R Notebook

library(MASS)  
library(MVN)  
library(biotools)

## ---  
## biotools version 4.2

library(readxl)  
library(ggplot2)  
library(tidyverse)

## Warning: package 'tidyverse' was built under R version 4.2.2

## ── Attaching packages  
## ───────────────────────────────────────  
## tidyverse 1.3.2 ──

## ✔ tibble 3.1.8 ✔ dplyr 1.0.9  
## ✔ tidyr 1.2.0 ✔ stringr 1.4.1  
## ✔ readr 2.1.2 ✔ forcats 0.5.2  
## ✔ purrr 0.3.4   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ✖ dplyr::select() masks MASS::select()

library(caret)

## Warning: package 'caret' was built under R version 4.2.2

## Loading required package: lattice  
##   
## Attaching package: 'caret'  
##   
## The following object is masked from 'package:purrr':  
##   
## lift

data13 <- read\_excel("C:/Users/lili/Downloads/data13.xlsx",sheet="Sheet2")  
knitr::kable(head(data13, 10))

| GPA | GMAT | admit |
| --- | --- | --- |
| 2.96 | 596 | 1 |
| 3.14 | 473 | 1 |
| 3.22 | 482 | 1 |
| 3.29 | 527 | 1 |
| 3.69 | 505 | 1 |
| 3.46 | 693 | 1 |
| 3.03 | 626 | 1 |
| 3.19 | 663 | 1 |
| 3.63 | 447 | 1 |
| 3.59 | 588 | 1 |

data13$admit <- as.factor(data13$admit)  
summary(data13)

## GPA GMAT admit   
## Min. :2.130 Min. :313.0 1:31   
## 1st Qu.:2.600 1st Qu.:425.0 2:28   
## Median :3.010 Median :482.0 3:26   
## Mean :2.975 Mean :488.4   
## 3rd Qu.:3.300 3rd Qu.:538.0   
## Max. :3.800 Max. :693.0

mnv <- manova(cbind(GPA,GMAT)~admit, data=data13)  
summary(mnv)

## Df Pillai approx F num Df den Df Pr(>F)   
## admit 2 1.0096 41.797 4 164 < 2.2e-16 \*\*\*  
## Residuals 82   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Terlihat bahwa tolak H0 sehingga ada perbedaan yang signifikan

#Uji Homogenitas H0 : ∑1=∑2=∑3 H1 : Minimal ada satu yang berbeda

boxM(data13[,-3],data13$admit)

##   
## Box's M-test for Homogeneity of Covariance Matrices  
##   
## data: data13[, -3]  
## Chi-Sq (approx.) = 16.074, df = 6, p-value = 0.01336

Terlihat bahwa tidak memenuhi asumsi homogenitas (tolak H0). Namun pada tutorial kali ini akan dianggap memenuhi asumsi homogenitas.

#Uji Normalitas

mvn(data13[, -3])

## $multivariateNormality  
## Test HZ p value MVN  
## 1 Henze-Zirkler 1.136135 0.01526607 NO  
##   
## $univariateNormality  
## Test Variable Statistic p value Normality  
## 1 Anderson-Darling GPA 0.5759 0.1307 YES   
## 2 Anderson-Darling GMAT 0.4648 0.2483 YES   
##   
## $Descriptives  
## n Mean Std.Dev Median Min Max 25th 75th Skew  
## GPA 85 2.974588 0.4289954 3.01 2.13 3.8 2.6 3.3 -0.04826749  
## GMAT 85 488.447059 81.5223466 482.00 313.00 693.0 425.0 538.0 0.38849916  
## Kurtosis  
## GPA -1.0091148  
## GMAT -0.1362411

Terlihat bahwa data semua variabel berdistribusi normal

set.seed(123)  
k <- length(levels(data13$admit))  
n <- nrow(data13)  
# data training sebanyak  
idx <- sample(1:n, 36)  
data13.tr <- data13[idx,] #data training  
data13.ts <- data13[-idx,] #data testing  
# Analisis Diskriminan  
fit<-lda(admit~.,data=data13.tr)  
# Output dari analisis diskriminan  
fit

## Call:  
## lda(admit ~ ., data = data13.tr)  
##   
## Prior probabilities of groups:  
## 1 2 3   
## 0.2777778 0.3611111 0.3611111   
##   
## Group means:  
## GPA GMAT  
## 1 3.428000 585.4000  
## 2 2.515385 461.5385  
## 3 2.923846 457.8462  
##   
## Coefficients of linear discriminants:  
## LD1 LD2  
## GPA -5.861113047 2.2030403  
## GMAT -0.009791428 -0.0154151  
##   
## Proportion of trace:  
## LD1 LD2   
## 0.9724 0.0276

Prior : Jumlah sampel yg masuk Jumlah sampel yang masuk 28% dari group 1, 36% dari group 2, 36% dari group 3

Means : rata-rata tiap kelompok. Rata-rata GPA group1 adalah 3.4, group2 adalah 2.5, group3 adalah 2.9 Rata-rata GMAT group1 adalah 585.4, group2 adalah 461.53, group3 adalah 457.84

koefisien diskriminannya LD1 : -5.86GPA - 0.0098GMAT LD2 = 2.2GPA - 0.015GMAT

proporsi of trace : proporsi menjelaskan LD1 menjelaskan 97% dari data, sedangkan LD2 menjelaskan 3% dari data

zscore <- predict(fit,data13[,-3])  
knitr::kable(head(zscore$x, 10))

| LD1 | LD2 |
| --- | --- |
| -1.248352 | -1.4668431 |
| -1.099006 | 0.8257618 |
| -1.656018 | 0.8632691 |
| -2.506910 | 0.3238023 |
| -4.635944 | 1.5441507 |
| -5.128677 | -1.8605879 |
| -1.952372 | -1.7750833 |
| -3.252433 | -1.9929557 |
| -3.716375 | 2.3060442 |
| -4.862521 | 0.0443931 |

kor1 <- t(cor(zscore$x[,1],data13[,-3]))  
kor2 <- t(cor(zscore$x[,2],data13[,-3]))  
tabkor<-data.frame(kor1,kor2)  
tabkor

## kor1 kor2  
## GPA -0.9710885 0.3119494  
## GMAT -0.6592011 -0.6996044

pred <- predict(fit,data13.ts[,-3])  
pred

## $class  
## [1] 3 3 3 1 1 1 1 1 1 1 1 1 1 1 1 1 1 3 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3  
## [39] 3 3 1 3 3 3 3 3 3 3 3  
## Levels: 1 2 3  
##   
## $posterior  
## 1 2 3  
## 1 2.423080e-01 5.248747e-03 7.524432e-01  
## 2 2.357844e-02 1.079055e-03 9.753425e-01  
## 3 1.957225e-01 2.307773e-04 8.040467e-01  
## 4 9.324878e-01 4.366146e-06 6.750786e-02  
## 5 9.999733e-01 3.545382e-12 2.667887e-05  
## 6 9.999998e-01 1.973107e-13 1.825904e-07  
## 7 9.995659e-01 4.441686e-08 4.340786e-04  
## 8 9.919158e-01 3.302344e-07 8.083831e-03  
## 9 9.985600e-01 1.300292e-08 1.439965e-03  
## 10 9.999472e-01 8.268495e-11 5.275590e-05  
## 11 9.990919e-01 2.824614e-09 9.080749e-04  
## 12 9.996905e-01 9.242793e-10 3.095306e-04  
## 13 9.736707e-01 6.941992e-07 2.632865e-02  
## 14 9.968919e-01 3.577356e-08 3.108048e-03  
## 15 8.984928e-01 7.959115e-06 1.014992e-01  
## 16 7.247398e-01 6.207388e-05 2.751981e-01  
## 17 9.999868e-01 6.184420e-12 1.319947e-05  
## 18 4.777424e-01 1.304243e-05 5.222446e-01  
## 19 9.979445e-01 2.847089e-08 2.055518e-03  
## 20 9.999991e-01 1.845609e-14 9.151377e-07  
## 21 1.000000e+00 3.427203e-17 2.548323e-09  
## 22 1.264447e-12 9.926243e-01 7.375656e-03  
## 23 3.759979e-15 9.997133e-01 2.866532e-04  
## 24 6.493805e-11 9.941372e-01 5.862825e-03  
## 25 1.750313e-14 9.982297e-01 1.770348e-03  
## 26 2.150635e-10 9.661961e-01 3.380387e-02  
## 27 1.661307e-14 9.980421e-01 1.957947e-03  
## 28 7.381813e-09 9.610313e-01 3.896867e-02  
## 29 3.064541e-07 7.422911e-01 2.577086e-01  
## 30 7.051007e-15 9.958686e-01 4.131398e-03  
## 31 4.003649e-18 9.999458e-01 5.423045e-05  
## 32 9.197553e-12 9.922619e-01 7.738065e-03  
## 33 3.217886e-11 9.908705e-01 9.129531e-03  
## 34 8.498610e-17 9.991278e-01 8.722495e-04  
## 35 1.786522e-07 1.945736e-01 8.054262e-01  
## 36 7.362820e-07 9.604532e-02 9.039539e-01  
## 37 1.198049e-04 7.431685e-02 9.255633e-01  
## 38 1.280637e-03 1.730317e-03 9.969890e-01  
## 39 2.280937e-03 2.777943e-04 9.974413e-01  
## 40 5.057836e-06 3.569133e-03 9.964258e-01  
## 41 6.479848e-01 2.276832e-06 3.520130e-01  
## 42 1.690044e-05 2.020070e-01 7.979761e-01  
## 43 1.375363e-04 2.308011e-01 7.690614e-01  
## 44 8.812797e-03 1.636376e-03 9.895508e-01  
## 45 1.021910e-03 3.769988e-03 9.952081e-01  
## 46 5.540205e-05 8.524615e-03 9.914200e-01  
## 47 4.187435e-04 1.025415e-02 9.893271e-01  
## 48 7.927687e-05 1.032527e-02 9.895955e-01  
## 49 5.683882e-04 6.763850e-03 9.926678e-01  
##   
## $x  
## LD1 LD2  
## 1 -1.2483517 -1.466843072  
## 2 -1.0990064 0.825761783  
## 3 -1.6560183 0.863269086  
## 4 -2.5069104 0.323802299  
## 5 -4.6359442 1.544150677  
## 6 -5.1286767 -1.860587853  
## 7 -3.2524333 -1.992955664  
## 8 -2.9180130 -0.209110986  
## 9 -3.4062100 0.165344070  
## 10 -4.1783584 0.092761155  
## 11 -3.6303133 0.726964855  
## 12 -3.8066965 0.334971994  
## 13 -2.7900372 0.563890434  
## 14 -3.2496846 0.297464691  
## 15 -2.4091336 0.363432305  
## 16 -2.0673957 0.101313766  
## 17 -4.5689161 0.392325199  
## 18 -2.2116558 2.045986567  
## 19 -3.2891252 0.006762246  
## 20 -5.4373295 1.539843472  
## 21 -6.4268135 -0.475165945  
## 22 3.5323724 0.001528079  
## 23 4.4006485 -1.260511212  
## 24 2.9047590 -1.786685601  
## 25 4.1872986 0.118171800  
## 26 2.7403663 -0.330927076  
## 27 4.1972275 0.248107921  
## 28 2.1818426 -1.553150968  
## 29 1.5866270 -0.775656951  
## 30 4.3451984 1.395502599  
## 31 5.4571606 -0.397327273  
## 32 3.2187719 -0.720797234  
## 33 3.0229433 -1.029099283  
## 34 5.0195700 1.428455507  
## 35 1.5015278 1.605069517  
## 36 1.1790979 1.668976225  
## 37 0.3364853 -0.114806655  
## 38 -0.5702692 1.658177315  
## 39 -0.9208365 2.706527875  
## 40 0.4090110 3.314208576  
## 41 -2.5138157 2.713328569  
## 42 0.7859289 -0.207359067  
## 43 0.4720535 -1.158726416  
## 44 -0.8838698 0.935852002  
## 45 -0.4242225 1.202277745  
## 46 0.1530595 1.768205736  
## 47 -0.1412331 0.847668592  
## 48 0.1234103 1.492918393  
## 49 -0.2485265 1.021665518

Posterior : peluang masuk kelas, diambil yang terbesar di setiap observasi.

ct <- table(data13.ts$admit, pred$class)  
ct

##   
## 1 2 3  
## 1 17 0 4  
## 2 0 13 2  
## 3 1 0 12

sum(diag(prop.table(ct)))

## [1] 0.8571429

Observasi yang masuk kelas 1 dan merupakan kelas 1 ada 17, sedangkan yang masuk kelas 3 dan merupakan kelas 1 ada 1 (ada 1 observasi yang salah masuk kelas). Observasi yang masuk kelas 2 dan merupakan kelas 2 ada 13. Observasi yang masuk kelas 3 dan merupakan kelas 3 ada 12, sedangkan yang masuk kelas 1 dan merupakan kelas 3 ada 4, yang masuk kelas 2 dan merupakan kelas 3 ada 2 (ada 6 observasi yang salah masuk kelas). Terlihat bahwa model berhasil mengklasifikasikan hampir semua data yaitu 85,7% dengan benar

lda.data <- cbind(data13.tr, predict(fit)$x)  
ggplot(lda.data, aes(LD1, LD2)) +  
geom\_point(aes(color = as.factor(admit))) +  
theme\_minimal()

