ANALISIS DISKRIMINAN

I. Prinsip Dasar dan Tujuan Analisis

Analisis diskriminan adalah salah satu teknik statistik yang bisa digunakan pada hubungan dependensi (hubungan antarvariabel dimana sudah bisa dibedakan mana variabel respon dan mana variabel penjelas). Lebih spesifik lagi, analisis diskriminan digunakan pada kasus dimana variabel respon berupa data kualitatif dan variabel penjelas berupa data kuantitatif. Analisis diskriminan bertujuan untuk mengklasifikasikan suatu individu atau observasi ke dalam kelompok yang saling bebas (*mutually exclusive/disjoint*) dan menyeluruh (*exhaustive*) berdasarkan sejumlah variabel penjelas.

Ada dua asumsi utama yang harus dipenuhi pada analisis diskriminan ini, yaitu:

1. Sejumlah p variabel penjelas harus berdistribusi normal.
2. Matriks varians-covarians variabel penjelas berukuran pxp pada kedua kelompok harus sama.

Jika dianalogikan dengan regresi linier, maka analisis diskriminan merupakan kebalikannya. Pada regresi linier, variabel respon yang harus mengikuti distribusi normal dan homoskedastis, sedangkan variabel penjelas diasumsikan *fixed,* artinya variabel penjelas tidak disyaratkan mengikuti sebaran tertentu. Untuk analisis diskriminan, variabel penjelasnya seperti sudah disebutkan di atas harus mengikuti distribusi normal dan homoskedastis, sedangkan variabel responnya *fixed*.

II. Format Data Dasar dan Program Komputer yang Digunakan

Data dasar yang digunakan otomatis adalah data yang kontinu (karena adanya asumsi kenormalan) untuk variabel penjelas (X_j) dan data kategorik/kualitatif/*nonmetric* untuk variabel respon (Y).

Tabel 1. Format Data untuk Analisis Diskriminan

X_1	X_2	.	.	.	Xp	Y
…	…	…	…	…	…	…
…	…	…	…	…	…	…

Secara aplikatif, data dilihat pada bagian Contoh Aplikasi Analisis (bagian IV).

Beberapa software yang bisa digunakan adalah SPSS, SAS, dan Minitab. Karena keterbatasan ilmu yang dimiliki penulis, kali ini hanya akan diberikan contoh bagaimana penggunaan SPSS untuk melakukan analisis diskriminan ini.

III. Algoritma Pokok Analisis dan Model Matematis

Secara ringkas, langkah-langkah dalam analisis diskriminan adalah sebagai berikut:

1. Pengecekan adanya kemungkinan hubungan linier antara variabel penjelas. Untuk *point* ini, dilakukan dengan bantuan matriks korelasi (pembentukan matriks korelasi sudah difasilitasi pada analisis diskriminan). Pada output SPSS, matriks korelasi bisa dilihat pada *Pooled Within-Groups Matrices*.
2. Uji Vektor Rata-rata Kedua Kelompok

   H_o: $\boldsymbol{\mu_1 = \mu_2}$

   H_1: $\boldsymbol{\mu_1 \neq \mu_2}$

   Diharapkan dari uji ini adalah hipotesis nol ditolak, sehingga kita mempunyai informasi awal bahwa variabel yang sedang diteliti memang membedakan kedua kelompok. Pada SPSS, uji ini dilakukan secara univariate (jadi yang diuji bukan berupa vektor), dengan bantuan tabel *Tests of Equality of Group Means*.
3. Dilanjutkan pemeriksaan asumsi homoskedastisitas, dengan uji **Box's M**. Diharapkan dari uji ini hipotesisi nol tidak ditolak (Ho**:** $\boldsymbol{\Sigma_1 = \Sigma_2}$).
4. Pembentukan model diskriminan

   a. Kriteria Fungsi Linier Fisher

   - Pembentukan Fungsi Linier (teoritis)

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

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

$\mathbf{X_k}$ menyatakan matriks data pada kelompok ke-k

$$\mathbf{X}_k = \begin{bmatrix} x_{11k} & x_{12k} & \cdot & \cdot & \cdot & x_{1pk} \\ x_{21k} & x_{21k} & \cdot & \cdot & \cdot & x_{2pk} \\ \cdot & \cdot & & & & \cdot \\ \cdot & \cdot & & & & \cdot \\ \cdot & \cdot & & & & \cdot \\ x_{n1k} & x_{n2k} & \cdot & \cdot & \cdot & x_{npk} \end{bmatrix}; \quad \begin{matrix} i = 1, 2, ..., n \\ j = 1, 2, ..., p \\ k = 1, 2 \end{matrix}$$

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

Di bawah asumsi $\mathbf{X_k} \sim \mathbf{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ maka

$$\boldsymbol{\mu} = \begin{bmatrix} \mathbf{E}(\mathbf{X}_1) \\ \mathbf{E}(\mathbf{X}_2) \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix} \text{ dan } \boldsymbol{\Sigma}_k = \mathrm{E}(\mathbf{X}_k - \boldsymbol{\mu}_k)\ (\mathbf{X}_k - \boldsymbol{\mu}_k)'; \ \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$$

$$\boldsymbol{\mu}_k = \begin{bmatrix} \boldsymbol{\mu}_{1k} \\ \cdot \\ \cdot \\ \cdot \\ \boldsymbol{\mu}_{pk} \end{bmatrix}; \ \boldsymbol{\mu}_k \text{ adalah vektor rata-rata tiap variabel X pada kelompok ke-k}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdot & \cdot & \cdot & \sigma_{1p} \\ & \sigma_{22} & \cdot & \cdot & \cdot & \sigma_{2p} \\ & & \cdot & \cdot & \cdot & \cdot \\ & & & \cdot & \cdot & \cdot \\ & * & & & \cdot & \cdot \\ & & & & & \sigma_{pp} \end{bmatrix}$$

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

Fisher mentransformasikan observasi-observasi x yang multivariate menjadi observasi y yang univariate. Dari persamaan $Y = \lambda' \mathbf{X}$ diperoleh

$$\boldsymbol{\mu}_{ky} = \mathrm{E}(Y_k) = \mathrm{E}(\lambda' \mathbf{X}) = \lambda' \boldsymbol{\mu}_k;$$

$$\sigma_Y^2 = \mathrm{var}(\boldsymbol{\ell}' \mathbf{X}) = \boldsymbol{\ell}' \boldsymbol{\Sigma}\, \boldsymbol{\ell}$$

$\boldsymbol{\mu}_{ky}$ adalah rata-rata Y yang diperoleh dari X yang termasuk dalam kelompok ke-k

σ_Y^2 adalah varians Y dan diasumsikan sama untuk kedua kelompok.

Kombinasi linier 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:

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

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

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

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

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

- Pembentukan Fungsi Linier (dengan bantuan SPSS)

  Pada output SPSS, koefisien untuk tiap variabel yang masuk dalam model dapat dilihat pada tabel *Canonical Discriminant Function Coefficient*. Tabel ini akan dihasilkan pada output apabila pilihan *Function Coefficient* bagian *Unstandardized* diaktifkan.

- Menghitung *discriminant score*

  Setelah dibentuk fungsi liniernya, maka dapat dihitung skor diskriminan untuk tiap observasi dengan memasukkan nilai-nilai variabel penjelasnya.

- Menghitung *cutting score*

  Cutting score (m) dapat dihitung dengan rumus sebagai berikut:

$$m = \frac{n_1\mu_{1Y} + n_2\mu_{2Y}}{n_1+n_2}$$

  n_k adalah jumlah sampel ada kelompok ke-k, k=1,2

  Kemudian nilai-nilai *discriminant score* tiap observasi akan dibandingkan dengan *cutting score*, sehingga dapat diklasifikasikan suatu observasi akan termasuk ke dalam kelompok yang mana. Suatu observasi dengan karakteristik **x** akan diklasifikasikan sebagai anggota kelompok kode 1 jika $y = (\mu_1-\mu_2)'\Sigma^{-1}x \geq m$,

selain itu dimasukkan ke dalam kelompok 2(kode nol). Penghitungan m dilakukan secara manual, karena SPSS tidak mengeluarkan output m. Namun, kita dapat menghitung m dengan bantuan tabel *Function at Group Centroids* dari output SPSS.

- Penghitungan *Hit Ratio* (dalam model regresi logistik disebut *percentage correct*)

  Setelah semua observasi diprediksi keanggotaannya, dapat dihitung *hit ratio*, yaitu rasio antara observasi yang tepat pengklasifikasiannya dengan total seluruh observasi.

  Seberapa valid model diskriminan yang telah dihasilkan? Jawaban pertanyaan ini terkait dengan validasi model. SPSS versi 10.0 menggunakan validasi dengan metode *Leave One Out*. Misalkan ada sebanyak n observasi, akan dibentuk fungsi linier dengan observasi sebanyak n-1. Observasi yang tidak disertakan dalam pembentukan fungsi linier ini akan diprediksi keanggotaannya dengan fungsi yang sudah dibentuk tadi. Proses ini akan diulang dengan kombinasi observasi yang berbeda-beda, sehingga fungsi linier yang dibentuk ada sebanyak n. Inilah yang disebut dengan metode *Leave One Out*.

b. Kriteria *posterior probability*

Aturan pengklasifikasian yang ekivalen dengan model linier Fisher adalah berdasarkan nilai peluang suatu observasi dengan karakteristik tertentu (x) berasal dari suatu kelompok. Nilai peluang ini disebut *posterior probability* dan bisa ditampilkan pada sheet SPSS dengan mengaktifkan option probabilities of group membership pada bagian Save di kotak dialog utama.

$$P(k|\mathbf{x}) \frac{p_k f_k(\mathbf{x})}{\sum_k p_k f_k(\mathbf{x})},$$

dimana

p_k adalah *prior probability* kelompok ke-k dan

$$f_k(x) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp -\frac{1}{2}\left(\mathbf{x}-\boldsymbol{\mu}_k\right)'\boldsymbol{\Sigma}^{-1}\left(\mathbf{x}-\boldsymbol{\mu}_\kappa\right); \quad k = 0,1$$

Suatu observasi dengan karakteristik x akan diklasifikasikan sebagai anggota kelompok 0 jika $P(k=0|\mathbf{x}) > P(k=1|\mathbf{x})$. Nilai-nilai posterior probability inilah yang mengisi kolom dis 1_1 dan kolom di 1_2 pada sheet SPSS.

IV. Contoh Aplikasi

Di sebuah laboratorium dilakukan penelitian untuk mengetahui apa saja yang membedakan bunga A dan bunga B yang masih satu species. Untuk itu, diambil sampel bunga A dan B masing-masing sebanyak 10 buah. Kedua bunga dihitung lebar kelopaknya (X_1)dan lebar daunnya (X_2). Diketahui juga bahwa kedua bunga dapat dijadikan indikator derajat keasaman suatu zat (pH), maka diteliti juga pada trayek pH berapa saja kedua bunga sensitif untuk mendeteksinya(X_3). Data yang telah diperoleh akan dianalisis dengan menggunakan analisis diskriminan.

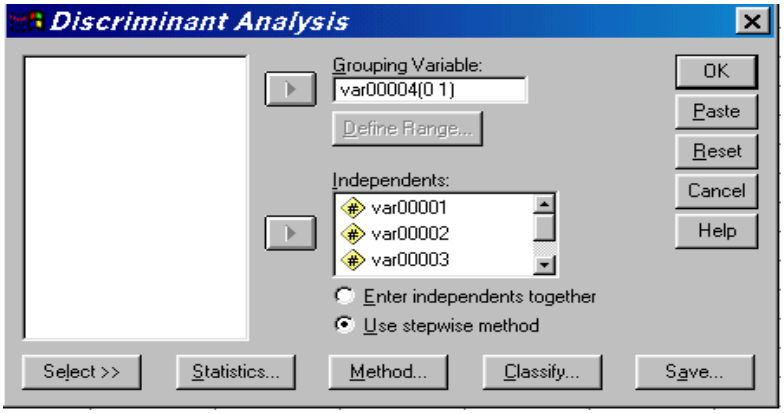
Tabel 3.1 Data karakteristik Bunga A dan Bunga B

X_1	X_2	X_3	Y
4,46209	4,27603	1,43488	0
5,17356	4,03402	1,48285	0
5,27081	3,36186	1,54692	0
4,49723	1,45367	1,27366	0
5,76719	2,13282	1,3265	0
5,91612	3,03981	1,36368	0
5,48373	3,37093	1,32595	0
5,0187	4,92126	1,43008	0
5,43291	3,54893	1,39074	0
4,34865	3,97278	1,38099	0
8,90377	3,51359	1,10593	1
8,37017	4,91499	1,06065	1
8,09676	5,23729	9,2296	1
9,36238	5,69686	1,13544	1
8,62503	5,4649	1,10277	1
9,22858	4,87046	0,10166	1
9,07482	4,90865	9,9952	1
9,84865	5,31779	1,13951	1
8,28943	5,69997	9,8791	1
8,5171	4,99028	8,8796	1

Sumber: Data bangkitan dari Minitab (telah dimodifikasi)

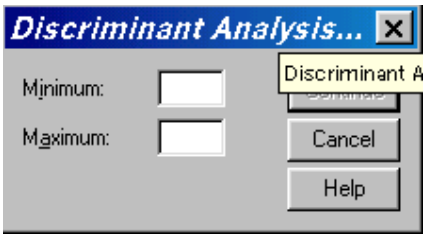
Untuk melakukan analisis diskriminan dengan bantuan SPSS, ikuti langkah-langkah berikut:

1. Pada menu **Analyze**, pilih submenu **Classify**, lalu pilih **Discriminant….**
2. Kemudian akan muncul kotak dialog.



Gambar 4.1 Tampilan Kotak Dialog Utama Analisis Diskriminan

- Bagian **Grouping Variable** diisi dengan variabel respon dan harus didefinisikan range- nya pada bagian **Define Range**.



Gambar 4.2 Tampilan Kotak Dialog Define Range

- Bagian Minimum diisi dengan kode terkecil dan Maximum diisi dengan kode terbesar dari variabel respon.
- Bagian **Independents** diisi dengan variabel penjelas. Metode yang sering dipaparkan pada literatur-literatur adalah metode bertatar (*stepwise*), maka kali ini hanya akan diberi contoh penggunaan metode ini. *Posterior probability* yang dihasilkan dengan metode *Enter* dan *Stepwise* agak berbeda, sehingga pada metode *Stepwise* nilai ketepatan klasifikasinya juga akan berbeda. Berdasarkan literatur-literatur yang pernah dibaca, penulis lebih menyarankan untuk menggunakan metode *Stepwise*. Untuk menampilkan nilai *hit ratio*, pada bagian **Classify** klik **Summary Table**.
- Bagian Save memungkinkan kita untuk menampilkan nilai-nilai *posterior probability* observasi untuk masuk ke kelompok kode nol(dis1_2), nilai-nilai *posterior probability* observasi untuk masuk ke kelompok kode satu (dis2_2), nilai-nilai *discriminant score* (dis1_1), dan pengklasifikasian observasi oleh model (dis_1) pada Sheet SPSS. Misalnya untuk observasi pertama, nilai

peluangnya untuk masuk ke dalam kelompok kode nol (1,00000) lebih besar daripada peluangnya untuk masuk dalam kelompok kode satu (0,00000), maka observasi ini akan dimasukkan oleh model ke dalam kelompok kode nol.

var00004	dis_1	dis1_1	dis1_2	dis2_2
,00	,00	-5,11743	1,00000	,00000
,00	,00	-3,73263	1,00000	,00000
,00	,00	-3,53395	1,00000	,00000

Gambar 2.3 Tampilan *Posterior Probability*, *Discriminant Score*, dan *Predicted Group Membership* pada sheet SPSS

Sampai di sini pengisian kotak dialog dirasa cukup untuk analisis diskriminan. Selanjutnya, kita akan mulai interpretasikan output-outputnya.

- Pengecekan multikolinieritas

Pooled Within-Groups Matrices

		VAR00001	VAR00002	VAR00003
Correlation	VAR00001	1,000	-,131	-,365
	VAR00002	-,131	1,000	,121
	VAR00003	-,365	,121	1,000

Dari matriks korelasi di atas, tidak ada angka yang mencapai 0,5 atau di atasnya sehingga kita mengidentifikasi tidak ada multikolinieritas pada data.

- Uji Kesamaan vektor rata-rata

Tests of Equality of Group Means

	Wilks' Lambda	F	df1	df2	Sig.
VAR00001	,074	225,080	1	18	,000
VAR00002	,487	18,983	1	18	,000
VAR00003	,801	4,467	1	18	,049

Dilihat dari nilai *p-value* nya, masing-masing variabel mempunyai rata-rata yang berbeda untuk kedua kelompok. Ingat, yang diuji adalah kesamaan rata-rata pada tiap kelompok (kelompok kode nol dan kode satu), bukan rata-rata antar variabel.

- Uji Kesamaan matriks *varians-covarians*(homoskedastisitas)

Test Results

Box's M		59,825
F	Approx.	17,558
	df1	3
	df2	58320,000
	Sig.	,000

Tests null hypothesis of equal population covariance matrices.

Tabel di atas memperlihatkan bahwa kita dapat menolak hipotesis nol karena nilai p-valuenya kurang dari 0,05 (dalam hal ini penelitian menggunakan tingkat kepercayaan 95%). Dari hasil pengujian ini, kita dapat mengatakan bahwa data kita berasal dari populasi yang mempunyai matriks varians-covarians yang sama.

- Pembentukan fungsi linier

Canonical Discriminant Function Coefficients

	Function
	1
VAR00001	1,935
VAR00003	,163
(Constant)	-13,988

Unstandardized coefficients

Dari tabel di atas, dapat kita bentuk fungsi liniernya sebagai berikut:

$$Y= -13,988+1,935X_1 +0,163X_3$$

- Penghitungan *discriminant score*

Misalnya untuk observasi pertama, dengan memasukkan nilai X_1=4,46209; dan X_3=14,3488 maka diperoleh discriminant scorenya sebesar -5,117.

- Penghitungan *cutting score*

Functions at Group Centroids

	Function
VAR00004	1
,00	-3,817
1,00	3,817

Unstandardized canonical discriminant functions evaluated at group means

Dari tabel di atas, dapat dihitung *cutting score* nya $= \frac{10(-3{,}817) + 10(3{,}817)}{20} = 0$

Untuk observasi pertama, karena *discriminant score* nya kurang dari *cutting score*, maka dimasukkan ke dalam kelompok kode 0 (pengklasifkasian tepat karena sebenarnya observasi pertama sebelumnya memang termasuk ke dalam anggota kelompok nol atau bunga A).

- *Hit Ratio*

Classification Results[b,c]

			Predicted Group Membership		
		VAR00004	,00	1,00	Total
Original	Count	,00	10	0	10
		1,00	0	10	10
	%	,00	100,0	,0	100,0
		1,00	,0	100,0	100,0
Cross-validated[a]	Count	,00	10	0	10
		1,00	0	10	10
	%	,00	100,0	,0	100,0
		1,00	,0	100,0	100,0

a. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.

b. 100,0% of original grouped cases correctly classified.

c. 100,0% of cross-validated grouped cases correctly classified.

Angka *hit ratio* di atas sudah mencapai 100% (pada kenyataannya sulit mencapai angka sebesar ini, ingat ini hanya data fiktif yang dibangitkan dengan bantuan komputer).

- Pengklasifikasian observasi baru

Jika ada bunga dari *species* yang sama, dapat diprediksi akan termasuk dalam kelompok mana berdasarkan karakteristik yang dimilikinya dengan fungsi linier yang sudah terbentuk. Inilah yang menjadi tujuan pembentukan fungsi diskriminan.

-Tita Rosy-