BAB III

ANALISIS DISKRIMINAN

3.1 Pengertian Analisis Diskriminan

Analisis diskriminan merupakan salah satu metode yang digunakan dalam analisis multivariat dengan metode dependensi (dimana hubungan antar variabel sudah bisa dibedakan mana variabel terikat dan mana variabel bebas). Analisis diskriminan digunakan pada kasus dimana variabel bebas berupa data metrik (interval atau rasio) dan variabel terikat berupa data nonmetrik (nominal atau ordinal). Analisis diskriminan adalah salah satu metode yang dapat digunakan untuk mengetahui variabel mana yang membedakan suatu kelompok dengan kelompok lain dalam suatu populasi. Analisis diskriminan juga dapat digunakan untuk mengklasifikasikan data berdasarkan perbedaan karakterikstik data tersebut.

Menurut Supranto (2004:78), teknik analisis diskriminan dibedakan menjadi dua, yaitu analisis diskriminan dua kelompok dan analisis diskriminan berganda. Untuk analisis diskriminan dua kelompok, jika variabel terikat (Y) dikelompokan menjadi dua maka diperlukan satu fungsi diskriminan. Untuk analisis diskriminan berganda, jika variabel dependen (Y) dikelompokan menjadi lebih dari dua kelompok maka diperlukan fungsi diskriminan sebanyak *(k-1)* untuk *k* kategori.

Analisis diskriminan bertujuan mengklasifikasikan suatu objek kedalam kelompok yang saling lepas (*mutually exclusive/disjoint*) dan menyeluruh

(*exhaustive*) berdasarkan sejumlah variabel bebas dan mengelompokkan objek baru ke dalam kelompok-kelompok yang saling lepas tersebut. Selain itu, beberapa tujuan dari analisis diskriminan ini, antara lain:

1. Menentukan apakah terdapat perbedaan yang nyata antara beberapa karakteristik yang diteliti dalam membedakan dua atau lebih kelompok.
2. Menentukan variabel bebas mana saja yang memberikan kontribusi penting (berarti) dalam membedakan nilai rata-rata diskriminan dari dua atau lebih kelompok.
3. Mengelompokkan data kedalam dua atau lebih kelompok berdasarkan karakteristik data yang diteliti.

Model analisis diskriminan berkenaan dengan kombinasi linear memiliki bentuk sebagai berikut:

$$Y = b_0 + b_1X_1 + b_2X_2 + \cdots + b_kX_k \tag{3.1}$$

Keterangan:

Y = nilai (skor) diskriminan dan merupakan variabel terikat.

X_k = variabel (atribut) ke-k dan merupakan variabel bebas.

b_k = koefisien diskriminan/bobot dari variabel (atribut) ke-k.

Dalam suatu populasi yang terdiri dari dua kelompok dan sejumlah observasi n_i untuk setiap kelompok ke-i, ditentukan kombinasi linear dari variabel bebas yang memisahkan kedalam dua kelompok. Kombinasi linear yang dapat dibentuk dari dua kelompok ini adalah

$$Y_{1i} = a'X_{1i} = a_1X_{1i1} + a_2X_{1i2} + \cdots + a_pX_{1ip} \quad i = 1,2,\ldots,n_1,$$

$$Y_{2i} = a'X_{2i} = a_1X_{2i1} + a_2X_{2i2} + \cdots + a_pX_{2ip} \quad i = 1,2,\ldots,n_2, \tag{3.2}$$

Dengan menggunakan persamaan $\lambda = \frac{a'Ha}{a'Ea}$ (3.3)

Di mana

$$H = \sum_{i=1}^{2} n_i (\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})' \quad (3.4)$$

$$E = \sum_{i=1}^{2} \sum_{j-i}^{n} (x_{ij} - \bar{x}_i)(x_{ij} - \bar{x}_i)' \quad (3.5)$$

dan **a** adalah vektor koefisien. $\bar{\boldsymbol{x}}_{\boldsymbol{i}}$ adalah vektor rata-rata kelompok ke-*i*. dan $\bar{\boldsymbol{x}}$ adalah vektor rata-rata keseluruhan dan n_1, n_2 adalah ukuran sampel dari kelompok 1 dan 2.

Dari persamaan (3.3) dapat dibentuk persamaan

$$a'Ha = \lambda a'Ea$$

$$a'(Ha - \lambda Ea) = 0 \quad (3.6)$$

$\boldsymbol{a}'$ tidak dibolehkan nol karena (3.3) akan menjadi bentuk $\lambda = 0/0$ sehingga solusi diperoleh dari $(Ha - \lambda Ea) = 0$. bentuk ini dapat dinyatakan dalam

$$(E^{-1}h - \lambda I) = 0 \quad (3.7)$$

3.2 Analisis Diskriminan Metode Fisher

Prinsip utama dari fungsi diskriminan Fisher adalah pemisahan sebuah populasi. Fungsi diskriminan yang terbentuk dapat digunakan untuk pengelompokan suatu observasi berdasarkan kelompok-kelompok tertentu. Metode Fisher ini tidak mengasumsikan data harus berdistribusi normal. tapi dalam perhitungan salah satu syarat yang harus diperhatikan adalah data yang

digunakan harus memiliki matriks kovarians yang sama untuk setiap kelompok populasi yang diberikan.

Misalkan terdapat suatu populasi yang terdiri atas h kelompok yang masing-masing mempunyai rata-rata μ_i, $i = 1, 2, \ldots, h$ dan matriks kovarians $\Sigma_1 = \Sigma_2 = \ldots = \Sigma_h = \Sigma$. Misalkan $\bar{\mu}$ adalah rata-rata keseluruhan atau rata-rata gabungan dari populasi tersebut *(overall mean)*, dan B_0 menyatakan *cross product* di antara kelompok:

$$B_0 = \sum_{i=1}^{h} (\bar{\mu}_i - \bar{\mu})(\bar{\mu}_i - \bar{\mu})' \text{ di mana } \bar{\mu} = \frac{1}{h}\sum_{i=1}^{h} \mu_i \tag{3.8}$$

Selanjutnya, kombinasi linear yang terbentuk untuk setiap kelompok dapat dinyatakan dalam bentuk

$$Y = a'X \tag{3.9}$$

Kombinasi linear ini dari tiap kelompok populasi memiliki nilai harapan sebagai berikut:

$$E(Y) = E(a'X) = a'E(X|\pi_i) = a'\mu_i = \mu_{iY} \text{.untuk kelompok } \pi_i \tag{3.10}$$

dan variansi

$$Var(Y) = Var(a'X) = a'Cov(X)a = a'\sum a \text{ . untuk semua populasi} \tag{3.11}$$

Dari beberapa rata-rata kelompok populasi, maka dapat diperoleh rata-rata keseluruhan untuk kombinasi linear gabungan, yaitu

$$\bar{\mu}_Y = \frac{1}{h}\sum_{i=1}^{h} \mu_{iY} = \frac{1}{h}\sum_{i=1}^{h} a'\mu_i = a'\left(\frac{1}{h}\sum_{i=1}^{h} \mu_i\right) = a'\bar{\mu} \tag{3.12}$$

Dalam populasi yang diteliti dapat dilakukan pengukuran keseragaman antara kelompok dari nilai relatif Y terhadap keseragaman dalam kelompok dari populasi yang diberikan tersebut dan diperoleh dengan cara:

$$\frac{\begin{pmatrix} jumlah\ kuadrat\ jarak\ dari \\ rata-rata\ keseluruhan\ populasi \\ terhadap\ Y \end{pmatrix}}{(variansi\ Y)} = \frac{\sum_{i=1}^{h}(\mu_{iY}-\bar{\mu}_Y)^2}{\sigma_Y^2}$$

$$= \frac{\sum_{i=1}^{h}(a'\mu_i - a'\bar{\mu})^2}{a'\sum a} = \frac{a'\left(\sum_{i=1}^{h}(\mu_i-\bar{\mu})(\mu_i-\bar{\mu})'\right)a}{a'\sum a}$$

atau dapat ditulis

$$\frac{\sum_{i=1}^{h}(\mu_{iY}-\bar{\mu}_Y)^2}{\sigma_Y^2} = \frac{a'B_0a}{a'\sum a} \tag{3.13}$$

Dalam perhitungannya besaran-besaran Σ dan μ_i biasanya tidak diketahui, sehingga untuk memperolehnya ditaksir dari sampel yang berukuran n_i dari kelompok populasi π_i, $i = 1, 2, \ldots, h$. Vektor rata-rata yang diperoleh dari tiap sampel diperoleh melalui persamaan berikut

$$\bar{x}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij} \tag{3.14}$$

Matriks kovarians sampel dinotasikan S_i, $i = 1, 2, \ldots, h$, dan vektor rata-rata keseluruhan sampel dapat diperoleh melalui

$$\bar{x} = \frac{\sum_{i=1}^{h} n_i\bar{x}_i}{\sum_{i=1}^{h} n_i} = \frac{\sum_{i=1}^{h}\sum_{j=1}^{n_i} x_{ij}}{\sum_{i=1}^{h} n_i} \tag{3.15}$$

Dari besaran-besaran penaksir di atas, maka diperoleh B_0 untuk menentukan ukuran sampel yaitu

$$B_0 = \sum_{i=1}^{h}(\bar{x}_i-\bar{x})(\bar{x}_i-\bar{x})' \tag{3.16}$$

Selain itu dapat ditentukan penaksir Σ dari sampel. yaitu

$$W = \sum_{i=1}^{h}(n_i - 1)S_i = \sum_{i=1}^{h}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)(x_{ij} - \bar{x}_i)' \tag{3.17}$$

Pada penjelasan sebelumnya penaksir dari Σ untuk populasi yang memiliki matriks kovarians yang sama adalah S_{pooled}. Selanjutnya dapat dinyatakan bahwa

$$\frac{W}{(n_1 + n_2 + \cdots + n_h - h)} = S_{pooled} \tag{3.18}$$

merupakan penaksir untuk Σ.

Dalam perhitungan untuk mencari vektor koefisien yang memaksimumkan keragaman di antara kelompok dari nilai relatif Y terhadap keragaman dalam kelompok. maka ditentukan vektor koefisien $\hat{a}$ yang memaksimumkan $\frac{\hat{a}'B_0\hat{a}}{\hat{a}'S_{pooled}\hat{a}}$ dan memaksimumkan $\frac{\hat{a}'B_0\hat{a}}{\hat{a}'W\hat{a}}$. Untuk mencari $\hat{a}$ yang memaksimumkan kasus ini dapat dinyatakan dalam bentuk vektor eigen $\hat{e}_i, i = 1,2,\ldots,h$ yang dapat dicari dari bentuk $W^{-1}\hat{B}_0$. Vektor-vektor eigen ini bersesuaian dengan nilai eigen dari bentuk persamaan $W^{-1}\hat{B}_0\hat{e} = \hat{\lambda}\hat{e}$ yang juga dapat dituliskan dalam bentuk

$$S_{pooled}^{-1}\hat{B}_0\hat{e} = \hat{\lambda}(n_1 + n_2 + \cdots + n_h - h)\hat{e} \tag{3.19}$$

3.3 Prosedur Analisis Diskriminan

3.3.1 Uji Normal Multivariat

Pengujian normal multivariat dilakukan dengan mencari nilai jarak kuadrat untuk setiap pengamatan yaitu: $d_i^2 = (X - \bar{X})'S^{-1}(X - \bar{X})$, di

mana X_j adalah pengamatan yang ke-j dan S^{-1} adalah kebalikan matriks ragam-peragam S.

Kemudian d_j^2 diurutkan dari yang paling kecil sampai yang paling besar, selanjutnya dibuat plot d_j^2 dengan nilai Chi-Kuadrat $\chi^2_{\frac{j-1/2}{n}}$ dimana j = urutan 1, 2, ... n dan p = banyak peubah. Bila hasil plot dapat didekati dengan garis lurus, maka dapat disimpulkan bahwa peubah ganda menyebar normal.

Untuk menguji normalitas dapat juga dilakukan dengan bantuan menggunakan program SPSS versi 17.0 dengan perumusan hipotesis sebagai berikut:

H_0 : pernyataan-pernyataan yang diteliti berdistribusi normal

H_1 : pernyataan-pernyataan yang diteliti tidak berdistribusi normal

Kriteria pengujian: H_0 ditolak jika nilai sig.< 0.05 atau sebaliknya.

3.3.2 Uji Kesamaan Matriks Kovarians

Uji kesamaan matriks kovarians dapat dilakukan sebagai berikut:

- Perumusan hipotesis

  H_0 : $\Sigma_1 = \Sigma_2$

  H_1 : $\Sigma_1 \neq \Sigma_2$

- Statistik uji

  Statitik uji yang digunakan untuk perhitungan adalah

$$M = \frac{|S_1|^{v_1/2}|S_2|^{v_2/2} \dots |S_k|^{v_k/2}}{|S_{pl}|^{\sum_l v_{pl}/2}} \qquad (3.20)$$

dengan

$$S_{pl} = \frac{\sum_{i=1}^{k} v_i S_i}{\sum_{i=1}^{k} v_i} \text{ dan } v_i = n_i - 1$$

dan M dihitung melalui pendekatan distribusi $\chi 2$ dan F.

Pendekatan distribusi $\chi 2$ dihitung melalui persamaan

$$u = -2(1 - c_1) \ln M$$

Berdistribusi

$$\chi^2_{\left[\frac{1}{2}(k-1)p(p+1)\right]} \text{ dengan } c_1 = \left[\sum_{i=1}^{k} \frac{1}{v_i} - \frac{1}{\sum_{i=1}^{k} v_i}\right]\left[\frac{2p^2 + 3p - 1}{6(p+1)(k-1)}\right]$$

Pendekatan distribusi F dihitung bergantung pada nilai c_1 dan c_2 dengan

$$c_2 = \frac{(p-1)(p+2)}{6(k-1)}\left[\sum_{i=1}^{k} \frac{1}{v_i^2} - \frac{1}{\left(\sum_{i=1}^{k} v_i\right)^2}\right]$$

Juga dengan menghitung

$$a_1 = \frac{1}{2}(k-1)p(p+1), \qquad a_2 = \frac{a_1 + 2}{|c_2 - {c_1}^2|}$$

$$b_1 = \frac{1 - c_1 - a_1 a_2}{a_1}, \qquad b_2 = \frac{1 - c_1 + 2/a_2}{a_2}$$

Jika $c_2 > c_1^2$ maka digunakan $F = -2b_1 \ln M$ yang didekati oleh F_{a_1, a_2}.

Jika $c_2 < c_1^2$ maka digunakan $F = \frac{2a_2 b_2 \ln M}{a_1(1 + 2b_2 \ln M)}$ yang didekati oleh F_{a_1, a_2}.

❖ Kriteria Pengujian

Tolak H_0 jika sign. < 0.05 . atau terima H_0 jika sign. > 0.05.

3.3.3 Uji Kesamaan Vektor Rata-rata

Uji kesamaan vektor rata-rata dari kelompok-kelompok (*Test of Equality of Group Means*) dapat dilakukan sebagai berikut:

❖ Pengujian Hipotesis

H_0 : $\mu_1 = \mu_2$ (pernyataan-pernyataan yang diteliti tidak memiliki perbedaan antar kelompok)

H_1 $\mu_1 \neq \mu_2$ (pernyataan-pernyataan yang diteliti memiliki perbedaan antar kelompok)

❖ Statistik Uji

Statistik uji yang digunakan dalam pengujian hipotesis tersebut adalah statistik *Wilk's Lambda*, yaitu:

$$\Lambda = \frac{|W|}{|W+B|} \tag{3.21}$$

dengan:

$$W = \sum_{i=1}^{k} \sum_{j=1}^{n_1} (X_{ij} - \bar{x}_i)(X_{ij} - \bar{x}_i)'$$

$$B = \sum_{i=1}^{k} n_i (\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})'$$

X_{ij} = pengamatan ke-j kelompok ke-i

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

n_i = banyak pengamatan pada kelompok ke-i

$\bar{x}$ = vektor rata-rata total

❖ Kriteria Pengujian

Tolak H_0 jika sign. < 0.05 . atau sebaliknya. Diharapkan dari uji ini adalah H_0 ditolak.

3.3.4 Pembentukan Fungsi Diskriminan

Fisher mengelompokkan suatu observasi berdasarkan nilai skor yang dihitung dari suatu fungsi linear $Y = \lambda' X$ dimana λ' menyatakan vektor yang berisi koefisien-koefisien variabel bebas yang membentuk persamaan linear terhadap variabel terikat. $\lambda' = [\lambda_1, \lambda_2, \ldots, \lambda_p]$.

$$X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$$

X_k menyatakan matriks data pada kelompok ke-k

$$X_k = \begin{bmatrix} x_{11k} & x_{12k} & x_{1pk} \\ x_{21k} & x_{22k} & x_{2pk} \\ \vdots & \vdots & \vdots \\ x_{n1k} & x_{n2k} & x_{npk} \end{bmatrix}, i = 1,2,\ldots,n \; ; j = 1,2,\ldots,p \; ; k = 1,2$$

x_{ijk} menyatakan observasi ke-i variabel ke-j pada kelompok ke-k.

Dengan asumsi $X_k \sim N(\mu_k, \Sigma_k)$ maka

$$\mu = \begin{bmatrix} E(X_1) \\ E(X_2) \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \text{ dan } E_k = E(X_k - \mu_k)(X_k - \mu_k)'; \; \Sigma_1 = \Sigma_2 = \Sigma$$

$\mu_k = \begin{bmatrix} \mu_{1k} \\ \vdots \\ \mu_{pk} \end{bmatrix}$: μ_k adalah vektor rata-rata tiap variabel X pada kelompok ke-k

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ & \sigma_{22} & \cdots & \sigma_{2p} \\ & & \ddots & \vdots \\ & & & \sigma_{pp} \end{pmatrix}$$

$$\sigma_{j1j2} = \begin{cases} varians\ vaariabel\ j\ apabila\ j_1 = j_2 \\ kovarians\ variabel\ j_1\ dan\ j_2\ apabila\ j_1 \neq j_2 \end{cases}$$

Fisher mentransformasikan observasi-observasi x yang multivariat menjadi observasi y yang univariat. Dari persamaan $Y = \lambda' X$ diperoleh

$$\mu_{ky} = E(Y_k) = E(\lambda' X) = \lambda' \mu_k;$$

$$\sigma_Y^2 = var(a' X) = a' \Sigma a$$

μ_{ky} adalah rata-rata Y yang diperoleh dari x yang termasuk dalam kelompok ke-k, sedangkan σ_Y^2 adalah varians Y dan diasumsikan sama untuk kedua kelompok.

Kombinasi linear yang terbaik menurut Fisher adalah yang dapat memaksimumkan rasio antara jarak kuadrat rata-rata Y yang diperoleh dari x dari kelompok 1 dan 2 dengan varians Y, atau dirumuskan sebagai berikut:

$$\frac{(\mu_{1Y} - \mu_{2Y})^2}{\sigma_Y^2} = \frac{\lambda'(\mu_1 - \mu_2)(\mu_1 - \mu_2)'\lambda}{\lambda'\Sigma\lambda} \tag{3.22}$$

Jika $(\mu_1 - \mu_2) = \delta$, maka persamaan di atas menjadi $\frac{(\lambda'\delta)^2}{\lambda'\Sigma\lambda}$. Karena Σ adalah matriks definit positif, maka menurut teori pertidaksamaan *Cauchy-Schwartz*, rasio $\frac{(\lambda'\delta)^2}{\lambda'\Sigma\lambda}$ dapat dimaksimumkan jika

$$\lambda' = c\Sigma^{-1}\delta = c\Sigma^{-1}(\mu_1 - \mu_2) \tag{3.23}$$

Dengan memilih $c = 1$, menghasilkan kombinasi linear yang disebut kombinasi linear Fisher sebagai berikut:

$$Y = \lambda'X = (\mu_1 - \mu_2)\Sigma^{-1}X \tag{3.24}$$

Setelah dibentuk fungsi linearnya, maka dapat dihitung skor diskriminan untuk setiap observasi dengan mensubstitusikan nilai-nilai variabel bebasnya.

Selanjutnya dilakukan pengujian signifikan dari fungsi diskriminan yang terbentuk, dengan perumusan hipotesis sebagai berikut:

H_0 : pernyataan-pernyataan yang diteliti tidak memiliki perbedaan antar kelompok

H_1 : pernyataan-pernyataan yang diteliti memiliki perbedaan antar kelompok

Kriteria pengujian: H_0 ditolak jika nilai $\chi_{hitung} > \chi_{tabel}$ atau sebaliknya.

Kemudian dilakukan uji kekuatan hubungan fungsi diskriminan untuk melihat seberapa besar hubungan nilai diskriminan dengan kelompok.

3.3.5 Penilaian Validitas Diskriminan

Bobot diskriminan diperkirakan dengan menggunakan *analysis sample* dikalikan dengan nilai variabel bebas di dalam *holdout sample* untuk mendapatkan skor diskriminan untuk kasus yang berada dalam *holdout sample*. Objek atau kasus tersebut kemudian dimasukan kedalam kelompok berdasarkan pada nilai fungsi diskriminan dan aturan-aturan yang tepat.

Secara teoritis terdapat dua prosedur yang dapat digunakan untuk mengevaluasi hasil pengelompokan, yaitu *Actual Error Rate* (AER) dan *Apparent Error Rate* (APER). Prosedur ini 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

Tabel 3.1 Klasifikasi *Actual Group* (Kelompok pada Kenyataan) dan *Predicted Group* (Kelompok yang Diprediksi)

		Predicted Group (Kelompok yang diprediksi)							
		π_1				π_2			
Actual Group	π_1	n_{11}			n_{1M}	n_1	n_{1c}		n_1
(Kelompok pada kenyataan)	π_2	n_{2M}	n_2	n_{2c}		n_{2c}			n_2

Dimana

n_{1C} = banyak pengamatan π_1 yang dikelompokan secara benar sebagai π_1

n_{1M} = banyak pengamatan π_1 yang salah dikelompokan sebagai π_2

n_{2C} = banyak pengamatan π_2 yang dikelompokan secara benar sebagai π_2

n_{2M} = banyak pengamatan π_2 yang salah dikelompokan sebagai π_1

a. *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), dimana taksiran dari ekspektasi *Actual Error Rate* (AER) adalah:

$$\hat{E}\,(AER) = \frac{\sum_{i=1}^{g} n_{iM}^{(H)}}{\sum_{i=1}^{g} n_i}, \quad i = 1, 2, \ldots, g \tag{3.25}$$

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

Ketepatan pengelompokannya adalah $1 - \hat{E}(AER)$

b. *Apparent Error Rate* (APER)

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

$$APER = \frac{\sum_{i=1}^{g} n_{iM}}{\sum_{i=1}^{g} n_i} \tag{3.26}$$

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

Ketepatan pengelompokannya adalah 1 – APER.

Selain secara teoritis, penilaian validitas diskriminan secara praktik (dengan menggunakan SPSS 17.0) data dilakukan dengan menghitung *hit ratio*, yaitu rasio antara observasi yang tepat pengklasifikasiannya dengan total seluruh observasi.