BAB III

ANALISIS DISKRIMINAN

3.1 Analisis Diskriminan

Analisis diskriminan (*discriminant analysis*) merupakan salah satu metode yang digunakan dalam analisis multivariat. Dalam analisis diskriminan terdapat dua jenis variabel yang terlibat yaitu variabel bebas dan variabel terikat. Variabel bebas dalam analisis diskriminan berupa data metrik (interval dan rasio) sedangkan variabel terikatnya berupa data nonmetrik (nominal dan ordinal). Oleh karena itu, analisis diskriminan termasuk ke dalam analisis multivariat metode dependensi (Sharma, 1996).

Analisis diskriminan adalah teknik multivariat untuk memisahkan objek-objek dalam kelompok yang berbeda dan mengelompokkan objek baru ke dalam kelompok-kelompok tersebut (Johnson, 1956). Analisis diskriminan dapat digunakan jika variabel terikatnya terdiri dari dua kelompok atau lebih. Apabila variabel terikatnya lebih dari dua kelompok, maka metode yang digunakan adalah analisis diskriminan multipel (*multiple discriminant analysis*).

Ada dua tujuan utama dalam pemisahan kelompok dalam analisis diskriminan, yaitu (Rencher, 2002) :

1. Aspek deskriptif atau menggambarkan pemisahan kelompok, di mana fungsi linier variabel (fungsi diskriminan) digunakan untuk menggambarkan atau menjelaskan perbedaan-perbedaan antara dua atau lebih kelompok. Tujuan dari gambaran analisis diskriminan meliputi identifikasi kontribusi *p* variabel

untuk memisahkan kelompok dan mencari hasil yang optimal di mana poin-poin tersebut dapat menjelaskan gambaran terbaik setiap kelompok.

2. Aspek prediksi atau mengelompokkan observasi ke dalam kelompok, di mana fungsi linier atau kuadratik variabel (fungsi pengelompokan) digunakan untuk menentukan unit sampel individu ke dalam salah satu dari beberapa kelompok. Nilai-nilai yang diukur dalam vektor observasi dari individu atau objek akan dievaluasi oleh fungsi pengelompokan untuk mencari kelompok di mana individu dipastikan termasuk di dalamnya.

Ada beberapa kasus analisis diskriminan, di antaranya:

1. Analisis Diskriminan Linier (*Linear Discriminant Analysis*).

   Analisis diskriminan linier digunakan jika data *p* buah variabel bebas berdistribusi normal multivariat dan setiap kelompoknya memiliki matriks varians kovarians yang sama.

2. Analisis Diskriminan Kuadratik (*Quadratic Discriminant Analysis*).

   Analisis diskriminan kuadratik digunakan jika data *p* buah variabel bebas berdistribusi normal multivariat tetapi matriks varians kovariansnya tidak sama dalam setiap kelompoknya.

3. Analisis Diskriminan Fisher (*Fisher Discriminant Analysis*).

   Analisis diskriminan Fisher digunakan jika data *p* buah variabel bebas tidak berdistribusi normal multivariat tetapi matriks varians kovariansnya sama dalam setiap kelompoknya.

4. Analisis Diskriminan Nonparametrik (*Nonparametric Discriminant Analysis).*

Analisis diskriminan nonparametrik digunakan jika data *p* buah variabel bebas tidak berdistribusi normal multivariat dan matriks varians kovariansnya tidak sama dalam setiap kelompoknya.

Analisis diskriminan melibatkan kombinasi linier dari dua atau lebih variabel bebas untuk membentuk suatu fungsi diskriminan yang dapat digunakan untuk membedakan suatu kelompok dengan kelompok lainnya. Kombinasi linier untuk analisis diskriminan memiliki bentuk persamaan linier, yaitu:

$$L = b_0 + b_1 X_1 + b_2 X_2 + ... + b_p X_p \qquad (3.1)$$

di mana, L = *score* diskriminan, b = bobot (*weight*) dan X = variabel bebas.

Dalam tujuan utama analisis diskriminan, yaitu aspek deskriptif, fungsi diskriminan yang terbentuk digunakan untuk membedakan suatu kelompok dengan kelompok lainnya dalam suatu populasi. Selain untuk membedakan kelompok, fungsi diskriminan juga dapat digunakan untuk masalah pengelompokan yaitu dalam aspek prediksi, fungsi yang terbentuk adalah fungsi pengelompokan yang digunakan untuk mengelompokkan observasi ke dalam kelompok yang telah ada. Fungsi pengelompokan ini disebut juga fungsi diskriminan, namun fungsi diskriminan ini tidak sama dengan fungsi diskriminan pada aspek deskriptif.

Pada proses pengelompokan analisis diskriminan, setiap observasi sebelumnya sudah diketahui masuk ke dalam salah satu kelompok dari beberapa kelompok yang ada. Dengan demikian, akan muncul konsep kesalahan

pengelompokan. Dari konsep inilah dapat diketahui seberapa baiknya pengelompokan yang dilakukan oleh analisis diskriminan tersebut.

Proses pengelompokan dalam analisis diskriminan dilakukan dengan cara membentuk suatu fungsi pengelompokan masing-masing kelompok, selanjutnya dihitung suatu skor setiap observasi dari masing-masing fungsi pengelompokan tersebut yang disebut dengan skor diskriminan.

Pengelompokan menggunakan skor diskriminan dilakukan dengan membuat suatu aturan pengelompokan untuk mengetahui observasi masuk ke dalam kelompok yang ada. Berikut akan dibahas aturan pengelompokan dalam analisis diskriminan.

3.2 Aturan Pengelompokan

Misalkan sebuah populasi Ω terdiri dari g kelompok $\pi_1, \pi_2, \ldots, \pi_g$ dengan masing-masing wilayah (*region*) $R_1, R_2, \ldots, R_g$. Suatu pengukuran terdiri dari p variabel bebas, $\mathbf{X}' = \left[\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p\right]$ dilakukan pada setiap kelompok sebanyak n observasi atau objek, $\mathbf{x}_m = \left\{x_{m1}, x_{m2}, \ldots, x_{mn_i}\right\}$; $m = 1, 2, \ldots, p$. Perbedaan antar kelompok dapat dilihat dari fungsi kepadatannya, $f_i(\mathbf{x})$ jika observasi berasal dari kelompok i, π_i ; $i = 1, 2, \ldots, g$ dengan peluang prior p_i di mana $\sum_{i=1}^{g} p_i = 1$. Besarnya biaya/resiko salah pengelompokan ada bila observasi yang berasal dari kelompok i (π_i) dikelompokkan sebagai kelompok k (π_k) dinotasikan dengan $c(k|i)$ dengan peluang $P(k|i)$; $i, k = 1, 2, \ldots, g$.

Berikut akan dibahas beberapa metode untuk memperoleh aturan pengelompokan observasi atau objek ke dalam salah satu kelompok dari beberapa kelompok yang ada pada analisis diskriminan.

3.2.1 Metode ECM Minimum

Nilai harapan dari salah pengelompokan (*Expected Cost of Misclassification* = ECM) dibangun oleh tiga komponen, yaitu peluang prior p_i, biaya/resiko salah mengelompokkan $c(k|i)$ dan peluang salah mengelompokkan $P(k|i)$.

Biaya/resiko salah pengelompokan akan bernilai sama dengan nol atau $c(k|i)=0$ jika $k=i$. Misalkan R_k adalah himpunan semua x yang dikelompokkan sebagai π_k, maka peluang salah pengelompokan ($P(k|i)$) adalah peluang bersyarat mengelompokkan observasi sebagai π_k padahal observasi tersebut berasal dari π_i, yaitu:

$$P(k|i)=\int_{R_k} f_i(\mathbf{x})\, d\mathbf{x},$$

untuk $k\neq i$; $i,k=1,2,\ldots,g$, peluang bersyarat $P(i|i)=1-\sum_{k=1}^{g} P(k|i)$.

Biaya/resiko salah pengelompokan dapat didefinisikan sebagai matriks biaya. Misalkan suatu populasi terdiri dari dua kelompok π_1 dan π_2, maka matriks biayanya adalah

		Diklasifikasikan sebagai	
		π_1	π_2
Populasi yang benar	π_1	0	$c(2\|1)$
	π_2	$c(1\|2)$	0

Untuk populasi yang terdiri dari g kelompok $\pi_1, \pi_2, \ldots, \pi_g$, maka ECM bersyarat dari x yang berasal dari π_1 yang dikelompokkan ke dalam π_2, atau π_3, …, atau π_g adalah

$$\text{ECM}(1) = P(2|1)c(2|1) + P(3|1)c(3|1) + \cdots + P(g|1)c(g|1)$$

$$= \left(\sum_{k=2}^{g} P(k|1)c(k|1) \right).$$

Dengan mengalikan setiap ECM bersyarat (ECM(i) ; $i = 1, 2, \ldots, g$) dengan masing-masing peluang priornya (p_i ; $i = 1, 2, \ldots, g$), maka diperoleh total biaya/resiko salah pengelompokan (*Total Cost of Misclassification* = TCM), yaitu:

$$\text{TCM} = p_1\text{ECM}(1) + p_2\text{ECM}(2) + \ldots + p_g\text{ECM}(g)$$

$$\text{TCM} = p_1 \left(\sum_{k=2}^{g} P(k|1)c(k|1) \right) + p_2 \left(\sum_{\substack{k=1 \\ k\neq 2}}^{g} P(k|2)c(k|2) \right) + \ldots + p_g \left(\sum_{k=1}^{g-1} P(k|g)c(k|g) \right)$$

$$\text{TCM} = \sum_{i=1}^{g} p_i \left(\sum_{\substack{k=1 \\ k\neq i}}^{g} P(k|i)c(k|i) \right) \qquad (3.2)$$

Pilih $R_1, R_2, \dots, R_g$ agar TCM bernilai minimum, sehingga diperoleh aturan pengelompokan yang optimal sebagai berikut:

***Result* 1.**

Aturan pengelompokan dengan metode ECM minimum adalah kelompokkan *x* ke dalam π_k, $k = 1, 2, \dots, g$ di mana

$$\sum_{\substack{i=1 \\ i \neq k}}^{g} p_i f_i(\mathbf{x}) c(k|i) \text{ bernilai minimum.} \tag{3.3}$$

Jika terdapat satu atau lebih, dipilih salah satu di antaranya.

Bukti.

Dalam metode ECM, peluang prior populasi diketahui. Oleh karena itu, dapat didefinisikan peluang posterior berdasarkan teori Bayesian (lampiran 7). Peluang posterior dari observasi yang berasal dari π_i, $P(\pi_i|\mathbf{x})$ adalah

$$P(\pi_i|\mathbf{x}) = \frac{p_i f_i(\mathbf{x})}{\sum_{l=1}^{g} p_l f_l(\mathbf{x})}.$$

Jika observasi tersebut dikelompokkan sebagai π_j, maka kerugian harapannya adalah

$$\sum_{\substack{i=1 \\ i \neq j}}^{g} \frac{p_i f_i(\mathbf{x})}{\sum_{l=1}^{g} p_l f_l(\mathbf{x})} c(j|i).$$

Untuk meminimumkan kerugian harapan tersebut, pilih j agar nilai kerugian harapan minimum. Kita mengganggap $\sum_{\substack{i=1 \\ i \neq j}}^{g} p_i f_i(\mathbf{x}) c(j|i)$ untuk semua j

dan memilih j yang memiliki biaya/resiko salah pengelompokannya ($c(j|i)$) minimum sehingga akan menyebabkan nilai $\sum_{\substack{i=1 \\ i\neq j}}^{g} p_i f_i(\mathbf{x}) c(j|i)$ menjadi minimum.

Oleh karena itu, kelompokkan *x* ke dalam $\pi_k,\ k = 1, 2, \ldots, g$ di mana

$$\sum_{\substack{i=1 \\ i\neq k}}^{g} p_i f_i(\mathbf{x}) c(k|i) < \sum_{\substack{i=1 \\ i\neq j}}^{g} p_i f_i(\mathbf{x}) c(j|i),$$

atau dengan kata lain kelompokkan *x* ke dalam $\pi_k,\ k = 1, 2, \ldots, g$ di mana

$\sum_{\substack{i=1 \\ i\neq k}}^{g} p_i f_i(\mathbf{x}) c(k|i)$ bernilai minimum. □

Bila komponen biaya/resiko salah pengelompokan $c(k|i)$ diabaikan atau diasumsikan sama untuk setiap kelompok, maka dari persamaan TCM akan dihasilkan aturan total peluang salah pengelompokan (*Total Probability of Misclassification* = TPM).

3.2.2 Metode TPM Minimum

Kriteria lain dari ECM yaitu bila biaya/resiko salah pengelompokan $c(k|i)$ diabaikan atau diasumsikan sama untuk setiap kelompoknya dapat digunakan untuk memperoleh aturan pengelompokan optimal, yaitu dengan memilih $R_1, R_2, \ldots, R_g$ untuk meminimumkan total peluang salah pengelompokan (*Total Probability of Misclassification* = TPM).

Misalkan suatu populasi terdiri dari dua kelompok π_1 dan π_2, maka TPMnya adalah

TPM = $P(\text{Salah pengelompokan observasi } \pi_1 \text{ atau salah pengelompokan observasi } \pi_2)$

TPM = $P(\text{observasi berasal dari } \pi_1 \text{ dan salah pengelompokan})$

$+P(\text{observasi berasal dari } \pi_2 \text{ dan salah pengelompokan})$

TPM = $p_1 \int_{R_2} f_1(\mathbf{x})\, d\mathbf{x} + p_2 \int_{R_1} f_2(\mathbf{x})\, d\mathbf{x}$.

Untuk populasi yang terdiri dari g kelompok $\pi_1, \pi_2, \ldots, \pi_g$, maka TPMnya adalah

$$\text{TPM} = \sum_{i=1}^{g} p_i \left(\int_{\substack{R_k \\ k \neq i}} f_i(\mathbf{x})\, d\mathbf{x} \right) \tag{3.4}$$

dengan

$$\int_{\substack{R_k \\ k \neq i}} f_i(\mathbf{x})\, d\mathbf{x} = P(k|i).$$

Kita juga dapat mengelompokkan observasi ke dalam kelompok yang memiliki peluang posterior maksimum. Menurut aturan Bayes (lampiran 7), peluang posterior dari observasi yang berasal dari π_i, $P(\pi_i|\mathbf{x})$ adalah

$$P(\pi_i|\mathbf{x}) = \frac{p_i f_i(\mathbf{x})}{\sum_{l=1}^{g} p_l f_l(\mathbf{x})}. \tag{3.5}$$

Aturan TPM adalah aturan ECM bila biaya/resiko salah pengelompokan $c(k|i)$ diabaikan atau diasumsikan sama untuk setiap kelompoknya, maka aturan pengelompokan yang optimal dengan metode yang meminimumkan TPM adalah:

***Result* 2.**

Aturan pengelompokan dengan metode TPM minimum adalah kelompokkan *x* ke dalam π_k jika

$$p_k f_k(\mathbf{x}) > p_i f_i(\mathbf{x}) \quad \text{untuk semua } i \neq k, \tag{3.6}$$

atau, setara dengan,

kelompokkan *x* ke dalam π_k jika

$$\ln p_k f_k(\mathbf{x}) > \ln p_i f_i(\mathbf{x}) \quad \text{untuk semua } i \neq k. \tag{3.7}$$

Bukti.

Andaikan semua biaya/resiko salah pengelompokan adalah sama atau diabaikan, maka persamaan pada result 1 (kasus aturan ECM minimum) menjadi

$$\sum_{\substack{i=1 \\ i \neq k}}^{g} p_i f_i(\mathbf{x})$$

yang merupakan aturan TPM minimum. TPM yaitu jumlah semua peluang pengelompokan yang bersifat salah pengelompokan.

Oleh karena itu, kelompokkan *x* ke dalam $\pi_k,\ k = 1, 2, \ldots, g$ di mana

$$\sum_{\substack{i=1 \\ i \neq k}}^{g} p_i f_i(\mathbf{x}) \text{ bernilai minimum.}$$

Nilai tersebut akan bernilai minimum jika $p_k f_k(\mathbf{x})$ bernilai maksimum, ini menyebabkan peluang posteriornya menjadi maksimum. Hal ini merupakan salah satu kriteria dalam mendapatkan aturan pengelompokan yang optimal.

Jadi kelompokkan *x* ke dalam π_k jika $p_k f_k(\mathbf{x})$ bernilai maksimum. Atau dengan kata lain kelompokkan *x* ke dalam π_k jika

$$p_k f_k(\mathbf{x}) > p_i f_i(\mathbf{x}) \quad \text{untuk semua } i \neq k,$$

atau, setara dengan,

kelompokkan *x* ke dalam π_k jika

$$\ln p_k f_k(\mathbf{x}) > \ln p_i f_i(\mathbf{x}) \quad \text{untuk semua } i \neq k. \qquad \square$$

3.2.3 Pengelompokan dengan Populasi Normal Multivariat

Pada kasus di mana $f_i(\mathbf{x}), i = 1,2,\ldots,g$ memiliki fungsi kepadatan normal multivariat dengan vektor rata-rata $\boldsymbol{\mu}_i$ dan matriks varians kovarians $\boldsymbol{\Sigma}_i$ dengan bentuk:

$$f_i(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}_i|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)' \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)\right], i = 1,2,\ldots,g. \quad (3.8)$$

Jika semua biaya/resiko salah pengelompokan adalah sama ($c(i|i) = 0, c(k|i) = 1$; $k \neq i$), maka aturan pengelompokan yang optimal yang meminimumkan ECM (sama dengan aturan TPM minimum) menjadi

kelompokkan *x* ke dalam π_k jika

$$\begin{aligned} \ln p_k f_k(\mathbf{x}) &= \ln p_k - \left(\tfrac{p}{2}\right)\ln(2\pi) - \tfrac{1}{2}\ln|\boldsymbol{\Sigma}_k| - \tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)' \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k) \\ &= \ln p_k - \tfrac{1}{2}\ln|\boldsymbol{\Sigma}_k| - \tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)' \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k) \quad (3.9) \\ &= \max_i \ln p_i f_i(\mathbf{x}) \end{aligned}$$

Catatan :

$-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)'\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k) = d_i^L(\mathbf{x}) = -\frac{1}{2}\mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{x} + \boldsymbol{\mu}_i'\boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_i'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i$.

Bukti.

$$\ln p_k f_k(\mathbf{x}) = \ln p_k \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}_k|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)'\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)\right]$$

$$\ln p_k f_k(\mathbf{x}) = \ln p_k + \ln \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}_k|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)'\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)\right]$$

$$\ln p_k f_k(\mathbf{x}) = \ln p_k + \ln \frac{1}{(2\pi)^{p/2}} + \ln \frac{1}{|\boldsymbol{\Sigma}_k|^{1/2}} + \left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)'\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)\right]$$

$$\ln p_k f_k(\mathbf{x}) = \ln p_k + \ln 1 - \ln(2\pi)^{p/2} + \ln 1 - \ln|\boldsymbol{\Sigma}_k|^{1/2} - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)'\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)$$

$$\ln p_k f_k(\mathbf{x}) = \ln p_k - \frac{p}{2}\ln(2\pi) - \frac{1}{2}\ln|\boldsymbol{\Sigma}_k| - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)'\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)$$

Konstanta $(p/2)\ln(2\pi)$ dapat diabaikan karena bernilai sama untuk semua kelompok. Maka persamaan di atas menjadi

$$\ln p_k f_k(\mathbf{x}) = \ln p_k - \tfrac{1}{2}\ln|\boldsymbol{\Sigma}_k| - \tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)'\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k).$$

Dari result 2 diperoleh $\ln p_k f_k(\mathbf{x}) > \ln p_i f_i(\mathbf{x})$, maka $\ln p_k f_k(\mathbf{x})$ merupakan nilai supremum dari $\ln p_i f_i(\mathbf{x})$, sehingga

$$\ln p_k f_k(\mathbf{x}) = \max_i \ln p_i f_i(\mathbf{x}).$$

□

Analisis diskriminan yang memenuhi asumsi distribusi normal multivariat terdiri dari dua macam, yaitu analisis diskriminan linier dan analisis diskriminan kuadratik.

3.2.4 Analisis Diskriminan Linier

Analisis diskriminan linier (*Linear Discriminant Analysis* = LDA) digunakan apabila observasi **X** memenuhi asumsi distribusi normal multivariat dan homogenitas matriks varians kovarians.

Berdasarkan persamaan (3.9) dapat didefinisikan skor diskriminan linier. Karena matriks varians kovarians sama untuk setiap kelompoknya maka substitusikan $\Sigma_i = \Sigma$, untuk $i = 1, 2, \ldots, g$. Oleh karena itu, untuk populasi ke-*i* skor diskriminan linier didefinisikan sebagai:

$$d_i^L(\mathbf{x}) = -\tfrac{1}{2}\ln|\boldsymbol{\Sigma}| - \tfrac{1}{2}\mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{x} + \boldsymbol{\mu}_i'\boldsymbol{\Sigma}^{-1}\mathbf{x} - \tfrac{1}{2}\boldsymbol{\mu}_i'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i + \ln p_i . \quad (3.10)$$

Dua suku pertama akan bernilai sama untuk $d_1^L(\mathbf{x}), d_2^L(\mathbf{x}), \ldots, d_g^L(\mathbf{x})$, maka suku tersebut dapat diabaikan. Skor diskriminan linier menjadi

$$d_i^L(\mathbf{x}) = \boldsymbol{\mu}_i'\boldsymbol{\Sigma}^{-1}\mathbf{x} - \tfrac{1}{2}\boldsymbol{\mu}_i'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i + \ln p_i \quad (3.11)$$

Jika $\boldsymbol{\mu}_i$ dan **Σ** tidak diketahui, maka gunakan $\bar{\mathbf{x}}_i$ sebagai taksiran vektor rata-rata $\boldsymbol{\mu}_i$ dan gunakan $\mathbf{S}_{\text{gab}}$ sebagai taksiran matriks varians kovarians gabungan **Σ**, yaitu:

$$\mathbf{S}_{\text{gab}} = \frac{(n_1-1)\mathbf{S}_1 + (n_2-1)\mathbf{S}_2 + \cdots + (n_g-1)\mathbf{S}_g}{n_1 + n_2 + \cdots + n_g - g} . \quad (3.12)$$

Maka taksiran $\hat{d}_i^L(\mathbf{x})$ dapat diperoleh dari skor diskriminan linier $d_i^L(\mathbf{x})$ yang dibentuk berdasarkan taksiran matriks varians kovarians gabungan **Σ** yaitu:

$$\hat{d}_i^L(\mathbf{x}) = \bar{\mathbf{x}}_i'\mathbf{S}_{\text{gab}}^{-1}\mathbf{x} - \tfrac{1}{2}\bar{\mathbf{x}}_i'\mathbf{S}_{\text{gab}}^{-1}\bar{\mathbf{x}}_i + \ln p_i \quad (3.13)$$

dengan

$\bar{\mathbf{x}}_i$ = vektor rata-rata sampel kelompok ke-*i*

$\mathbf{S}_i$ = matriks varians kovarians sampel kelompok ke-*i*

n_i = ukuran sampel kelompok ke-*i*.

Oleh karena itu, taksiran aturan pengelompokannya adalah kelompokkan *x* ke dalam π_k jika

Skor diskriminan linier $\hat{d}_k^L(\mathbf{x}) = \text{maks}\left(\hat{d}_1^L(\mathbf{x}), \hat{d}_2^L(\mathbf{x}), \ldots, \hat{d}_g^L(\mathbf{x})\right)$. (3.14)

3.2.5 Analisis Diskriminan Kuadratik

Analisis diskriminan kuadratik (*Quadratic Discriminant Analysis* = QDA) digunakan apabila observasi $\mathbf{X}$ memenuhi asumsi distribusi normal multivariat tetapi tidak memenuhi homogenitas matriks varians kovarians (Σ_i tidak sama).

Berdasarkan persamaan (3.9) dapat didefinisikan skor diskriminan kuadratik. Karena matriks varians kovarians tidak sama untuk setiap kelompoknya, maka untuk populasi ke-*i* skor diskriminan kuadratik didefinisikan sebagai:

$$d_i^Q(\mathbf{x}) = -\tfrac{1}{2}\ln|\mathbf{\Sigma}_i| - \tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)'\,\mathbf{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) + \ln p_i,\ i = 1, 2, \ldots, g. \quad (3.15)$$

Maka aturan pengelompokannya adalah kelompokkan *x* ke dalam π_k jika

skor diskriminan kuadratik $d_k^Q(\mathbf{x}) = \text{maks}\left(d_1^Q(\mathbf{x}), d_2^Q(\mathbf{x}), \ldots, d_g^Q(\mathbf{x})\right)$. (3.16)

Jika $\boldsymbol{\mu}_i$ dan $\mathbf{\Sigma}_i$ tidak diketahui, maka taksiran $\hat{d}_i^Q(\mathbf{x})$ dari skor diskriminan kuadratik $d_i^Q(\mathbf{x})$ adalah

$$\hat{d}_i^Q(\mathbf{x}) = -\tfrac{1}{2}\ln|\mathbf{S}_i| - \tfrac{1}{2}(\mathbf{x}-\bar{\mathbf{x}}_i)'\,\mathbf{S}_i^{-1}(\mathbf{x}-\bar{\mathbf{x}}_i) + \ln p_i,\ i = 1, 2, \ldots, g, \qquad (3.17)$$

dengan

$\bar{\mathbf{x}}_i$ = vektor rata-rata sampel kelompok ke-*i*

$\mathbf{S}_i$ = matriks varians kovarians sampel kelompok ke-*i*

n_i = ukuran sampel kelompok ke-*i*.

Oleh karena itu, taksiran aturan pengelompokannya adalah kelompokkan *x* ke dalam π_k jika

skor diskriminan kuadratik $\hat{d}_k^Q(\mathbf{x}) = \text{maks}\left(\hat{d}_1^Q(\mathbf{x}), \hat{d}_2^Q(\mathbf{x}), \ldots, \hat{d}_g^Q(\mathbf{x})\right)$. (3.18)

3.2.6 Metode Jarak Kuadrat

Jarak kuadrat diperoleh dari persamaan (3.15) dengan mengabaikan suku konstan, $-\tfrac{1}{2}\ln|\boldsymbol{\Sigma}|$. Jika nilai populasi tidak diketahui, maka bentuk taksiran jarak kuadrat dari $\mathbf{x}$ ke vektor rata-rata sampel ke-*i* $\bar{\mathbf{x}}_i$ adalah

1. Untuk kasus matriks varians kovarians yang sama dalam setiap kelompoknya atau $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \cdots = \boldsymbol{\Sigma}_g = \boldsymbol{\Sigma}$, yaitu:

$$D_i^2(\mathbf{x}) = (\mathbf{x}-\bar{\mathbf{x}}_i)'\,\mathbf{S}_{gab}^{-1}(\mathbf{x}-\bar{\mathbf{x}}_i), \quad i = 1, 2, \ldots, g. \qquad (3.19)$$

2. Untuk kasus matriks varians kovarians yang tidak sama dalam setiap kelompoknya, yaitu:

$$D_i^2(\mathbf{x}) = (\mathbf{x}-\bar{\mathbf{x}}_i)'\,\mathbf{S}_i^{-1}(\mathbf{x}-\bar{\mathbf{x}}_i), \quad i = 1, 2, \ldots, g. \qquad (3.20)$$

Maka aturan pengelompokannya adalah

kelompokkan $\mathbf{x}$ ke dalam π_i jika $-\frac{1}{2}D_i^2(\mathbf{x}) + \ln p_i$ bernilai maksimum. (3.21)

atau,

kelompokkan $\mathbf{x}$ ke dalam π_i jika $D_i^2(\mathbf{x})$ bernilai minimum (3.22)

Jika peluang prior kelompok ke-*i* tidak diketahui, maka aturan pengelompokan biasa menetapkan $p_1 = p_2 = \cdots = p_g = 1/g$ atau suku $\ln p_i$ dapat dihilangkan. (3.23)

3.3 Evaluasi Hasil Pengelompokan

Ada suatu prosedur untuk mengetahui tingkat ketepatan pengelompokan, di antaranya *Actual Error Rate* (AER) dan *Apparent Error Rate* (APER).

Prosedur tersebut berdasarkan dari matriks konfusi. Matriks konfusi menunjukkan keanggotaan kelompok pada kenyataan melawan keanggotaan kelompok yang diprediksi. Untuk n_1 observasi dari π_1 dan n_2 observasi dari π_2, maka matriks konfusinya adalah

		Keanggotaan yang diprediksi		
		π_1	π_2	
Keanggotaan pada kenyataan	π_1	n_{1C}	$n_{1M} = n_1 - n_{1C}$	n_1
	π_2	$n_{2M} = n_2 - n_{2C}$	n_{2C}	n_2

Di mana

n_{1C} = jumlah item π_1 yang dikelompokkan secara benar sebagai item π_1

n_{1M} = jumlah item π_1 yang salah dikelompokkan sebagai item π_2

n_{2C} = jumlah item π_2 yang dikelompokkan secara benar sebagai item π_2

n_{2M} = jumlah item π_2 yang salah dikelompokkan sebagai item π_1

3.3.1. *Apparent Error Rate* (APER)

Error Rate pada *Apparent Error Rate* (APER) merupakan proporsi salah pengelompokan pada data *training sample*. APER dapat dengan mudah dihitung dengan matriks konfusi. Maka evaluasi hasil pengelompokan menggunakan *Apparent Error Rate* (APER) adalah

$$\text{APER} = \frac{\sum_{i=1}^{g} n_{iM}}{\sum_{i=1}^{g} n_i}. \tag{3.24}$$

Di mana

n_{iM} adalah banyaknya observasi *training sample* yang salah pengelompokan pada kelompok ke-*i*.

n_i adalah banyaknya observasi pada kelompok ke-*i*.

Ketepatan pengelompokannya adalah

$$1 - \text{APER} \tag{3.25}$$

3.3.2. *Actual Error Rate* (AER)

Error Rate pada *Actual Error Rate* (AER) merupakan proporsi salah pengelompokan pada data sampel validasi atau *holdout sample*. Prosedur *holdout*

Lachenbruch dapat digunakan untuk mengetahui tingkat ketepatan pengelompokan melalui *Actual Error Rate* (AER), di mana taksiran dari ekspektasi *Actual Error Rate* (AER) adalah:

$$\hat{E}\left(\text{AER}\right) = \frac{\sum_{i=1}^{g} n_{iM}^{(H)}}{\sum_{i=1}^{g} n_i}, \quad i = 1, 2, ..., g \,. \tag{3.26}$$

Di mana

$n_{iM}^{(H)}$ adalah banyaknya observasi *holdout* yang salah pengelompokan pada kelompok ke-*i*.

n_i adalah banyaknya observasi pada kelompok ke-*i*.

Ketepatan pengelompokannya adalah

$$1 - \hat{E}\left(\text{AER}\right). \tag{3.27}$$