maka sampel dapat dikatakan mewakili populasi atau sampel representatif. KMO dirumuskan sebagai berikut (Widarjono, 2010) (dalam Kurniawati et al. 2015):

$$KMO = \frac{\sum_{j=1}^p \sum_{k=1}^p r_{X_j X_k}^2}{\sum_{j=1}^p \sum_{k=1}^p r_{X_j X_k}^2 + \sum_{j=1}^p \sum_{k=1}^p \rho_{X_j X_k, X_l}^2} \quad (3)$$

dimana:

$$r_{X_j X_k} = \frac{\frac{1}{n} \sum_{i=1}^n (X_{ij} - \bar{X}_j)(X_{ik} - \bar{X}_k)}{\sqrt{\frac{\sum_{i=1}^n (X_{ij} - \bar{X}_j)^2}{n}} \sqrt{\frac{\sum_{i=1}^n (X_{ik} - \bar{X}_k)^2}{n}}} \quad (4)$$

$$\rho_{X_j X_k, X_l} = \frac{r_{X_j X_k} - r_{X_j X_l} r_{X_k X_l}}{\sqrt{(1 - r_{X_j X_l}^2)(1 - r_{X_k X_l}^2)}} \quad (5)$$

dengan:

$r_{X_j X_k}$  = korelasi antar variabel  $X_j$  dan  $X_k$

$\bar{X}_j$  = rata-rata variabel  $X_j$

$\bar{X}_k$  = rata-rata variabel  $X_k$

$n$  = banyaknya observasi (objek)

## 2. Tidak Ada Multikolinearitas (Non Multikolinearitas)

Multikolinearitas adalah adanya hubungan linear yang sempurna atau pasti di antara beberapa atau semua variabel (Gujarati, 2010). Multikolinearitas berkenaan dengan terdapatnya lebih dari satu hubungan linear pasti. Untuk mengetahui adanya multikolinearitas salah satunya adalah dengan menghitung nilai *Variance Inflation Factor* (VIF) dengan rumus :

$$VIF_j = \frac{1}{1 - R_j^2} \quad (6)$$

Dikatakan terjadi multikolinearitas apabila nilai  $VIF_j \geq 10$ . Selain menggunakan nilai VIF juga bisa menggunakan matriks korelasi. Menurut yamin dan Kurniawan (dalam Puspitasari dan Susanti, 2016) dikatan terjadi multikolinearitas apabila nilai korelasi antar variabel  $> 0,70$ . Apabila terjadi multikolinearitas maka harus dilakukan tindakan perbaikan multikolinearitas.

### 2.6 *Principal Component Analysis* (PCA)

*Principal Component Analysis* (PCA) atau Analisis Komponen Utama (AKU) merupakan salah satu solusi dalam analisis kluster jika terjadi multikolinearitas dalam data. PCA bertujuan untuk mereduksi variabel menjadi lebih sedikit dari jumlah variabel sebelumnya dimana  $K$  variabel baru  $< K$  variabel lama. *Principal component* (PC) merupakan suatu kombinasi linear dari variabel-variabel asal. Pembentukan PC berdasarkan dua cara yaitu matriks kovarian atau matriks korelasi (Johnson dan Wichern, 2007). Tahapan menentukan PC berdasarkan matriks korelasi adalah sebagai berikut:

1. Membuat matriks  $Z$  yang berisi data dari variabel  $X$  yang telah distandarisasi.
2. Membuat matriks korelasi dari  $Z$  yaitu  $Z'Z$ . Pereduksian PC dimulai dengan cara mencari nilai eigen  $\lambda_1, \lambda_2, \dots, \lambda_p$  yang diperoleh dari persamaan:

$$|Z'Z - \lambda I| = 0 \quad (7)$$

dimana jumlahan nilai eigen merupakan jumlah diagonal matriks korelasi, yaitu:

$$\sum_{j=1}^n \lambda_j = tr(Z'Z) \quad (8)$$

Nilai eigen selalu diurutkan dari yang terbesar sampai nilai terkecil. Nilai eigen menunjukkan besarnya total varian yang dijelaskan oleh  $PC$  yang terbentuk.  $PC_j$  saling orthogonal dan dibentuk berdasarkan persamaan:

$$PC_j = \gamma_{1j}z_1 + \gamma_{2j}z_2 + \dots + \gamma_{pj}z_p \quad (9)$$

Vektor eigen  $\gamma_j$  diperoleh dari setiap nilai eigen  $\lambda_j$  yang memenuhi persamaan:

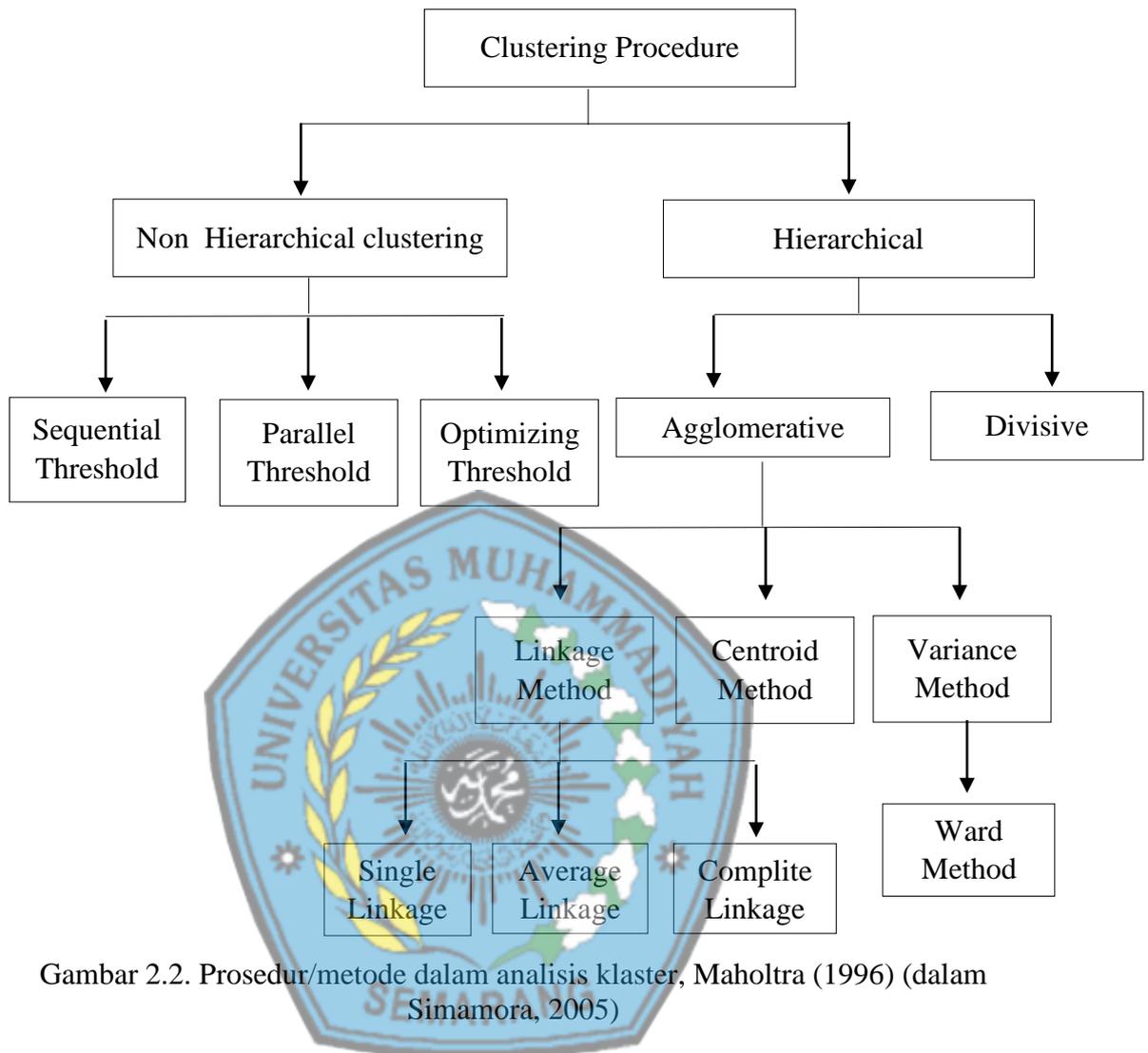
$$(Z'Z - \lambda_j I)\gamma_j = 0 \quad (10)$$

Jumlah *Principal Component* (PC) yang terpilih berdasarkan nilai eigen ( $\lambda$ ) dimana nilai  $\lambda > 1$  maka PC tersebut akan dipilih (Hardika et al, 2013)

## 2.7 Metode Pengklasteran

Dalam analisis data mining, analisis kluster merupakan metode pengelompokan *Unsupervised* yakni pengelompokan tanpa label atau arahan. Selain itu metode *Unsupervised* memiliki beberapa asumsi yang harus dipenuhi seperti kecukupan sampel dan non-multikolinearitas. Dalam analisis kluster secara umum ada dua metode yang sering digunakan yaitu:

1. *Hierarchical Clustering* (metode Hirarki)
2. *Non-Hierarchical Clustering/Partitioning*



Gambar 2.2. Prosedur/metode dalam analisis kluster, Maholtra (1996) (dalam Simamora, 2005)

## 2.8 Hierarchical Clustering (Metode Hirarki)

*Hierarchical Clustering* adalah analisis pengelompokan yang hasil pengelompokannya disajikan secara berjenjang dari  $n$ ,  $(n-1)$  sampai terdapat satu kelompok. Metode kluster hirarki secara umum terbagi dua yakni metode *agglomerative* dan *divisive*. Terdapat beberapa teknik metode *agglomerative* diantaranya, metode single linkage, complete linkage, average linkage, dan ward's. Adapun langkah-langkah analisis kluster dalam metode *agglomerative* adalah sebagai berikut:

1. Dimulai dengan  $N$  kluster dimana setiap objek dianggap sebagai kluster, kemudian membuat matriks jarak berukuran  $N \times N$  (matriks similaritas)  $D = \{d_{ij}\}$ .

2. Menemukan pasangan kluster dalam matriks yang mempunyai jarak paling dekat, seperti kluster U dan V maka jaraknya menjadi  $d_{ij}$ .
3. Gabungkan kluster U dan V menjadi sebuah kluster baru dan beri nama kluster UV, kemudian perbaharui jarak matriks.
4. Mengulangi langkah di atas sampai semua objek tergabung dalam satu kluster.

(Johnson dan Wichern, 2007).

### 2.8.1 Single Linkage

Single Linkage (metode pautan tunggal) merupakan metode analisis kluster yang didasarkan pada jarak minimum atau aturan tetangga terdekat. Dua objek pertama yang dikelompokkan adalah yang memiliki jarak terdekat diantara keduanya. Selanjutnya jarak terdekat lainnya dideteksi. Dalam metode ini, bila ada dua kluster hendak digabungkan atau tidak maka yang perlu diperhatikan cukup satu anggota dari masing-masing kluster yang keduanya saling berhubungan atau berjarak paling dekat (Simamora, 2005).

Untuk menentukan jarak antar kluster dengan menggunakan *single linkage* dapat dilakukan dengan melihat jarak antar dua kluster yang ada kemudian memilih jarak paling dekat atau aturan tetangga dekat (*nearest-neighbour rule*). Dihitung dengan cara:

$$d_{(UV)W} = \min \{d_{UW}, d_{VW}\} \quad (11)$$

Dimana nilai  $d_{UW}$  dan  $d_{VW}$  merupakan jarak minimum antara kluster U dan W serta kluster V dan W (Johnson dan Wichern, 2007). Berikut langkah-langkah penyelesaian dengan metode *single linkage*:

1. Bentuk matriks jarak untuk matriks data yang diberikan, misalkan

$$D(1)_{n \times n} = \begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1n} \\ d_{21} & d_{22} & \cdots & d_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \cdots & d_{nn} \end{bmatrix} \quad (12)$$

2. Asumsikan setiap data dianggap sebagai kluster, kemudian tentukan kluster yang mempunyai jarak terdekat, misal kluster U dan kluster V mempunyai

jarak terdekat kemudian digabungkan, hasil gabungannya adalah kluster UV.

3. Dari kluster UV yang telah terbentuk, cari jarak minimum antar kluster UV dengan kluster (objek) lainnya yang belum bergabung. Matriks jarak baru yang diperoleh yaitu  $D(2)$ . Misalkan  $d_{(UV)W} = \min \{d_{UW}, d_{VW}\}$  maka kluster yang baru terbentuk adalah (UVW).
4. Ulangai langkah 2 sampai semua objek bergabung menjadi satu kelompok.

### 2.8.2 Metode Ward

Penggunaan metode Ward untuk mengelompokkan suatu objek tertentu, secara khusus pertama kali diperkenalkan oleh Joe H. Ward tahun 1963. Metode *Ward* adalah satu-satunya di antara metode pengelompokan aglomeratif yang didasarkan pada kriteria *sum-of-squares* klasik (Murtagh dan Legendre, 2014). Metode *Ward* Disebut dikenal dengan metode varians sebab metode ini bertujuan untuk memperoleh cluster yang memiliki varians internal cluster yang sekecil mungkin. Metode varians yang umum dipakai adalah metode *ward* dimana rata-rata untuk setiap kluster dihitung. Lalu, dihitung jarak Euclidean antara setiap obyek dan nilai rata-rata itu, lalu jarak itu dihitung semua. Pada setiap tahap, dua kluster yang memiliki kenaikan '*sum of squares* dalam kluster' yang terkecil digabungkan (Simamora, 2005). *Sum of Square Error* (SSE) dapat dihitung dengan persamaan berikut:

$$SSE = \sum_{j=1}^p \left( \sum_{i=1}^n X_{ij}^2 - \frac{1}{n} \left( \sum_{i=1}^n X_{ij} \right)^2 \right) \quad (13)$$

Adapun langkah-langkah pengelompokan menggunakan metode *ward* adalah sebagai berikut:

1. Dimulai dari setiap objek yang dianggap sebagai sebuah kluster tersendiri, maka akan terdapat N kluster yang mempunyai satu objek. Pada tahap ini SSE bernilai nol.
2. Menghitung nilai SSE untuk setiap kombinasi dua pasangan kluster dari N kluster, lalu memilih dua pasangan kluster yang memiliki nilai SSE terkecil

untuk digabungkan menjadi satu kluster. Secara sistematis, N kluster akan berkurang 1 pada setiap tahap (N-1).

3. Membuat kombinasi dua pasangan kluster baru yang terdiri dari satu kluster yang telah terbentuk dan kluster yang lain, kemudian menghitung kembali nilai SSE. Memilih dua pasang kluster yang memiliki nilai SSE terkecil untuk digabungkan menjadi satu kluster.
4. Mengulangi langkah (3) sampai semua objek bergabung menjadi satu kluster.

(Johnson dan Wichern, 2007)

## 2.9 Non-Hierarchical Clustering (Kluster Tak Hirarki)

*Non-Hierarchical Clustering* atau *Partitioning* adalah salah satu metode pengelompokan dimana objek akan dikelompokkan ke dalam K kelompok yang telah ditentukan sebelumnya.

### 2.9.1 K-Means

Dalam statistik dan mesin pembelajaran, pengelompokan *K-Means* merupakan metode analisis kelompok yang mengarah pada pemartisan N objek pengamatan ke dalam K kelompok (*cluster*) dimana setiap objek pengamatan dimiliki oleh sebuah kelompok dengan *mean* (rata-rata) terdekat.

*K-Means* merupakan salah satu metode pengelompokan data non hirarki (sekatan) yang berusaha mempartisi data yang ada kedalam kelompok sehingga data berkarakteristik sama dimasukkan kedalam satu kelompok yang sama dan data yang berkarakteristik berbeda dikelompokkan kedalam kelompok lain (Prasetyo, 2012). Menurut MacQueen (1967), metode ini dimulai dengan menentukan jumlah cluster terlebih dahulu yang diinginkan misalkan *c cluster*. Langkah berikutnya adalah K-Means.

Pengelompokan data dengan metode K-Means secara umum dilakukan dengan algoritma sebagai berikut (Prasetyo, 2012):

1. Menentukan k sebagai jumlah kluster yang ingin dibentuk.
2. Mengalokasikan data secara acak.

- Menentukan pusat kluster dari data yang ada pada masing-masing kluster dengan persamaan:

$$C_{kj} = \frac{x_{1j} + x_{2j} + \dots + x_{nj}}{n} \quad (14)$$

dimana:

$C_{kj}$  = pusat kluster ke- $k$  pada variabel ke- $j$  ( $j=1,2,\dots,p$ )

$n$  = banyak data pada kluster ke- $k$

- Menentukan jarak setiap objek dengan setiap centroid dengan perhitungan jarak setiap objek dengan setiap centroid menggunakan jarak euclidean

- Menghitung fungsi objektif

$$J = \sum_{i=1}^n \sum_{j=1}^k a_{ij} d(x_i, C_{kj})^2 \quad (15)$$

- Mengalokasikan masing-masing data ke centroid/rata-rata terdekat yang dirumuskan.

$$a_{ij} = \begin{cases} 1, & s = \min \{d(x_i, C_{kj})\} \\ 0, & \text{lainnya} \end{cases} \quad (16)$$

$a_{ij}$  adalah nilai keanggotaan titik  $x_i$  ke pusat kluster  $C_{kj}$ ,  $s$  adalah jarak terpendek dari data  $x_i$  ke pusat kluster.

- Mengulangi langkah 3 sampai 6 hingga tidak adalagi objek yang berpindah.

### 2.9.2 Fuzzy C-Means (FCM) clustering

Pengelompokan dengan metode Fuzzy C-Means (FCM) didasarkan pada teori logika fuzzy. Teori ini pertama kali diperkenalkan oleh Lotfi Zadeh pada tahun 1965 dengan nama himpunan fuzzy (*fuzzy set*). Dalam teori fuzzy, keanggotaan sebuah data tidak diberi nilai secara tegas dengan nilai 1 (menjadi anggota) dan 0 (tidak menjadi anggota) melainkan dengan suatu nilai derajat keanggotaan yang jangkauan nilainya 0 sampai 1. Nilai keanggotaan suatu data dalam sebuah himpunan menjadi 0 ketika data sama sekali bukan anggota, dan 1 ketika data

menjadi anggota secara penuh dalam suatu himpunan. Umumnya nilai keanggotaannya antara 0 dan 1. Semakin tinggi nilai keanggotaannya maka semakin tinggi derajat keanggotaannya dan sebaliknya (Prasetyo, 2012).

FCM juga merupakan suatu teknik pengklasteran data yang mana keberadaan tiap-tiap data dalam suatu kluster ditentukan oleh nilai keanggotaan. Teknik ini pertama kali diperkenalkan oleh Dunn tahun 1973 kemudian Jim Bezdek pada tahun 1981. Konsep dasar FCM, pertama kali adalah menentukan pusat kluster yang akan menandai lokasi rata-rata untuk tiap kluster. Pada kondisi awal, pusat kluster masih belum akurat. Tiap-tiap data yang memiliki derajat keanggotaan untuk tiap-tiap kluster. Dengan cara memperbaiki pusat kluster dan nilai keanggotaan tiap-tiap data secara berulang, maka akan dapat dilihat bahwa pusat kluster akan bergerak menuju lokasi yang tepat (Kusumadewi, 2006) .

Adapun Algoritma FCM adalah sebagai berikut (Kusumadewi, 2006):

1. Tentukan:
  - a. Matriks  $X$  berukuran  $n \times m$ , dengan  $n$  = jumlah data yang akan dikluster; dan  $m$  = jumlah variabel.
  - b. Jumlah kluster yang akan dibentuk =  $C (\geq 2)$ .
  - c. Pangkat (pembobot) =  $W (>1)$
  - d. Maksimum iterasi
  - e. Kriteria penghentian =  $\xi$  (nilai positif yang sangat kecil)
  - f. Iterasi awal,  $t=1$  dan  $\Delta = 1$ .
2. Bentuk matriks partisi awal  $U^0$  sebagai berikut:

$$U = \begin{bmatrix} \mu_{11}(x_1) & \mu_{12}(x_2) & \dots & \mu_{1n}(x_n) \\ \mu_{21}(x_1) & \mu_{22}(x_2) & \dots & \mu_{2n}(x_n) \\ \vdots & & & \vdots \\ \mu_{c1}(x_1) & \mu_{c2}(x_2) & \dots & \mu_{cn}(x_n) \end{bmatrix} \quad (17)$$

(matriks partisi awal biasanya dipilih secara acak)

3. Hitung pusat kluster  $V$  untuk setiap kluster:

$$V_{ij} = \frac{\sum_{k=1}^n (\mu_{ik})^w x_{kj}}{\sum_{k=1}^n (\mu_{ik})} \quad (18)$$

4. Perbaiki derajat keanggotaan setiap data pada setiap kluster (perbaiki matriks partisi), sebagai berikut:

$$\mu_{ik} = \left[ \sum_{j=1}^c \left( \frac{d_{ik}}{d_{jk}} \right)^{2/(w-1)} \right]^{-1} \quad (19)$$

Dengan :

$$d_{ik} = d(x_k - V_i) = \left[ \sum_{j=1}^m (x_{kj} - V_{ij}) \right]^{1/2} \quad (20)$$

5. Tentukan kriteria berhenti yaitu perubahan matriks partisi pada iterasi sekarang dengan iterasi sebelumnya sebagai berikut :

$$\Lambda = \|U^t - U^{t-1}\| \quad (21)$$

Apabila  $\Lambda \leq \xi$  maka iterasi dihentikan, namun apabila  $\Lambda > \xi$  maka naikan iterasi ( $t = t+1$ ) dan kembali ke langkah 3

## 2.10 Pemilihan Metode Terbaik

Pemilihan metode terbaik dimaksudkan untuk melihat kinerja masing-masing pengelompokan khususnya dalam mengelompokan kabupaten/kota di Jawa Tengah. Untuk mengukur kinerja metode analisis yang digunakan, menurut Bunkers (dalam Komariah dan Akbar, 2011) dilakukan evaluasi kelompok dengan kriteria dua nilai simpangan baku, yaitu simpangan baku dalam kelompok ( $S_w$ ) dan antar kelompok ( $S_B$ ) sehingga diperoleh metode terbaik, berikut persamaan  $S_w$  :

$$S_w = K^{-1} \sum_{k=1}^K S_k \quad (22)$$

$S_w$  = rata-rata simpangan baku dalam kluster

$S_k$  = simpangan baku kluster ke- $k$

Dimana  $K$  adalah banyaknya kelompok yang terbentuk dan  $S_k$  merupakan simpangan baku kelompok ke- $k$ . Nilai  $S_B$  dicari dengan persamaan berikut:

$$S_B = \left[ (K-1)^{-1} \sum_{k=1}^K (\bar{x}_k - \bar{x})^2 \right]^{1/2} \quad (23)$$

$S_B$  = simpangan baku antar kelompok

Dimana  $\bar{x}_k$  adalah rata-rata kelompok ke-k dan  $\bar{x}$  adalah rata-rata keseluruhan kelompok. Semakin kecil nilai  $S_w$  dan semakin besar nilai  $S_B$  maka metode tersebut memiliki kinerja yang baik, artinya mempunyai homogenitas yang tinggi (Komariah dan Akbar, 2011).

