

BAB II

LANDASAN TEORI

2.1 Analisis Multivariat

Zikmund (2000) mengatakan bahwa berdasarkan banyaknya variabel yang terlibat, analisis dapat dibagi menjadi tiga tingkatan. Jika hanya ada satu variabel yang terlibat maka analisis tersebut disebut sebagai analisis univariat. Sedangkan jika ada dua variabel yang terlibat maka analisis tersebut dinamakan analisis bivariat dan jika variabel yang terlibat berjumlah lebih dari dua, maka dinamakan analisis multivariat. Analisis multivariat secara sederhana dapat diartikan sebagai analisis yang dilakukan untuk melihat hubungan diantara lebih dari dua variabel. Hair, dkk (1998) mengartikan analisis multivariat sebagai metode-metode statistik yang mengolah beberapa pengukuran menyangkut individu atau objek sekaligus (*simultaneously*). Analisis multivariat merupakan jenis analisis statistik yang dapat digunakan untuk menganalisis data yang memiliki banyak variabel. Hubungan diantara variabel-variabel yang ada dapat diketahui dengan melakukan analisis multivariat.

Wijaya dan Budiman (2016) membagi teknik analisis multivariat menjadi dua, yaitu analisis dependensi dan interdependensi. Analisis dependensi berfungsi menjelaskan atau memprediksi variabel terikat (*dependent variable*) dengan melibatkan dua atau lebih variabel bebas. Adanya keberadaan variabel terikat dan variabel bebas merupakan ciri penting dari dependensi. Analisis regresi berganda, analisis multivariat varian, analisis diskriminan, analisis korelasi kanonikal, dan analisis *conjoint* merupakan bagian dari analisis dependensi. Sedangkan analisis interdependensi berfungsi melakukan pengelompokan terhadap sejumlah data atau memberikan makna terhadap seperangkat variabel. Berbeda dengan analisis dependensi, dalam analisis interdependensi variabel yang ada tidak dipisahkan menjadi variabel terikat dan variabel bebas. Analisis *cluster*, analisis faktor, dan *multidimensional scaling* termasuk dalam analisis ini.

2.2 Matriks Data Multivariat

Data yang dikumpulkan dari dua atau lebih observasi dimana observasi tersebut diukur dengan beberapa karakteristik merupakan pengertian dari data multivariat. Data multivariat muncul ketika peneliti berusaha untuk memahami berbagai fenomena dimana pengamatannya dilakukan terhadap lebih dari satu variabel. Data dalam analisis multivariat dapat dinyatakan dalam bentuk matriks \mathbf{X} dimana terdapat n objek dan p variabel (Johnson & Wichern, 2002), sebagai berikut:

$$\mathbf{X} = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix} = (x_{il}), i = 1, 2, \dots, n \text{ dan } l = 1, 2, \dots, p \quad (2.1)$$

dimana:

- x_{il} : nilai objek ke- i pada variabel ke- l (elemen dari matriks \mathbf{X})
- n : banyak objek
- p : banyaknya variabel yang diamati

2.3 Analisis Cluster

Teknik multivariat yang mempunyai tujuan utama untuk mengelompokkan objek-objek berdasarkan karakteristik yang dimilikinya merupakan pengertian dari analisis *cluster* (Wijaya & Budiman, 2016). Sekumpulan objek yang satu sama lain saling memiliki kesamaan dan ditempatkan dalam sebuah kelompok yang sama dinamakan sebagai *cluster*. *Cluster-cluster* terbentuk dengan homogenitas internal dan heterogenitas eksternal yang tinggi. Hal ini dikarenakan analisis *cluster* menjadikan objek yang saling berdekatan sebagai satu *cluster*. Objek yang saling berdekatan dapat ditunjukkan dengan dekatnya jarak antar objek dan kedekatan keseluruhan objek dalam suatu *cluster* dapat dilihat dari seberapa dekat jarak objek terhadap *centroid cluster*. Analisis *cluster* berfokus pada perbandingan objek berdasarkan set variabel. Set variabel dapat merepresentasikan karakteristik dari objek-objek yang ada sehingga seperangkat variabel yang digunakan akan menentukan solusi *cluster* secara keseluruhan.

Dikelompokkannya objek-objek berdasarkan kemiripan karakteristik tertentu bertujuan untuk meringkas data dan mengatur penerimaan informasi sesuai dengan

kepentingan. Dua metode yang umum digunakan untuk melakukan pengelompokan objek-objek ke dalam suatu *cluster* dinamakan metode hierarki dan non-hierarki. Aglomerasi (*agglomerative*) dan pemecahan (*divisive*) merupakan klasifikasi metode dalam analisis *cluster* hierarki. Metode aglomerasi sendiri terdiri atas metode *linkage*, *variance*, dan *centroid*, dimana *linkage* terdiri dari metode *Single Linkage*, *Complete Linkage*, dan *Average Linkage*, sedangkan *variance* terdiri dari metode *Ward*. Sementara itu, metode pemecahan (*divisive*) dilakukan dengan menggunakan metode *Splinter Average Distance*. Metode *Sequential Threshold*, *Parallel Threshold*, dan Optimalisasi merupakan bagian dari analisis *cluster* non-hierarki. Penentuan mengenai metode yang akan digunakan tergantung kepada keinginan peneliti dan konteks penelitian tanpa mengabaikan teori, substansi, dan konsep yang berlaku.

2.3.1 Metode Pengelompokan Hierarki (*Hierarchical Clustering Method*)

Teknik hierarki (*hierarchical methods*) adalah teknik *clustering* membentuk konstruksi hierarki atau berdasarkan tingkatan tertentu seperti struktur pohon (struktur pertandingan) (Gunawan, 2016). Apabila informasi mengenai jumlah kelompok belum diketahui maka metode pengelompokan hierarki dapat digunakan. Pengelompokan dengan metode hierarki dimulai dengan menyatukan dua atau lebih objek yang mempunyai kemiripan paling dekat. Kemudian proses dilanjutkan pada objek yang lain hingga *cluster* membentuk semacam pohon dengan tingkatan (hierarki) yang tampak jelas antar objek, dari yang paling dekat kemiripannya hingga yang paling jauh. Dendogram merupakan alat yang dapat membantu memperjelas proses hierarki.

Tipe dasar dalam metode ini dapat dijelaskan sebagai berikut:

1. Metode Aglomerasi (*Agglomerative Methods*)

Metode ini memulai pengelompokan dengan anggapan bahwa masing-masing objek berada di dalam *cluster* yang berbeda (membentuk *cluster* tersendiri), sehingga banyaknya *cluster* adalah sebanyak jumlah objek. Kemudian sepasang objek yang memiliki jarak terdekat bergabung. Setelah itu objek selanjutnya akan bergabung dengan *cluster* yang sudah terbentuk atau bergabung bersama objek lain membentuk *cluster* baru. Jarak kedekatan antar

objek tetap diperhitungkan. Proses akan diteruskan hingga akhirnya terbentuk sebuah *cluster* yang beranggotakan keseluruhan objek. Dalam proses pembentukan *cluster*, terdapat lima macam metode aglomerasi, yaitu:

a) *Single Linkage*

Jarak terdekat menjadi prinsip dalam metode ini. Proses pengelompokan diawali dengan menemukan jarak terkecil dalam $\mathbf{D} = \{d_{ij}\}$, kemudian objek yang saling berdekatan tersebut digabung, misal A dan B sehingga kedua objek membentuk *cluster* (AB). Selanjutnya jarak antara *cluster* (AB) dengan *cluster* lain misalnya C dihitung menggunakan Persamaan (2.2) (Gunawan, 2016).

$$d_{(AB)C} = \min(d_{AC}, d_{BC}) \quad (2.2)$$

dimana:

d_{AC} : jarak antara *cluster* A dan C

d_{BC} : jarak antara *cluster* B dan C

b) *Complete Linkage*

Jarak terjauh menjadi prinsip dalam metode ini. Algoritma *Complete Linkage* diawali dengan menemukan jarak terkecil dalam $\mathbf{D} = \{d_{ij}\}$, kemudian objek yang saling berdekatan tersebut digabung, misal A dan B sehingga kedua objek membentuk *cluster* (AB). Selanjutnya jarak antara *cluster* (AB) dengan *cluster* lain misalnya C dihitung menggunakan Persamaan (2.3) (Gunawan, 2016).

$$d_{(AB)C} = \max(d_{AC}, d_{BC}) \quad (2.3)$$

c) *Average Linkage*

Jarak rata-rata antar tiap pasangan objek yang mungkin merupakan prinsip dalam metode ini. Proses diawali dengan menemukan jarak terkecil dalam $\mathbf{D} = \{d_{ij}\}$, kemudian objek yang saling berdekatan tersebut digabung, misal A dan B sehingga kedua objek membentuk *cluster* (AB). Selanjutnya jarak antara *cluster* (AB) dengan *cluster* lain misalnya C dihitung menggunakan Persamaan (2.4) (Dani, Wahyuningsih, & Rizki, 2019).

$$d_{(AB)C} = \frac{d_{(AC)} + d_{(BC)}}{n_{(AB)}n_C} \quad (2.4)$$

dimana:

n_{AB} : banyaknya objek dalam *cluster* (AB)

n_C : banyaknya objek dalam *cluster* C

d) Metode *Ward*

Ward merupakan metode analisis *cluster* yang berusaha meminimumkan variasi antar objek yang berada dalam satu *cluster*. Pengelompokan objek menggunakan metode *Ward* diawali dengan menemukan jarak terkecil dalam $\mathbf{D} = \{d_{ij}\}$, kemudian objek yang saling berdekatan tersebut digabung, misal A dan B sehingga kedua objek membentuk *cluster* (AB). Selanjutnya jarak antara *cluster* (AB) dengan *cluster* lain misalnya C dihitung menggunakan Persamaan (3.1) (Dani, Wahyuningsih, & Rizki, 2019).

e) Metode *Centroid*

Metode *Centroid* merupakan teknik dalam analisis *cluster* yang proses pengklasterannya didasarkan pada jarak antar *centroid*-nya. *Centroid* adalah rata-rata dari semua anggota dalam satu *cluster* (Arifin, Ermawati, & Hartinah, 2018). Rata-rata dari seluruh objek dalam *cluster* dilibatkan saat penggabungan antar *cluster* sehingga metode *Centroid* dapat memperkecil *variance within cluster*. Algoritma *Centroid* diawali dengan menemukan objek yang paling mirip dalam $\mathbf{D} = \{d_{ij}\}$, kemudian objek yang saling berdekatan tersebut digabung, misal A dan B sehingga kedua objek membentuk *cluster* (AB). Penambahan anggota membuat *centroid* berubah sehingga *centroid* baru dihitung menggunakan Persamaan (2.5) (Dani, Wahyuningsih, & Rizki, 2019).

$$d_{(AB)C} = \frac{n_A d_{(AC)} + n_B d_{(BC)}}{n_{(AB)}} - \frac{n_A n_B d_{(AB)}}{n_{(AB)}^2} \quad (2.5)$$

dimana:

- n_A : banyaknya objek pada *cluster A*
 n_B : banyaknya objek pada *cluster B*
 $n_{(AB)}$: banyaknya objek pada *cluster (AB)*

2. Metode Pemecahan (*Divisive Methods*)

Pengklasteran dengan *divisive methods* memiliki proses yang berlawanan dengan *agglomerative methods*. Pada awalnya seluruh objek berada dalam satu *cluster*. Selanjutnya objek dengan ketidakmiripan yang besar akan dipisahkan dan membentuk *cluster* baru. Pemisahan terus dilanjutkan hingga terbentuk sejumlah *cluster* yang diinginkan. Metode *Splinter Average Distance* merupakan teknik dalam melakukan *divisive methods*. Metode *Splinter Average Distance* ini didasarkan pada perhitungan jarak rata-rata masing-masing objek dengan objek pada *group splinter* (kelompok yang memisahkan diri dari induk) dan jarak rata-rata objek tersebut dengan objek lain pada kelompoknya (Gunawan, 2016). Apabila jarak objek ke *group splinter* lebih dekat daripada jarak objek ke kelompoknya sendiri, maka objek tersebut dikeluarkan dari kelompoknya dan bergabung ke *group splinter*. Jika suatu objek selalu lebih dekat dengan kelompoknya sendiri maka proses berhenti. Selanjutnya dilakukan tahap pemisahan dalam *group*.

2.3.2 Metode Pengelompokan Non-Hierarki (*Non-Hierarchical Clustering Method*)

Pengelompokan non-hierarki bertujuan untuk mengelompokkan n objek ke dalam k *cluster* ($k < n$). Prosedur non-hierarki diawali dengan menentukan jumlah *cluster* yang diinginkan. Objek kemudian akan digabungkan ke dalam sejumlah *cluster* tersebut. *K-Means clustering* merupakan penamaan yang biasa digunakan untuk menyebut metode pengelompokan non-hierarki. Teknik atau metode pengelompokan non-hierarki terdiri dari *Sequential Threshold*, *Parallel Threshold*, dan Optimalisasi (Gunawan, 2016).

1. *Sequential Threshold*

Pengelompokan dengan metode ini diawali dengan memilih satu objek dasar sebagai nilai awal *cluster*. Kemudian seluruh objek yang memiliki jarak

terdekat dengan *cluster* ini akan bergabung ke dalam satu *cluster*. Proses berlanjut hingga terbentuk sejumlah *cluster* yang dikehendaki.

2. *Parallel Threshold*

Metode ini memiliki prinsip yang sama dengan *Sequential Threshold*. Akan tetapi *Parallel Threshold* memilih beberapa objek sekaligus sebagai nilai awal *cluster* dan selanjutnya penggabungan objek ke dalamnya dilakukan secara bersamaan dengan jarak tetap dijadikan patokan dalam penggabungan objek ke dalam *cluster-cluster*.

3. Optimalisasi

Sequential Threshold dan *Parallel Threshold* yang dikembangkan menciptakan metode baru yaitu optimalisasi. Metode ini melakukan optimasi pada penempatan objek yang ditukar untuk *cluster* lainnya dengan pertimbangan kriteria optimasi.

2.4 Tahapan Analisis Cluster

Tahap-tahap yang dilakukan dalam mengelompokkan objek menggunakan analisis *cluster* adalah sebagai berikut (Supranto, 2010):

1. Merumuskan Masalah

Perumusan masalah dilakukan dengan mendefinisikan variabel-variabel yang digunakan sebagai dasar pengklasteran. Proses pemilihan variabel-variabel yang akan digunakan dalam membentuk *cluster* merupakan hal yang paling penting dalam perumusan masalah. Masuknya variabel yang tidak sesuai dengan permasalahan yang diangkat akan mendistorsi hasil pengklasteran sehingga set variabel yang dipilih haruslah relevan dengan permasalahan yang menjadi topik penelitian.

2. Standardisasi Data

Proses standardisasi dilakukan apabila terdapat variabilitas satuan diantara variabel-variabel yang diteliti. Perbedaan satuan yang cukup besar dapat berakibat pada tidak validnya perhitungan yang dilakukan dalam analisis *cluster*. Standardisasi dilakukan pada data asli dengan cara mentransformasinya sebelum

dianalisis lebih lanjut. Transformasi dilakukan terhadap variabel yang relevan ke dalam bentuk *z-score* (Gunawan, 2016).

$$Z = \frac{x - \bar{x}}{s} \quad (2.6)$$

dimana:

x : data

\bar{x} : rata-rata data

s : standar deviasi data

3. Memilih Ukuran Jarak atau Similaritas

Jarak (*distance*) antar pasangan objek merupakan pendekatan yang umum digunakan untuk mengukur kemiripan. Objek dengan jarak yang saling berdekatan akan memiliki karakteristik yang lebih mirip dibandingkan dengan yang saling berjauhan. *Euclidean* merupakan ukuran jarak yang umum digunakan. Jarak *euclidean* merupakan besarnya jarak suatu garis lurus yang menghubungkan antar objek (Gunawan, 2016). Besarnya jarak antara objek ke- i dan objek ke- j disimbolkan dengan d_{ij} dan untuk variabel ke- l dengan $l = 1, 2, \dots, p$. Jarak *euclidean* dihitung menggunakan Persamaan (2.7) (Simamora, 2005).

$$d_{ij} = \sqrt{\sum_{l=1}^p (x_{il} - x_{jl})^2} \quad (2.7)$$

dimana:

d_{ij} : jarak *euclidean* antara objek ke- i dan objek ke- j

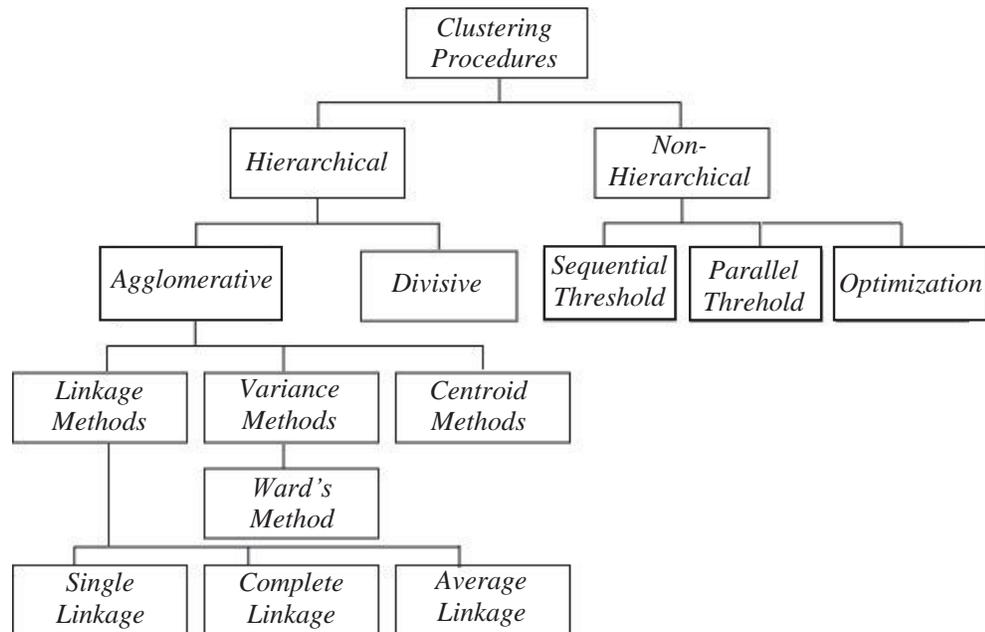
p : banyaknya variabel yang diamati

x_{il} : nilai objek ke- i pada variabel ke- l

x_{jl} : nilai objek ke- j pada variabel ke- l

4. Memilih Suatu Prosedur Pengklasteran

Metode pengelompokan dalam analisis *cluster* dibedakan menjadi dua, yaitu metode hierarki dan non-hierarki. Klasifikasi prosedur pengklasteran dalam analisis *cluster* dapat dilihat pada Gambar 2.1.



Sumber: Supranto, 2010

Gambar 2.1 Klasifikasi Prosedur Pengklasteran

5. Menentukan Banyaknya *Cluster*

Permasalahan utama dalam mengelompokkan objek dengan analisis *cluster* adalah menentukan banyaknya *cluster* yang terbentuk. Hal ini dikarenakan tidak adanya aturan baku yang menjelaskan berapa banyak *cluster* yang baik untuk digunakan. Penentuan mengenai banyaknya *cluster* penting dilakukan agar objek dapat terklasterkan dengan tepat.

6. Menginterpretasikan Profil *Cluster*

Pada tahap ini dilakukan interpretasi terhadap karakteristik dari masing-masing *cluster* untuk menunjukkan bahwa *cluster-cluster* yang terbentuk memiliki perbedaan pada dimensi yang relevan. Tiap *cluster* diberi penamaan sesuai dengan sifat khas yang dapat menggambarkan *cluster* tersebut. Rata-rata masing-masing variabel di tiap *cluster* dijadikan ukuran untuk menjabarkan karakteristik *cluster*.