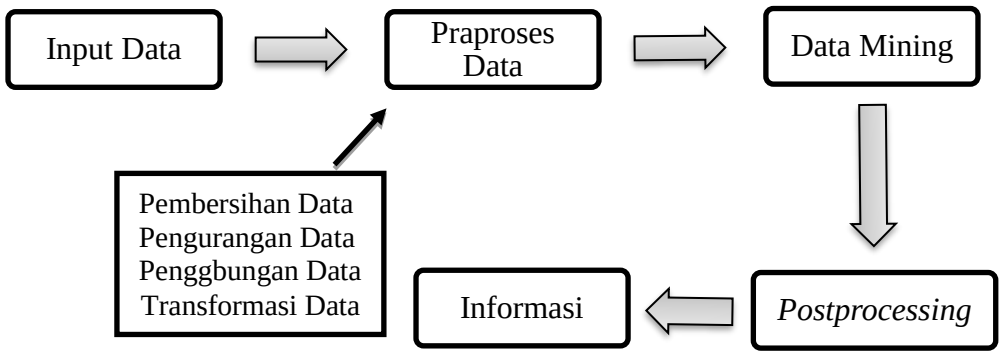
# BAB II

# TINJAUAN PUSTAKA

## 2.1 Data Mining

Menurut Tan, Steinbach, dan Kumar (2006), data mining merupakan suatu proses untuk menemukan informasi yang menarik dan tersembunyi dari suatu kumpulan data yang berukuran besar yang tersimpan dalam suatu basis data, data warehouse atau tempat penyimpanan data lainnya. Teknik-teknik data mining yang digunakan bertugas untuk menemukan pola baru dan bermakna di dalam basis data yang mungkin masih belum diketahui. Data mining juga merupakan bagian integral dari Knowledge Discovery in Databases (KDD).




**Gambar 2.1. Proses dari KDD (Tan, Steinbach, dan Kumar, 2006)**

Berdasarkan gambar tersebut, data masukan mengalami 3 proses sebelum menjadi hasil yang berupa informasi yaitu praproses data, data mining, dan

postprocessing. Praproses bertujuan untuk mentransformasi data ke dalam format sesuai dengan kebutuhan. Tahapan praproses antara lain adalah pembersihan data untuk membuang data-data yang tidak digunakan dan data duplikat, pengurangan data, penggabungan data, dan transformasi atau normalisasi data. Postprocessing bertujuan untuk membantu pengguna dalam memahami informasi. Kualitas informasi yang dihasilkan oleh proses KDD sangat dipengaruhi oleh kualitas data, pengetahuan tentang data, dan teknik pengolahan data yang akan digunakan.

### 1.1.1 Tipe Data

Menurut Kamus Besar Bahasa Indonesia (KBBI), data merupakan keterangan yang benar dan nyata. Menurut Anderson, Sweeney, dan Williams (2012), data adalah fakta dan gambaran yang dikumpulkan, dianalisis, dan dirangkum untuk diinterpretasikan dalam presentasi. Menurut Anderson dan Sclove (1974), data yang digunakan dalam analisis statistik (data statistik) berdasarkan jenis variabelnya dikelompokkan menjadi dua, yaitu data numerik dan data kategorik. Data numerik merupakan data dengan variable kuantitatif yang menghasilkan informasi numerik. Data numerik dapat dikelompokkan menjadi dua yaitu data diskrit (hasil pencacahan) dan data kontinu (hasil pengukuran). Agresti (2007) menyatakan bahwa data kategorik memiliki skala pengukuran yang terdiri atas satu set kategori, misalnya filsafat politik yang dapat diukur sebagai kategori liberal, kategori moderat, atau kategori konservatif.

### 1.1.2 Praproses Data

Praproses data dilakukan karena data awal biasanya tidak bersih, tidak lengkap dan tidak konsisten. Sehingga, praproses data bertujuan untuk meningkatkan kualitas data sehingga diharapkan dapat membantu meningkatkan akurasi, efektifitas, dan efisiensi dari suatu proses analisis data mining. Praproses data juga merupakan langkah yang sangat penting dalam proses KDD karena kualitas hasil akhir suatu proses data mining sangat dipengaruhi oleh kualitas data. Praproses data juga bertujuan untuk mentranformasi data input ke dalam format sesuai dengan kebutuhan. Pembersihan data, pengurangan data, penggabungan data, dan transformasi data merupakan bagian dari praproses data (Han dan Kamber, 2001).

a. Pembersihan data

Pembersihan data dilakukan karena data penelitian seringkali memiliki *record* dengan nilai atribut yang tidak lengkap, nilai kosong, tidak konsisten, dan noisy. Data yang memiliki atribut dengan nilai tidak lengkap atau kosong dapat diatasi dengan beberapa cara yaitu menghapus data tersebut, isi atribut kosong dengan rata-rata nilai atribut atau isi atribut kosong dengan nilai atribut yang paling sering muncul (Han dan Kamber, 2001). Nilai tidak konsisten adalah nilai yang berada diluar kesepakatan. Data noisy adalah kesalahan tidak berpola atau perbedaan yang terjadi pada peubah yang diukur (Tan, Steinbach, dan Kumar, 2006).

b. Pengurangan data

Pengurangan data biasanya dikaitkan dengan data yang sangat besar yang merupakan suatu usaha yang digunakan untuk mengurangi ukuran data dengan tujuan untuk memperoleh data dengan volume yang relatif kecil tetapi dapat mewakili kondisi data asli. Memproses data hasil pengurangan seharusnya jauh lebih efisien dibanding dengan memproses data asli tetapi mendapatkan hasil yang relatif sama. Seleksi atribut dan seleksi *record* merupakan sebagian dari teknik pengurangan data

c. Seleksi atribut

Data yang akan dianalisis memungkinan memiliki atribut dengan jumlah yang cukup banyak tetapi sebagian dari atribut data tersebut tidak relevan dengan kebutuhan penelitian sehingga diperlukan seleksi atribut. Sebagai contoh, dilakukan analisis clustering terhadap data siswa sekolah untuk menemukan karakteristik siswa yang berkaitan dengan Indeks Prestasi Akademik, maka atribut data pribadi seperti Nama, Alamat, atau Nomor Telepon merupakan atribut yang tidak relevan dengan kebutuhan penelitian. Jika atribut data pribadi tersebut diikutkan ke dalam proses clustering, maka dapat memperlambat proses penelitian dan akan mendapatkan hasil yang kurang berkualitas. Seleksi atribut adalah suatu usaha untuk mengurangi ukuran data dengan cara menghapus atribut yang tidak relevan dengan kebutuhan penelitian (Han dan Kamber, 2006).

d. Seleksi *Record*

Secara umum, karakteristik data mining adalah menganalisis data dengan ukuran yang sangat besar berdasarkan sampel dari data tersebut. Sampel digunakan untuk memberikan informasi terkait dengan keseluruhan data. Kualitas dari informasi yang dihasilkan tergantung dari data objek yang akan dipilih sebagai sampel. Seleksi *record* adalah suatu usaha untuk mendapatkan data sampel yang representatif dengan data asli.

e. Penggabungan data

Pada proses data mining seringkali dibutuhkan suatu proses penggabungan data. Penggabungan dilakukan karena data yang akan dianalisis berasal dari beberapa sumber yang berbeda. Sumber tersebut dapat berupa multiple databases, data cubes, atau flat file.

f. Tranformasi

Secara prinsip, data kategori dapat ditransformasi/dikonversi ke dalam bilangan numerik, dimana satu bilangan numerik mewakili satu nilai kategori. Atribut kategori yang demikian disebut dengan "*dummy variable*" (Kandardzic, 2011). Dalam suatu data numerik kadang-kadang terdapat atribut yang memiliki nilai dengan rentang yang berbeda dengan atribut lain atau memiliki satuan yang berbeda. Untuk beberapa algoritma data mining, kondisi demikian dapat mengacaukan hasil perhitungan proximity (Tan, Steinbach, dan Akasapu, 2006). Atribut dengan rentang nilai besar menjadi sangat dominan, dan akan

mempengaruhi hasil secara tidak proporsional. Maka dari itu, perlu dilakukan standarisasi terhadap semua atribut sehingga setiap atribut penelitian memiliki kontribusi secara proporsional terhadap hasil akhir suatu proses data mining

## 2.2 Analisis Multivariat

Analisis multivariat merupakan analisis yang berkaitan dengan jumlah variabel lebih dari dua yang dianalisis secara simultan pada masing-masing pengamatan (Johnson dan Wichern, 2007). Terdapat beberapa teknik analisis multivariat yang sering digunakan seperti Analisis Faktor, *Multiple Regression*, Analisis Diskriminan, Analisis Korelasi Kanonikal, Analisis Multivariat dari varian dan kovarian, Analisis Klaster, Analisis Korespondensi, Analisis Gabungan, *Multidimensional Scalling*, *Principal Component Analysis* (PCA), dan *Structural Equation Modelling* (SEM) (Hair *et al*, 2010).

## 2.3 Analisis Kelompok (*Cluster Analysis*)

Analisis klaster merupakan suatu metode multivariat yang bertujuan untuk mengelompokan sampel subyek atas dasar satu set peubah yang diukur menjadi beberapa kelompok yang berbeda sehingga subyek yang sama ditempatkan dalam kelompok yang sama (Cornish, 2007). Analisis klaster atau pengelompokan berbeda dengan metode klasifikasi. Metode klasifikasi berkaitan dengan sejumlah kelompok yang telah diketahui sebelumnya dan tujuan operasionalnya adalah untuk menetapkan pengamatan baru. Sedangkan analisis klaster merupakan suatu teknik data mining

untuk mengelompokan himpunan objek ke dalam beberapa *cluster* hanya berdasarkan kemiripan karakteristik dari atribut yang dimiliki oleh data objek sedemikian sehingga data objek yang berada di dalam *cluster* yang sama memiliki kemiripan satu sama lain tetapi tidak mirip dengan data objek yang berada dalam *cluster* yang berbeda menurut Han dan Kamber (2001).

a. Ukuran Kemiripan

Ukuran kemiripan digunakan untuk mencari pasangan antar objek yang mirip dalam data. Kemiripan antar pasangan objek $x$ dan $y$ dinyatakan dengan $si(x , y)$. $si(x , y)$ akan bernilai besar jika $x$ dan $y$ merupakan pasangan objek yang berkarakteristik mirip, sebaliknya $si(x , y)$ akan bernilai kecil jika $x$ dan $y$ merupakan pasangan objek memiliki karakteristik tidak mirip. Untuk setiap pasangan objek dan, berlaku 3 kondisi berikut (Kandardzic, 2011):

1. $0 \leq si\ (x , y) \leq 1$, kemiripan bernilai 0 dan 1
2. $si\ (x , x) = 1$, setiap objek mirip dengan dirinya sendiri.
3. $si\ (x , y) = si\ (y , x)$, kemiripan bersifat simetri.

b. Ukuran Tidak Kemiripan

Ukuran ketidakmiripan digunakan untuk mencari jarak antara pasangan objek di dalam data. Jarak antara pasangan objek dan dinyatakan dengan $d(x , y)$. $d(x , y)$ akan bernilai besar jika $x$ dan $y$ merupakan pasangan objek yang memiliki karakteristik tidak mirip, sebaliknya jika $d(x , y)$ akan bernilai kecil jika $x$ dan $y$

merupakan pasangan objek yang memiliki karakteristik mirip. Untuk setiap objek x dan y berlaku kondisi berikut (Han dan Kamber, 2001):

1. $d(x, y) \geq 0$, jarak merupakan bilangan non-negatif.
2. $d(x, x) = 0$, jarak suatu objek dengan dirinya sendiri = 0.
3. $d(x, y) = d(y, x)$, jarak bersifat simetri.

Terkait dengan pengertian dan tujuan dilakukannya analisis klaster, dapat dinyatakan bahwa suatu kelompok (*cluster*) yang baik adalah kelompok yang mempunyai, (Hair, Black, Babin, dan Anderson, 2009) :

1. Homogenitas (kesamaan) yang tinggi antara anggota dalam satu kelompok (*within-cluster*),
2. Heterogenitas (perbedaan) yang tinggi antara kelompok yang satu dengan kelompok yang lain (*between cluster*).

## 2.4 Metode Pengelompokan (*Clustering*)

Terdapat dua teknik yang sering digunakan dalam analisis *cluster* yaitu teknik hierarki dan non-hierarki. Pengelompokan dengan metode hierarki dapat dilakukan berdasarkan metode *agglomerative* (penggabungan) dan *divisive* (pemisahan). Metode *agglomerative* menggabungkan satu per satu observasi menjadi kelompok-kelompok yang ditentukan berdasarkan kemiripan antar kelompok. Penentuan kemiripan dilakukan dengan menghitung jarak antar kelompok. Sedangkan metode

devisive adalah memisahkan sebuah kelompok menjadi beberapa kelompok (Johnson dan Wichern, 2007).

Metode pengelompokan hieararki dan non-hierarki berfokus pada data dengan skala kontinu sehingga dapat digunakan dengan mudah untuk menghitung fungsi jarak antara dua objek, namun pada beberapa kasus terdapat data yang berskala kategori bahkan terdapat kasus dengan campuran data numerik dan kategorik Tahap pengelompokan dalam analisis *cluster* ini dibedakan menurut jenis data yang dimiliki. Analisis kelompok pada data kategorik tidak dapat diperlakukan seperti pada data numerik. Hal tersebut dikarenakan sifat khusus data kategorik, sehingga pengelompokan data kategorik menjadi lebih rumit dibandingkan pengelompokan untuk data numerik. (Hair, et.al, 2009).

### 2.4.1 *Clustering* Data Kategorik

Metode *clustering* hirarki dan non-hirarki dinilai tidak tepat digunakan pada data kategorik sehingga dikembangkan metode ROCK (*Robust Clustering using linKs*) untuk *clustering* data kategorik tersebut (Guha, Rastogi, dan Shim, 2000).

Metode ROCK menggunakan konsep *link* sebagai ukuran kemiripan antar sepasang data. Pengamatan yang memiliki tingkat hubungan (*link*) tinggi digabungkan dalam satu kelompok, sedangkan yang rendah dipisahkan dari data yang dikelompokkan. Algoritma ROCK akan berhenti jika jumlah kelompok yang diharapkan sudah terpenuhi atau tidak ada *link* antar kelompok. Metode ini cukup

efektif untuk menangani *outlier*. Pemangkasan *outlier* memungkinkan untuk membuang yang tidak ada tetangga, sehingga titik tersebut tidak berpartisipasi dalam pengelompokan. Namun dalam beberapa situasi, *outlier* dapat hadir sebagai *cluster-cluster* yang kecil (Guha, Rastogi, dan Shim, 2000).

*Clustering* untuk data kategorik dengan algoritma ROCK dapat dilakukan dengan tiga langkah. Adapun langkahnya sebagai berikut:

1. Menghitung kemiripan (similaritas) menggunakan rumus *Jaccard coefficient* (Rahayu, 2009). Ukuran kemiripan antara pasangan objek ke−i dan objek ke−j dihitung dengan rumusan yang didefinisikan pada persamaan (2.4)

   $$si(Xi, Xj) = \left|\frac{Xi \cap Xj}{Xi \cup Xj}\right| \ , Xi \neq Xj \qquad (2.1)$$

   Dimana :

   $i$ = 1, 2, 3, …, n d  j = 1, 2, 3, …., n

   $Xi$ = himpunan pengamatan ke-$i$ dan X$i$ = {$x_{1j}$,$x_{2j}$,…., $x_{mk\ j}$}

   $|X|$ = bilangan kardinal atau jumlah anggota dari himpunan
2. Langkah kedua adalah menentukan tetangga. Pengamatan dapat dinyatakan sebagai tetangga jika nilai si($X_i$ , $X_j$) ≥ θ
3. Langkah terakhir adalah menghitung *link* antar objek pengamatan. Besarnya *link* dipengaruhi oleh nilai *threshold* (θ) yang merupakan parameter yang ditentukan oleh peneliti yang dapat digunakan untuk mengontrol seberapa dekat hubungan antar objek. Besarnya nilai *threshold* (θ) yang diinputkan adalah 0 < θ < 1.

Metode ROCK menggunakan informasi tentang *link* sebagai ukuran kemiripan antar objek. Jika terdapat objek pengamatan $X_i$, $X_j$, dan $X_k$ dimana *Xi* tetangga dari *Xj*, dan *Xj* tetangga dari *Xk*, maka dapat dikatakan *Xi* memiliki *link* dengan *Xk* walaupun *Xi* bukan tetangga dari *Xk*. Cara untuk menghitung *link* untuk semua kemungkinan pasangan pada *n* objek, maka dapat menggunakan matriks *A*. Matriks *A* merupakan matriks yang berukuran *n x n* dan bernilai 1 jika *Xi* dan *Xj* dinyatakan mirip (tetangga) dan bernilai 0 dan jika *Xi* dan *Xj* tidak mirip (bukan tetangga). Jumlah *link* antar pasangan *Xi* dan *Xj* diperoleh dari hasil kali antara baris ke *Xi* dan kolom ke *Xj* pada matriks A. Jika *link* antara *Xi* dan *Xj* semakin besar maka semakin besar pula kemungkinan *Xi* dan *Xj* berada dalam satu kelompok (*cluster*) yang sama. Adapun metode penggabungan kelompok pada algoritma ROCK yang dapat digunakan yang didasarkan atas ukuran kebaikan (*goodness measure*) antar kelompok dengan rumusan pada persamaan (2.2) *Goodness measure* adalah persamaan yang digunakan untuk menghitung jumlah *link* dibagi dengan kemungkinan *link* yang terbentuk berdasarkan ukuran kelompoknya (Tyagi dan Sharma, 2012).

Persamaan pada ukuran kebaikan (2.2) :

$$g(Ci, Cj) = \frac{li\ [Ci,Cj]}{(ni+nj)^{1+2f(\theta)} - ni^{1+2f(\theta)} - nj^{1+2f(\theta)}} \qquad (2.2)$$

Dengan *li* [*Ci, Cj*] = $\sum_{X \epsilon ci\ xj \epsilon cj}$ , (*Xi,Xj*) yang menyatakan jumlah *link* dari semua kemungkinan pasangan objek yang ada dalam $C_i$ dan $C_j$ , serta $n_i$ dan $n_j$

masing-masing menyatakan jumlah anggota dalam kelompok ke-*i* dan ke-*j* , sedangkan $f(\theta) = \frac{1-\theta}{1+\theta}$

### 2.4.2 *Clustering* Data Numerik

*Clustering* data numerik dilakukan berdasarkan ukuran ketidakmiripan atau jarak untuk data numerik dimana jarak yang dapat digunakan adalah jarak *euclidean*. Misalkan terdapat dua observasi dengan variabel-variabel berdimensi m yaitu $x_i = [x_1, x_{2,\dots}, x_m]^T$ dan $x_j = [x_1, x_{2,\dots}, x_m]^T$ . Konsep jarak *euclidean* yang mengukur jarak antara observasi $x_i$ dan $x_j$ pada persamaan (2.3) berikut

$$d_i = \sqrt{\sum_{k=1}^{p}(X_i - X_j)^2} \qquad (2.3)$$

Dimana,

$d_i$ : jarak antara objek ke-i dan objek ke-j

p : jumlah peubah cluster

$X_{ij}$ : data dari subyek ke-i pada peubah ke-k

$X_j$ : data dari subyek ke-j pada peubah ke-k

Hasil pengelompokan disajikan dalam bentuk dendrogram (diagram pohon) yang memungkinkan penelusuran objek-objek yang diamati menjadi lebih mudah dan informatif. Teknik yang dapat digunakan untuk pengelompokan meliputi metode hirarki dan metode non hirarki.

Pengelompokan hirarki dimulai dengan dua atau lebih objek yang memiliki kesamaan paling dekat, kemudian proses diteruskan ke objek lain yang mempunyai kedekatan kedua. Analisis dilakukan hingga kelompok membentuk semacam "pohon", dimana ada hirarki (tingkatan) yang jelas antar objek pengamatan, dari yang paling mirip sampai paling tidak mirip. Dendrogram umumnya digunakan untuk membantu memperjelas proses hirarki tersebut (Hair, et.al, 2009). Terdapat dua teknik pengelompokan dalam analisis kelompok hirarki yaitu teknik pembagian (*divisive*) dan teknik penggabungan (*agglomerative*).

Teknik pembagian berawal dari satu kelompok yang berunsurkan semua objek yang ada. Kelompok ini kemudian dibagi menjadi dua kelompok, dan kemudian masing-masing kelompok dibagi lagi menjadi dua kelompok, dan kemudian masing-masing kelompok dibagi lagi menjadi dua kelompok, dan seterusnya. Berbeda dengan teknik pembagian, dalam teknik penggabungan setiap objek merupakan satu kelompok tersendiri. Lalu dua kelompok yang terdekat digabungkan dan seterusnya sehingga diperoleh satu kelompok yang berunsurkan semua objek pengamatan. Bila suatu kelompok merupakan penggabungan dari beberapa kelompok sebelumnya, maka diperlukan ukuran ketidakmiripan antar kelompok, kelompok-kelompok dengan ukuran ketidakmiripan terkecil digabungkan menjadi sebuah kelompok yang baru. (Alvionita, 2017).

Andaikan $d_{uv}$ merupakan ukuran ketakmiripan antara kelompok ke-u dengan kelompok ke-v dan $d_{w(u,v)}$ merupakan ukuran ketakmiripan antara kelompok ke-w

dengan kelompok (u,v) yang merupakan penggabungan antara kelompok ke-u dengan kelompok ke-v, maka beberapa teknik pengelompokan antara kelompok dinyatakan sebagai berikut, (Johnson dan Wichern, 2007).

a. Pautan Tunggal (*Single Linkage/Nearest Neighbor*), prosedur ini didasarkan pada jarak terkecil atau jarak terdekat antar objek. Jika dua objek terpisah oleh jarak yang pendek maka kedua objek tersebut digabung menjadi satu kelompok dan demikian saterusnya. Ukuran jarak yang digunakan adalah sebagai berikut

   $d_{w(u,v)} = \min (d_{wu}.d_{wv})$ (2.4)

   $d_{w(uv)}$ : jarak minimum antara kelompok UV dan W

   $d_{wu}$ : jarak kelompok U dan kelompok W

   $d_{wv}$ : jarak kelompok V dan kelompok W

b. Pautan Lengkap (*Complete Linkage/Farthest Neighbor*), berlawanan dengan *single linkage*, prosedur ini pengelompokannya berdasarkan jarak terbesar atau jarak terjauh pada antar objek pengamatan. Ukuran jarak yang digunakan adalah sebagai berikut

   $d_{w(u,v)} = \text{maks} (d_{wu}.d_{wv})$ (2.5)

   $d_{w(uv)}$ : jarak maksimum antara kelompok UV dan W

   $d_{wu}$ : jarak kelompok U dan kelompok W

   $d_{wv}$ : jarak kelompok V dan kelompok W

c. Pautan Rataan (*Average Linkage Between Method*/BAVERAGE), procedure ini hampir sama dengan *single linkage* maupun *complete linkage*, namun kriteria yang digunakan adalah rata-rata jarak seluruh objek dalam suatu kelompok dengan jarak seluruh objek dalam kelompok yang lain. Dengan $n_u$ dan $n_v$ merupakan jumlah pengamatan dalam kelompok ke-u dan ke-v, ukuran jarak yang digunakan adalah pada persamaan (2.4) berikut ini

$$d_{w(u,v)} = \frac{n_u}{n_u+n_v} d_{wu} + \frac{n_v}{n_u+n_v} d_{wv} \quad (2.6)$$

$d_{wu}$ : jarak kelompok U dan kelompok W

$d_{wv}$ : jarak kelompok V dan kelompok W

$n_u$ : jumlah pengamatan pada kelompok U

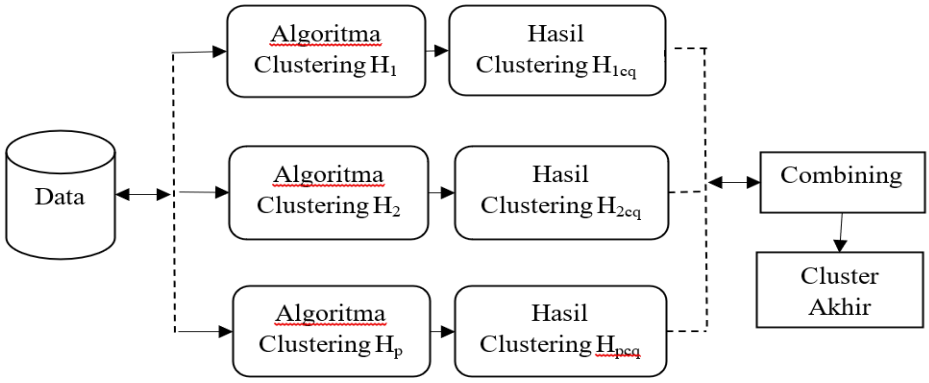
$n_v$ : jumlah pengamatan pada kelompok V

Kelebihan dalam metode hirarki adalah mempercepat pengolahan dan menghemat waktu karena data yang diinputkan membentuk hirarki (tingkatan) sehingga mempermudah dalam penafsiran.

### 2.4.3 *Clustering* Data Campuran

Analisis klaster terhadap data campuran diawali dengan membagi data menjadi dua, yaitu data murni numerik dan data murni kategori. Apabila terdapat data dengan variabel berskala campuran sebanyak m, dengan $m_{numerik}$ merupakan jumlah variable numerik, dan $m_{kategori}$ merupakan jumlah variabel kategori, sehingga $m = m_{numerik} + m_{kategori}$. Selanjutnya dilakukan pengelompokan data sesuai dengan jenis data secara terpisah. Hasil pengelompokan tersebut kemudian

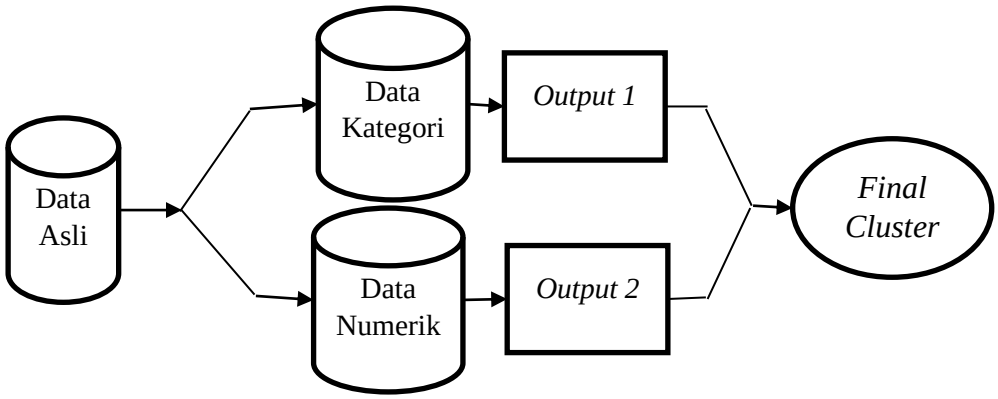
digabungkan menggunakan metode pengelompokan *ensemble*. Pengelompokan *ensemble* merupakan metode yang menggabungkan algoritma yang berbeda untuk mendapatkan partisi umum dari data dengan tujuan untuk konsolidasi dari portofolio hasil pengelompokan individu (Suguna dan Selvi, 2012). Tujuan pengelompokan *ensemble* adalah untuk menggabungkan hasil pengelompokan dari beberapa algoritma pengelompokan untuk mendapatkan hasil pengelompokan yang lebih baik dan *robust* (Yoon, Ahn, Lee, Cho, dan Kim, 2006). Secara umum dijelaskan dengan gambar 2.2 berikut




**Gambar 2.2 *Cluster* Campuran (Herliyasari, 2018)**

Tidak banyak algoritma pada metode *clustering* yang dikembangkan untuk memproses data dengan tipe campuran. Namun, Salah satu metode yang dapat digunakan adalah algCEBMDC (*Cluster Ensemble Based Mixed Data Clustering)* yang merupakan suatu algoritma *clustering* dengan pendekatan *cluster ensemble*. Algoritma algCEBMDC dikembangkan untuk menyelesaikan masalah yang berkaitan dengan *clustering* data dengan tipe campuran (kategorik dan numerik)

Pertama, data asli yang bertipe campuran dipisah menjadi dua yaitu data dengan tipe kategorik dan data dengan tipe numerik. Selanjutnya, kedua data tersebut diproses secara terpisah dengan menggunakan algoritma *clustering* yang sesuai dengan tipe masing-masing data. Terakhir, cluster yang dihasilkan oleh kedua algoritma digabungkan sebagai data baru dengan tipe kategorik, kemudian diproses dengan menggunakan algoritma *clustering* data kategorik untuk mendapatkan hasil akhir (Hee, Xu, dan Deng, 2005). Secara umum ditunjukkan oleh Gambar (2.3).




**Gambar (2.3). Tahapan *Cluster Ensemble* (Herliyasari, 2018)**

Berikut ini adalah langkah-langkah dalam analisis data campuran menggunakan Algortima CEBMDC (He, Xu, dan Deng, 2005).

1. Membagi data menjadi dua bagian, yaitu data murni skala numerik dan murni skala kategori.
2. Melakukan pengelompokan objek dengan variabel numerik dengan algoritma pengelompokan data skala numerik, serta melakukan pengelompokan objek dengan skala kategori dengan algoritma pengelompokan data kategori.

3. Menggabungkan (*combining*) hasil pengelompokan dari variabel data numerik dan data kategori yang disebut dengan proses *ensemble*.
4. Melakukan pengelompokan *ensemble* menggunakan algoritma pengelompokan data kategori untuk mendapatkan hasil dari kelompok akhir (*final cluster*).

## 2.5 Validasi Hasil *Clustering*

Setelah hasil kelompok diperoleh, untuk melihat kualitas struktur data dari kelompok yang terbentuk dapat dinilai menggunakan berbagai jenis ukuran validasi (Iam-on dan Garret 2010). Pengukuran validasi yang dapat digunakan yaitu *Compactness* (CP). CP merupakan salah satu kategori pengukuran yang paling umum digunakan. CP mengukur jarak rata-rata antara setiap pasang titik data yang termasuk dalam cluster yang sama. Berikut rumus yang digunakan untuk menghitung CP untuk data numerik

$$CP = \frac{1}{N}\sum_{k=1}^{K} n_k \left(\frac{\sum_{x_i,x_j \in C_k} d(x_i,x_j)}{n_k(n_k-1)/2}\right), \quad (2.7)$$

dengan K adalah banyaknya cluster yang terbentuk, $n_k$ adalah jumlah objek pengamatan yang termasuk ke dalam cluster ke-K. d(xi; xj) adalah jarak antara objek ke-i dan objek ke-j, dan N adalah jumlah seluruh objek. Semakin kecil nilai CP, maka cluster yang dihasilkan semakin baik. Pengukuran validasi untuk data kategorik juga dapat menggunakan *compactness* dengan mengukur kemiripan rata-rata antara setiap

pasang titik data yang termasuk dalam cluster yang sama. Rumus yang digunakan sebagai berikut :

$$CP* = \frac{1}{N}\sum_{k=1}^{K} n_k \left(\frac{\sum_{x_i,x_j \in c_k} sim(x_i,x_j)}{n_k(n_k-1)/2}\right), \quad (2.8)$$

dengan *sim(xi; xj)* adalah kemiripan antara objek ke-i dan objek ke-j. Semakin besar nilai CP* maka cluster yang dihasilkan semakin baik.

## 2.6 *One-Way* MANOVA

MANOVA adalah sebuah teknik pengembangan dari metode ANOVA (univariate Analysis of Variance) yang melibatkan beberapa variabel sekaligus (Multivariate). Pengujian hipotesis uji MANOVA adalah sebagai berikut :

$H_0 : \mu 1 = \mu 2 = \cdots = \mu K = \mu$ (tidak ada perbedaan yang signifikan antar rata-rata setiap kelompok)

$H_1 : \mu K \neq \mu$ (terdapat perbedaan yang signifikan pada rata-rata setiap kelompok)

Salah satu alat uji stastistik yang dapat membantu dalam proses pengambilan keputusan pada rata-rata perbedaan antar kelompok adalah Wilk's lambda. Jika nilai stastistik dari uji Wilk's lambda semakin rendah, maka perbedaan rata-rata yang terjadi antar kelompok akan semakin signifikan. Nilai Wilk's lambda berkisar diantara nilai 0 hingga 1.

Pengujian signifikansi menggunakan MANOVA harus memenuhi syarat dari asumsi ditribusi normal multivariate.

## 2.7 Gambaran umum Provinsi Jawa Tengah

Menurut publikasi BPS Jawa Tengah, secara astronomis Provinsi Jawa Tengah terletak diantara 5°40' dan 8°30' Lintang Selatan dan antara 108°30' dan 111°30' Bujur Timur (termasuk Pulau Karimunjawa). Berdasarkan posisi geografis, Jawa Tengah memiliki batas-batas : Utara – Laut Jawa; Selatan – Provinsi Daerah Istimewa Yogyakarta dan Samudra Hindia; Barat – Provinsi Jawa Barat; Timur – Provinsi Jawa Timur. Jawa Tengah memiliki 35 Kabupaten/Kota terdiri dari 29 Kabupaten dan 6 Kota. Berdasarkan proyeksi jumlah penduduk Indonesia 2015-2045 penduduk di Pulau Jawa pada 2019 mencapai 150,4 juta jiwa. Jumlah tersebut setara dengan separuh penduduk Indonesia yang mencapai 266,91 juta jiwa. Adapun jumlah penduduk laki-laki lebih banyak dari perempuan, yakni masing-masing 75,23 juta jiwa dan 75,17 juta jiwa.

## 2.8 Kesejahteraan Rakyat

Kesejahteraan merupakan titik ukur bagi suatu masyarakat bahwa telah berada pada kondisi sejahtera. Kesejahteraan tersebut dapat diukur dari kesehatan, keadaan ekonomi, kebahagiaan dan kualitas hidup rakyat (Segel dan Bruzy, 1998:8). Menurut Badan Pusat Statistik Jawa Tengah aspek spesifik yang dapat dijadikan indikator untuk mengamati kesejahteraan rakyat adalah :

1. Kemiskinan

   Di setiap negara termasuk Indonesia kemiskinan merupakan masalah utama dalam hal pembangunan, angka kemiskinan merupakan salah satu indikator terpenting dalam tolak ukur kesejahteraan rakyat. Indikator kemiskinan meliputi jumlah penduduk (dalam persen).

2. Ketenagakerjaan

   Ketenagakerjaan merupakan salah satu aspek penting bagi pembangunan. Dengan terus bertambahnya angkatan kerja maka dari itu pemerintah harus berupaya dalam hal perluasan lapangan pekerjaan guna mengurangi pengangguran dan kemiskinan. Indikator yang digunakan adalah tingkat pengangguran terbuka (TPT) dan Tingkat Partisipasi Angkatan Kerja (TPAK)

3. Kesehatan

   Tingkat kesehatan merupakan indikator penting untuk menggambarkan mutu pembangunan manusia suatu wilayah. Indikator yang digunakan meliputi angka kematian bayi (AKB) dan angka kematian ibu (AKI)

4. Pendidikan

   Berdasarkan UUD 1945 Pasal 28C, ayat (1) dinyatakan bahwa setiap orang berhak mengembangkan diri melalui pemenuhan kebutuhan dasarnya, berhak mendapatkan pendidikan, memperoleh manfaat dari IPTEK, seni dan budaya demi meningkatkan kualitas hidup dan demi kesejahteraan umat manusia. Indikator yang dapat digunakan adalah angka partisipasi kasar dan angka partisipasi murni pada jenjang SMP/Sederajat dan SMA/Sederajat.

5. Indeks Pembangunan Manusia

   Pencapaian pembangunan manusia diukur dengan memperhatikan tiga aspek esensial yaitu umur panjang dan hidup sehat, pengetahuan, dan standar hidup layak. IPM merupakan indikator komposit yang digunakan untuk melihat perkembangan pembangunan dalam jangka panjang salah satunya adalah kesejahteraan rakyat.