

KREDENSIAL MIKRO MAHASISWA INDONESIA (KMMI) COURSE DATA SAINS UNTUK BISNIS DAN PERKANTORAN

Modul 4: Analisis Diskriminan

Tujuan Pembelajaran

Setelah mempelajari modul ini, mahasiswa diharapkan:

1. Mampu memahami dan menjelaskan konsep dasar mengenai Analisis Diskriminan,
2. Mampu memahami dan menjelaskan prosedur mengenai Analisis Diskriminan,
3. Mampu menerapkan Analisis Diskriminan dalam pengolahan data bisnis dan perkantoran dengan bantuan software statistik R,
4. Mampu menginterpretasikan hasil pengolahan data yang diperoleh.

A. Pendahuluan

Saat ini kebutuhan akan penggunaan metode statistika multivariat semakin meningkat. Hal ini disebabkan karena pada umumnya permasalahan yang ada lebih kompleks, dimana melibatkan lebih dari satu variabel. Karena itu dibutuhkan bantuan analisis yang lebih komprehensif dari metode statistika univariat. Statistika univariat hanya memperhatikan satu variabel terikat dalam sebuah analisis, sementara statistika multivariat melibatkan lebih dari satu variabel terikat.

Tujuan Analisis Multivariat (Johnson & Wichern, 2002), antara lain, yaitu:

1. untuk mereduksi jumlah variabel
2. untuk mengelompokkan variabel atau atribut.
3. untuk mengidentifikasi hubungan antar variabel.

Metode statistika multivariat terbagi atas dua, yaitu model dependensi dan model interdependensi. Model dependensi berkaitan dengan hubungan antara variabel terikat (dependent) dan variabel bebas (independent), dimana salah satu atau kedua-dua variabel tersebut merupakan multivariabel. Contohnya Analisis Regresi Berganda, Analisis Diskriminan, Analisis Konjoin, Analisis Varians Multivariat, Analisis Korelasi Kanonik, dan sebagainya. Sementara model interdependensi tidak membedakan antara variabel terikat dan variabel bebas, keduanya saling bergantung (interdependensi). Contohnya Analisis Komponen Utama, Analisis Faktor, Analisis Klaster, Penskalaan Multidimensi, dan sebagainya. Pembahasan tentang analisis multivariat tentunya tidak lepas dengan teori distribusi peluang, seperti distribusi normal multivariat, distribusi Chi-Square, distribusi-t (distribusi Student), dan distribusi F (distribusi Snecdor F), dan distribusi Wishart.

B. Materi dan Prosedur

Dalam kursus ini, kita akan mempelajari tentang Analisis Diskriminan.

1. Analisis Diskriminan

Analisis diskriminan merupakan salah satu metode statistika multivariat yang digunakan untuk mengelompokkan sejumlah objek ke dalam beberapa kelompok berdasarkan beberapa variabel, sedemikian sehingga setiap objek menjadi anggota salah satu kelompok. Analisis

KREDENSIAL MIKRO MAHASISWA INDONESIA (KMMI) COURSE DATA SAINS UNTUK BISNIS DAN PERKANTORAN

diskriminan dilakukan dengan pembentukan kombinasi linear dari variabel-variabel tersebut, sehingga skor kombinasi linear tersebut dari setiap objek yang terlibat merupakan petunjuk dari kelompok mana objek tersebut menjadi anggota. Analisis diskriminan juga dapat digunakan untuk mengalokasikan suatu objek baru ke dalam suatu kelompok yang telah terbentuk

Secara umum, tujuan dari analisis diskriminan (Johnson dan Wichern, 2002) adalah

- untuk mendapatkan kombinasi linear saling bebas dari beberapa variabel pembeda (fungsi diskriminan) yang digunakan sebagai dasar pengelompokan suatu objek ke dalam suatu kelompok tertentu,
- untuk memasukkan objek menjadi anggota salah satu kelompok berdasarkan skor yang dicapai pada variabel pembeda,
- untuk mendapatkan variabel pembeda untuk tiap-tiap kelompok yang memberikan sumbangan terbesar terhadap perbedaan antar kelompok.

Analisis diskriminan mampu mengelompokkan objek menjadi dua, tiga, atau lebih kelompok. Jika hanya dua kelompok maka disebut sebagai analisis diskriminan dua kelompok, sedangkan jika mengelompokkan tiga atau lebih kelompok disebut analisis diskriminan ganda.

1.1 Fungsi Diskriminan

Fungsi diskriminan adalah kombinasi linear variabel-variabel bebas yang akan menghasilkan cara terbaik untuk pemisahan kelompok-kelompok. Fungsi ini memberikan nilai sedekat mungkin dalam kelompok dan sejauh mungkin antar kelompok. Fungsi ini digunakan untuk menjelaskan perbedaan antar kelompok dan juga untuk pengklasifikasian.

Untuk kasus dua kelompok, fungsi diskriminan linear Fisher dapat ditulis sebagai berikut:

$$Y = w_1X_1 + w_2X_2 + \dots + w_pX_p \quad (1)$$

di mana:

Y menyatakan skor diskriminan

w_i menyatakan pembobot diskriminan untuk variabel ke- i

X_i menyatakan variabel bebas ke- i .

Persamaan (1) dapat dinyatakan dalam bentuk matrix (Johnson dan Wichern, 2002), sebagai berikut:

$$Y = \tilde{w}^t \tilde{X} = (\tilde{\mu}_1 - \tilde{\mu}_2)^t \Sigma^{-1} \tilde{X} \quad (2)$$

KREDENSIAL MIKRO MAHASISWA INDONESIA (KMMI) COURSE DATA SAINS UNTUK BISNIS DAN PERKANTORAN

di mana $\tilde{w} = \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_p \end{pmatrix}$, $\tilde{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix}$,

$\tilde{\mu}_1$ adalah masing-masing vektor mean dari \tilde{X} di kelompok 1

$\tilde{\mu}_2$ adalah masing-masing vektor mean dari \tilde{X} di kelompok 2

Σ adalah matriks varians kovarians variabel X

Melalui fungsi diskriminan ini, objek dapat dikelompokkan berdasarkan variabel bebasnya. Hasilnya berupa *group mean* atau disebut juga dengan *centroid*. Jika ada dua kelompok, maka akan ada dua centroid, begitu seterusnya. Untuk mengetahui tingkat signifikansi perbedaan antara group yang ada, maka diperlukan informasi jarak di antara centroid yang ada.

Pada analisis diskriminan ini ada dua asumsi yang perlu diperhatikan (Johnson dan Wichern, 2002), yaitu

- 1) p variabel bebas mengikuti distribusi normal multivariat,
- 2) Matriks varians kovarians dari variabel bebas dalam setiap kelompok adalah sama (homogen).

1.2 Pengklasifikasian objek

Misalkan pengamatan baru adalah \tilde{X}_0 dan nilai fungsi diskriminannya (Johnson dan Wichern, 2002) adalah

$$Y_0 = (\tilde{\mu}_1 - \tilde{\mu}_2)^t \Sigma^{-1} \tilde{X}_0,$$

$$m = \frac{1}{2} \tilde{w}^t (\tilde{\mu}_1 + \tilde{\mu}_2), \text{ merupakan titik tengah antara nilai rata-rata } Y$$

Jika $Y_0 \geq m$, maka \tilde{X}_0 merupakan anggota dari kelompok 1, tetapi jika $Y_0 < m$, maka \tilde{X}_0 merupakan anggota dari kelompok 2.

2. Prosedur Analisis Diskriminan

Tahapan yang digunakan pada analisis Diskriminan adalah

- a. Menentukan objek dan variabel yang akan dianalisis
- b. Melakukan pengujian asumsi-asumsi, yaitu
 - 1) tidak terjadi kasus multikolinearitas,
 - 2) variabel yang digunakan mengikuti distribusi normal multivariat,
 - 3) kesamaan matriks varians kovarians dari variabel (homogen).
- c. Menguji kesamaan vektor-vektor mean.
- d. Mengestimasi Fungsi Diskriminan.
- e. Mengecek ketepatan pengklasifikasian dari fungsi diskriminan
- f. Memprediksi kelompok suatu objek baru berdasarkan fungsi diskriminan

KREDENSIAL MIKRO MAHASISWA INDONESIA (KMMI) COURSE DATA SAINS UNTUK BISNIS DAN PERKANTORAN

C. Studi Kasus

Sebuah perusahaan yang bergerak dalam penjualan Air Mineal mengumpulkan data untuk mengetahui faktor-faktor yang menyebabkan perbedaan perilaku konsumen dalam mengkonsumsi Air Mineral, yaitu apakah banyak mengkonsumsi air mineral atau sedikit berdasarkan variabel-variabel/atribut-atribut yang diberikan (Santoso, 2002). Selain informasi tentang perilaku konsumen dalam mengkonsumsi air mineral, atribut-atribut lainnya yang juga ditanyakan dalam daftar kuesioner adalah:

1. Usia Konsumen (tahun)
2. Berat badan konsumen (kg)
3. Tinggi Badan Konsumen (cm)
4. Penghasilan konsumen (ribuan rupiah/bulan)
5. Jam kerja konsumen (jam)

Gunakan informasi data tersebut untuk mengetahui

- a. apakah terdapat perbedaan perilaku konsumen dalam mengkonsumsi air mineral?
- b. faktor-faktor apa saja yang mempengaruhi perbedaan perilaku tersebut?

Data responden tersebut terangkum dalam file data yang dapat didownload di LMS Syam OK. Tujuan dari riset tersebut dapat diperoleh melalui penggunaan analisis diskriminan dengan bantuan software statistic, salah satunya software R. Langkah-langkah yang dilakukan berdasarkan prosedur yang telah dijelaskan pada bagian B.2.

Sebelum kita melakukan analisis diskriminan terlebih dahulu kita memanggil package R di *library* yang diperlukan dalam analisis ini, yaitu:

```
> library(Hmisc)          ### untuk uji Multikolinearitas
> library(mvShapiroTest)  ### untuk uji Normal Multivariat
> library(MVTests)        ### untuk uji Kesamaan Matriks Varians Kovarians
> library(MASS)           ### untuk uji Kesamaan Vektor Mean dan Analisis Diskriminan
> library(rrcov)          ### untuk Analisis Diskriminan
```

1. Menentukan objek dan variabel yang akan dianalisis

Objek dalam contoh ini adalah “minum” dengan 2 kategori, yaitu sedikit (0) dan banyak (1). Sementara variabel bebasnya ada 4, yaitu usia, berat, tinggi, dan jam kerja. Data dan cara mengimpor datanya dalam R dilakukan, sebagai berikut:

```
> dt.disk<-read.csv(file.choose(),header=T,sep=";",dec=".") ## mengimpor data
> dt.disk1<-dt.disk[1:40,]  ### data yang dianalisis dengan ukuran sampel=40
> dt.disk2<-dt.disk[41:50,] ### data untuk prediksi
> dt.disk1
```

usia berat tinggi jamkerja minum

KREDENSIAL MIKRO MAHASISWA INDONESIA (KMMI) COURSE DATA SAINS UNTUK BISNIS DAN PERKANTORAN

1	40	65	154	5.33	0
2	40	70	157	5.30	0
3	40	60	158	5.27	0
4	39	75	160	5.33	0
5	40	50	159	5.50	0
6	28	62	158	5.00	1
7	29	50	160	5.07	0
8	40	52	165	5.13	0
9	35	68	150	5.17	1
10	36	70	152	5.23	0
11	39	50	154	5.33	0
12	30	62	155	5.30	0
13	34	60	157	5.27	0
14	35	51	160	5.33	1
15	40	62	165	5.50	1
16	30	51	162	5.00	1
17	35	80	157	5.33	0
18	22	52	154	5.30	1
19	40	72	155	5.27	0
20	41	45	164	5.33	1
21	32	42	160	5.50	0
22	29	54	157	5.30	0
23	21	35	150	5.27	1
24	25	50	154	5.07	1
25	30	60	158	5.20	0
26	45	40	159	5.13	0
27	35	45	158	5.17	1
28	35	42	152	5.23	0
29	30	51	156	5.30	1
30	24	75	154	5.13	0
31	28	42	155	5.17	1
32	27	51	157	5.23	1
33	20	55	159	5.30	1
34	26	70	160	5.33	0
35	29	40	162	5.40	0
36	20	42	156	5.20	1
37	35	51	153	5.10	1
38	31	70	162	5.40	0
39	34	55	164	5.47	1
40	28	52	160	5.33	1

2. Pengujian asumsi multikolinearitas

```
> uji.multi<-rcorr(as.matrix(dt.disk1[,-5],type="pearson"))
> uji.multi
```

KREDENSIAL MIKRO MAHASISWA INDONESIA (KMMI) COURSE DATA SAINS UNTUK BISNIS DAN PERKANTORAN

	usia	berat	tinggi	jamkerja
usia	1.00	0.19	0.23	0.21
berat	0.19	1.00	-0.05	0.08
tinggi	0.23	-0.05	1.00	0.30
jamkerja	0.21	0.08	0.30	1.00

n= 40

P

	usia	berat	tinggi	jamkerja
usia		0.2330	0.1463	0.1972
berat	0.2330		0.7799	0.6448
tinggi	0.1463	0.7799		0.0567
jamkerja	0.1972	0.6448	0.0567	

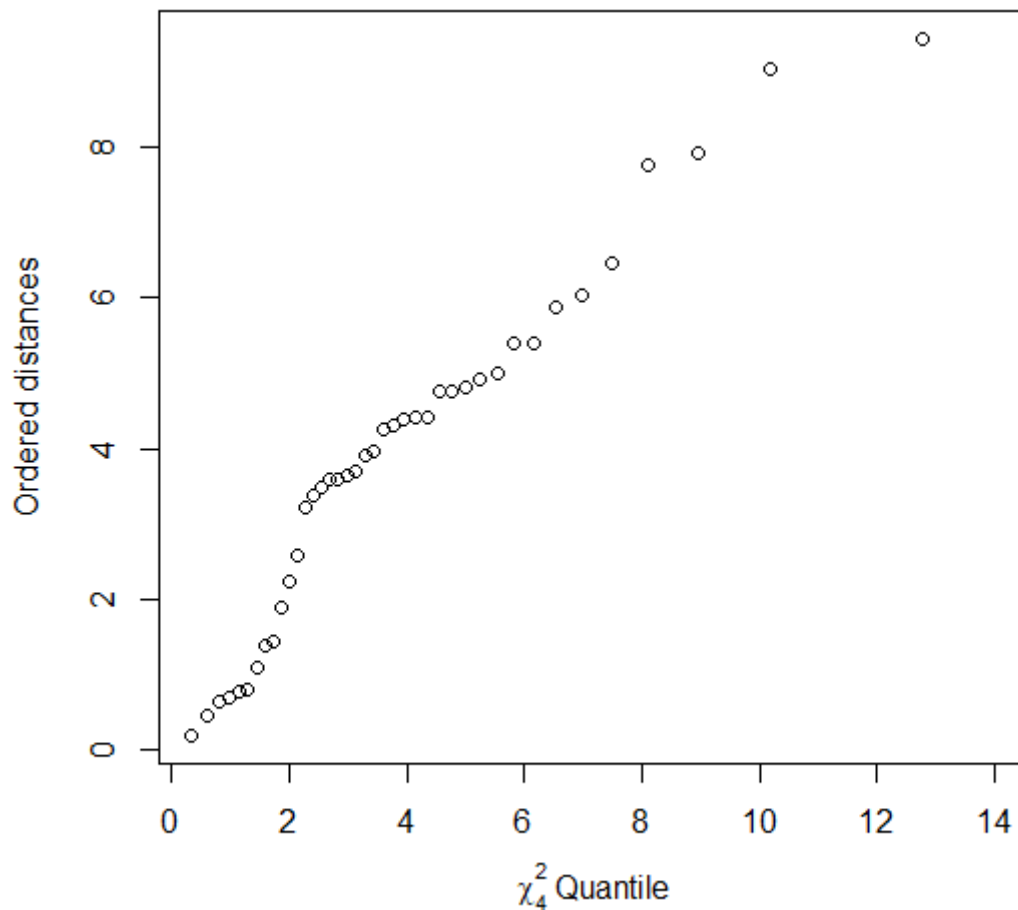
Berdasarkan hasil uji multikolinearitas diperoleh bahwa semua variabel bebas mempunyai nilai $P > 0.05$. Hal ini menunjukkan bahwa tidak terjadi kasus multikolinearitas pada data.

3. Pengujian asumsi distribusi normal multivariate

Berikut diberikan syntax untuk membuat plot Chi-Kuadrat untuk melihat secara visual, apakah data mengikuti distribusi normal multivariat. Jika plot tersebut membentuk pola garis lurus, maka dapat diduga data mengikuti distribusi normal multivariat.

```
> x <- dt.disk1[,-5]
> cm <- colMeans(x)
> S <- cov(x)
> d <- apply(x, 1, function(x) t(x - cm) %*% solve(S) %*% (x - cm))
> plot(qc <- qchisq((1:nrow(x) - 1/2) / nrow(x), df = 4),
+ sd <- sort(d),
+ xlab = expression(paste(chi[4]^2, " Quantile")),
+ ylab = "Ordered distances", xlim = range(qc) * c(1, 1.1))
```

KREDENSIAL MIKRO MAHASISWA INDONESIA (KMMI) COURSE DATA SAINS UNTUK BISNIS DAN PERKANTORAN



Gambar 1. Plot Chi-Kuadrat

Berdasarkan Gambar 1, terlihat bahwa pola data sepertinya membentuk sebuah garis lurus. Untuk lebih meyakinkan bahwa data mengikuti distribusi normal multivariat, dilakukan pengujian, sebagai berikut:

```
> uji.mvn<-mvShapiro.Test(as.matrix(dt.disk1[,-5]))  
> uji.mvn
```

Generalized Shapiro-Wilk test for Multivariate Normality by
Villasenor-Alva and Gonzalez-Estrada

data: as.matrix(dt.disk1[, -5])

MVW = 0.96976, **p-value = 0.3491**

Hasil yang diperoleh adalah nilai- $p > 0.05$. Ini berarti bahwa data mengikuti distribusi normal multivariat.

4. Pengujian asumsi kesamaan matriks varians kovarian

KREDENSIAL MIKRO MAHASISWA INDONESIA (KMMI)

COURSE DATA SAINS UNTUK BISNIS DAN PERKANTORAN

```
> uji.KMV<-BoxM(dt.disk1[,5],dt.disk1[,5])
> uji.KMV
```

```
$Chisq
[1] 10.81696
$df
[1] 10
$p.value
[1] 0.3719557
$Test
[1] "BoxM"
attr(,"class")
[1] "MVTests" "list"
```

Hasil yang diperoleh adalah nilai- $p > 0.05$. Ini berarti bahwa data pada kelompok “minum sedikit” dan data pada kelompok “minum banyak” adalah homogen.

5. Pengujian asumsi kesamaan vektor mean

```
> uji.KVM<-manova(cbind(usia,berat,tinggi,jamkerja)~minum,data=dt.disk1)
> summary(uji.KVM)
```

	Df	Pillai	approx F	num Df	den Df	Pr(>F)
minum	1	0.26299	3.1223	4	35	0.02686 *
Residuals	38					

 Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Karena nilai- $p < 0.05$, maka diperoleh bahwa terdapat perbedaan vektor mean dari data. Ini menunjukkan bahwa pemisahan kelompok “minum sedikit” dan kelompok “minum banyak” signifikan.

6. Mengestimasi fungsi diskriminan

```
> model.disk<-lda(minum~usia+berat+tinggi+jamkerja,data=dt.disk1)
> model.disk
```

Call:

```
lda(minum ~ usia + berat + tinggi + jamkerja, data = dt.disk1)
```

Prior probabilities of groups:

```
0 1
0.55 0.45
```

Group means:

KREDENSIAL MIKRO MAHASISWA INDONESIA (KMMI) COURSE DATA SAINS UNTUK BISNIS DAN PERKANTORAN

	usia	berat	tinggi	jamkerja
0	34.68182	59.50000	157.5909	5.285455
1	29.66667	51.11111	157.5000	5.235556

Coefficients of linear discriminants:

LD1

usia -0.11250838

berat -0.05836712

tinggi 0.05367353

jamkerja -2.43125609

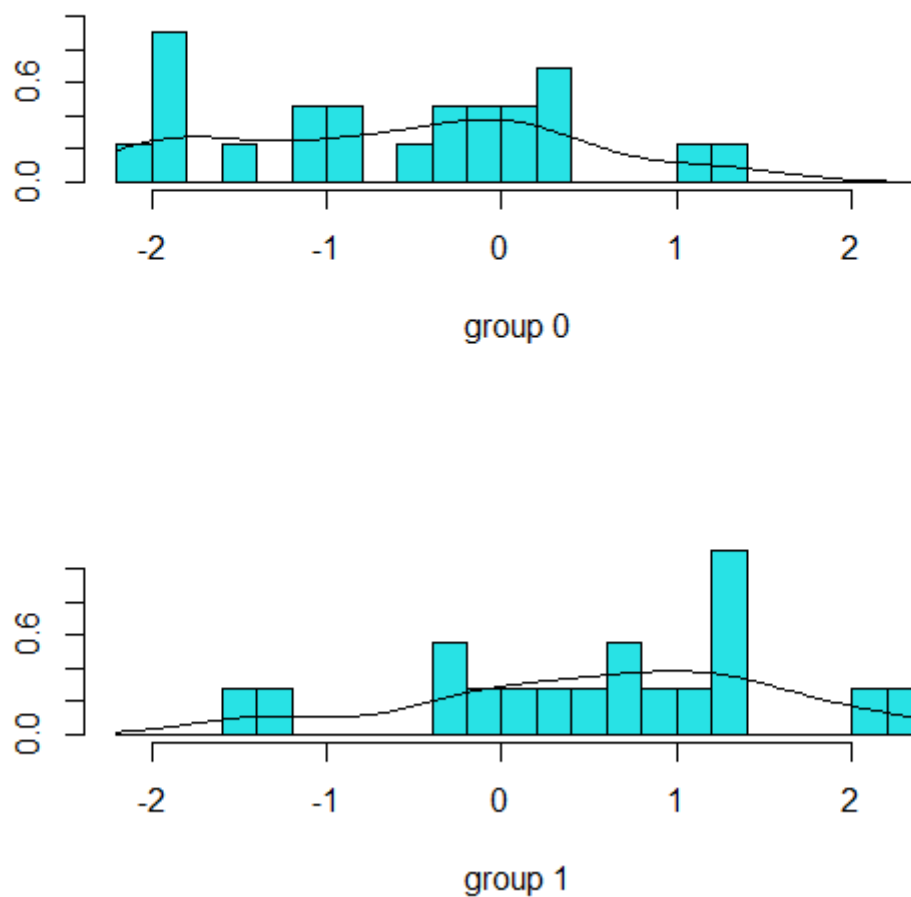
Berdasarkan hasil pemodelan tersebut diperoleh fungsi diskriminan

$$Y = -0.113 \text{ usia} - 0.058 \text{ berat} + 0.054 \text{ tinggi} - 2.431 \text{ jamkerja}$$

Berdasarkan nilai *Groups means* bahwa variabel “tinggi” dan “jamkerja” untuk kelompok “minum sedikit” dan “minum banyak” mempunyai nilai yang hampir sama, sehingga dapat dikatakan bahwa dua variabel tersebut tidak mempengaruhi perilaku seseorang dalam mengonsumsi air mineral. Sementara variabel usia dan berat mempunyai masing-masing nilai mean yang berbeda antara kelompok “minum sedikit” dan “minum banyak”, sehingga dapat dikatakan bahwa kedua variabel tersebut dapat mempengaruhi perilaku seseorang dalam mengonsumsi air mineral.

```
>plot(model.disk)
```

KREDENSIAL MIKRO MAHASISWA INDONESIA (KMMI) COURSE DATA SAINS UNTUK BISNIS DAN PERKANTORAN



Gambar 2. Plot hasil fungsi diskriminan

7. Mengecek hasil pengklasifikasian

```
> lda.klasik<-LdaClassic(dt.disk1[, -5], dt.disk1[, 5])  
> pred.klas<-predict(lda.klasik)  
> pred.klas
```

Apparent error rate 0.275

Classification table

Actual	Predicted	
	0	1
0	17	5
1	6	12

Confusion matrix

KREDENSIAL MIKRO MAHASISWA INDONESIA (KMMI) COURSE DATA SAINS UNTUK BISNIS DAN PERKANTORAN

Actual	Predicted	
	0	1
0	0.773	0.227
1	0.333	0.667

Berdasarkan hasil pengklasifikasian diperoleh tingkat keakuratan $\frac{(0.773+0.667)}{2} = 0.72$ (72%) di mana terdapat 11 pengamatan yang salah klasifikasi.

8. Memprediksi objek baru

```
> dt.baru<-dt.disk2[, -5]
> pred.class<-predict(model.disk,dt.baru)$class
> cbind(dt.baru,pred.class)
```

	usia	berat	tinggi	jamkerja	pred.class
41	29	51	165	5.50	1
42	21	40	162	5.40	1
43	22	70	179	5.23	1
44	22	65	159	5.30	1
45	25	47	154	5.13	1
46	30	40	158	5.27	1
47	45	49	159	5.30	0
48	35	59	156	5.20	0
49	39	70	175	5.13	0
50	34	45	155	5.17	1

Berdasarkan hasil diperoleh bahwa analisis diskriminan dapat digunakan untuk memisahkan kelompok “minum sedikit” dan “minum banyak”, berdasarkan variabel usia, berat badan, tinggi badan dan jam kerja dengan tingkat akurasi 72%. Dua variabel, yaitu usia dan berat badan yang paling berpengaruh terhadap pemisahan tersebut.

Tutorial ringkas tentang analisis diskriminan juga dapat dilihat di link berikut:

<https://www.youtube.com/watch?v=s8pvp2Ctxfc>

D. Latihan Mandiri

Contoh kasus yang digunakan dalam latihan ini diadaptasi dari Santoso dan Tjiptono (2002), dimana pengelola Minimarket X menduga bahwa ada dua kelompok pembeli, yaitu mereka yang sering berbelanja di Minimarket X dan mereka yang dapat dikategorikan jarang berbelanja di minimarket tersebut. Kategori konsumen sering belanja, jika konsumen berbelanja di Minimarket X minimal 7 kali dalam sebulan, selain dari itu dikategorikan jarang berbelanja.

KREDENSIAL MIKRO MAHASISWA INDONESIA (KMMI) COURSE DATA SAINS UNTUK BISNIS DAN PERKANTORAN

Pihak pengelola minimarket ingin mengetahui faktor-faktor apa saja yang mendorong seorang konsumen untuk sering berbelanja atau jarang berbelanja di minimarket tersebut. Untuk mencapai tujuan tersebut, setiap responden diberikan kuesioner yang memuat pertanyaan tentang pendapatnya terhadap 10 atribut minimarket, serta frekuensi berbelanja di minimarket tersebut. Atribut-atribut tersebut adalah

1. Penempatan produk yang dijual (Layout)
2. Kebersihan minimarket (Bersih)
3. Harga produk yang ditawarkan (Harga)
4. Pelayanan Kasir (Kasir)
5. Kelengkapan produk yang dijual (Lengkap)
6. Fasilitas AC yang dingin (AC)
7. Adanya diskon harga produk (Diskon)
8. Ketersedian Lahan Parkir yang aman dan luas (Parkir)
9. Pelayanan Karyawan (Staf)
10. Citra tentang minimarket tersebut (Citra)

Gunakan informasi data tersebut untuk mengetahui

- a. apakah terdapat perbedaan seorang konsumen untuk sering berbelanja atau jarang berbelanja?
- b. faktor-faktor apa saja yang mempengaruhi perbedaan perilaku tersebut?

*Data lengkap untuk latihan mandiri ini dapat diakses di LMS Syam Ok.

E. Rangkuman

Modul ini telah memperkenalkan analisis diskriminan yang dapat digunakan untuk mengelompokkan sejumlah objek berdasarkan variabel-variabel/atribut. Selain itu, analisis diskriminan juga dapat digunakan untuk mengalokasikan suatu objek baru ke dalam satu kelompok. Modul ini juga telah dilengkapi dengan studi kasus yang penyelesaiannya/pengolahan datanya menggunakan software R.

F. Daftar Pustaka

1. Johnson, R.A., dan Wichern, D.W. 2002. *Applied Multivariate Statistical Analysis*, Fifth Edition. Prentice Hall: New Jersey.
2. Santoso, S. 2002. *SPSS Statistik Multivariat*. PT. Elex Media Komputindo: Jakarta.
3. Santoso, S., dan Tjiptono, F. 2002. *Riset Pemasaran: Konsep dan Aplikasi dengan SPSS*. PT. Gramedia: Jakarta.