TUGAS 3

LOGISTIC REGRESSION DAN MULTIDIMENSIONAL SCALING

Nama : Surya Abdillah

NRP : 5025201229

Kelas : Analisis Data Multivariat A (2023)

Dosen Pengampu : Dr. Ahmad Saikhu, S.Si., MT.

LOGISTIC REGRESSION

Pertanyaan

Terapkan logistic regression dengan sumber pada buku Joseph halaman 333

TABLE 1 Group Descriptive Statistics and Tests of Equality for the Estimation Sample

	Dependent Variable Group Means: X_4 Region			
Independent Variables	**Group 0: USA/North America ($n = 26$)**	**Group 1: Outside North America ($n = 34$)**	***F* Value**	**Significance**
X_6 Product Quality	8.527	7.297	14.387	.000
X_7 E-Commerce Activities	3.388	3.626	2.054	.157
X_8 Technical Support	5.569	5.050	1.598	.211
X_9 Complaint Resolution	5.577	5.253	.849	.361
X_{10} Advertising	3.727	3.979	.775	.382
X_{11} Product Line	6.785	5.274	25.500	.000
X_{12} Salesforce Image	4.427	5.238	9.733	.003
X_{13} Competitive Pricing	5.600	7.418	31.992	.000
X_{14} Warranty & Claims	6.050	5.918	.453	.503
X_{15} New Products	4.954	5.276	.600	.442
X_{16} Order & Billing	4.231	4.153	.087	.769
X_{17} Price Flexibility	3.631	4.932	31.699	.000
X_{18} Delivery Speed	3.873	3.794	.152	.698

Pengerjaan

Pada buku tersebut digunakan dataset HBAT dengan jumlah 100 baris data. Berdasarkan teori yang sudah dijelaskan, logistic regression memiliki skema tersendiri dalam menilai mana kolom yang memiliki pengaruh signifikan dalam model prediksinya. Sehingga, dalam percobaan ini akan dilakukan scenario pengujian, yaitu:

a. tanpa droping,
b. droping fitur berdasar nilai korelasi
c. droping fitur dengan pengaruh yang tidak siginifikan

Beberapa Langkah yang dilakukan dalam percobaan Logistic Regression ini, yaitu:

1. Import library yang diperlukan
2. Load dataset, yakni data HBAT 100 baris data dengan kolom sesuai pada **table 1**
3. Pembuatan matriks korelasi untuk scenario pengujian 3:

	X4	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15	X16	X17	X18
X4	1.00	-0.52	0.19	-0.17	-0.01	0.19	-0.55	0.40	0.55	-0.15	0.11	0.04	0.62	0.02
X6	-0.52	1.00	-0.14	0.10	0.11	-0.05	0.48	-0.15	-0.40	0.09	0.03	0.10	-0.49	0.03
X7	0.19	-0.14	1.00	0.00	0.14	0.43	-0.05	0.79	0.23	0.05	-0.03	0.16	0.27	0.19
X8	-0.17	0.10	0.00	1.00	0.10	-0.06	0.19	0.02	-0.27	0.80	-0.07	0.08	-0.19	0.03
X9	-0.01	0.11	0.14	0.10	1.00	0.20	0.56	0.23	-0.13	0.14	0.06	0.76	0.39	0.87
X10	0.19	-0.05	0.43	-0.06	0.20	1.00	-0.01	0.54	0.13	0.01	0.08	0.18	0.33	0.28
X11	-0.55	0.48	-0.05	0.19	0.56	-0.01	1.00	-0.06	-0.49	0.27	0.05	0.42	-0.38	0.60
X12	0.40	-0.15	0.79	0.02	0.23	0.54	-0.06	1.00	0.26	0.11	0.03	0.20	0.35	0.27
X13	0.55	-0.40	0.23	-0.27	-0.13	0.13	-0.49	0.26	1.00	-0.24	0.02	-0.11	0.47	-0.07
X14	-0.15	0.09	0.05	0.80	0.14	0.01	0.27	0.11	-0.24	1.00	0.04	0.20	-0.17	0.11
X15	0.11	0.03	-0.03	-0.07	0.06	0.08	0.05	0.03	0.02	0.04	1.00	0.07	0.09	0.11
X16	0.04	0.10	0.16	0.08	0.76	0.18	0.42	0.20	-0.11	0.20	0.07	1.00	0.41	0.75
X17	0.62	-0.49	0.27	-0.19	0.39	0.33	-0.38	0.35	0.47	-0.17	0.09	0.41	1.00	0.50
X18	0.02	0.03	0.19	0.03	0.87	0.28	0.60	0.27	-0.07	0.11	0.11	0.75	0.50	1.00

Dari matriks korelasi tersebut, akan diambil fitur dengan nilai korelasi diatas 0.3 atau dibawah -0,3, yakni fitur X6, X11, X12, X13, dan X17

4. Dilakukan oembuatan model logistic regression menggunakan fungsi glm dari glmnet

```
1  logistic <- glm(X4 ~ ., data = df_classification, family = "binomial")
2  tidy(logistic)
```

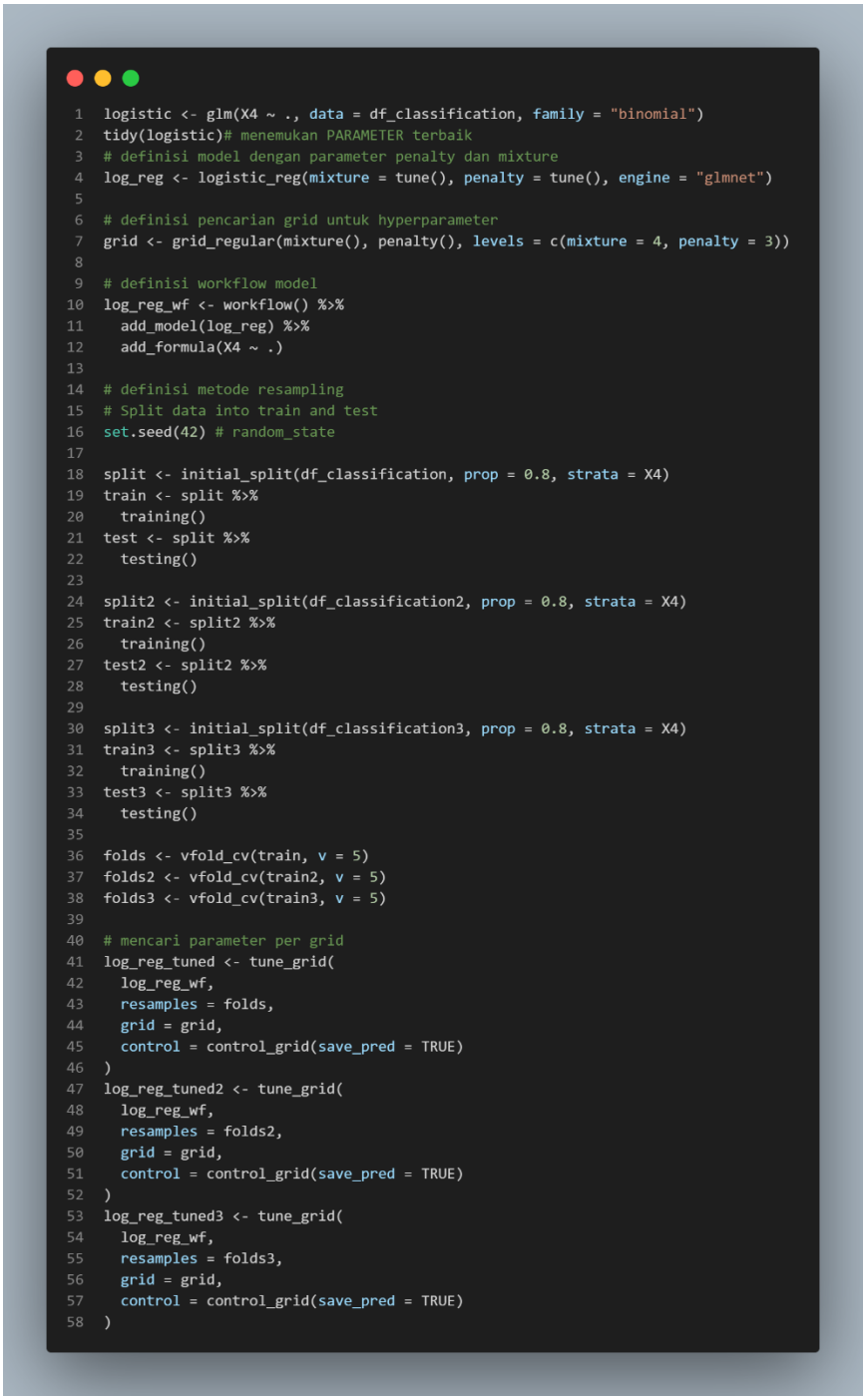
Didapatkan hasil sebagai berikut:

```
> tidy(logistic)
# A tibble: 14 × 5
   term        estimate std.error statistic p.value
   <chr>          <dbl>     <dbl>     <dbl>   <dbl>
 1 (Intercept) -23.8       17.1      -1.40  0.163
 2 X6           -1.07       0.646    -1.66  0.0968
 3 X7           -5.49       2.27     -2.42  0.0156
 4 X8            0.268      0.666     0.403 0.687
 5 X9            0.134      0.824     0.162 0.871
 6 X10          -1.91       0.898    -2.13  0.0332
 7 X11           2.64       5.50      0.480 0.631
 8 X12           8.20       3.11      2.64  0.00835
 9 X13           0.0711     0.532     0.134 0.894
10 X14           0.367      1.17      0.312 0.755
11 X15           0.364      0.355     1.03  0.305
12 X16           1.45       1.08      1.35  0.177
13 X17           8.88       6.41      1.39  0.166
14 X18         -12.2       11.3      -1.08  0.280
```

Nilai siginificance masing-masing atribut dapat dilihat melalui nilai p.value, dimana semakin kecil nilai p.value, maka semakin signifikan pengaruh predictor tersebut terhadap fitur target. Berdasar hasil ini, didapati bahwa predictor dengan pengaruh yang cukup signifikan adalah X4, X6, X7, X12, dan X13. Sehingga, fitur predictor yang akan digunakan pada scenario pengujian 3 adalah 5 prediktor teersebut.

5. Dilakukan pencarian parameter terbaik dengan alur sebagai berikut:

a. Pendefinisian model
b. Pencarian grid untuk hyper parameter
c. Pendefinisian workflow model
d. Penentuan model resampling
e. Splitting data
f. Pembuatan skema cross validation
g. Pencarian parameter per-grid



```r
logistic <- glm(X4 ~ ., data = df_classification, family = "binomial")
tidy(logistic)# menemukan PARAMETER terbaik
# definisi model dengan parameter penalty dan mixture
log_reg <- logistic_reg(mixture = tune(), penalty = tune(), engine = "glmnet")

# definisi pencarian grid untuk hyperparameter
grid <- grid_regular(mixture(), penalty(), levels = c(mixture = 4, penalty = 3))

# definisi workflow model
log_reg_wf <- workflow() %>%
  add_model(log_reg) %>%
  add_formula(X4 ~ .)

# definisi metode resampling
# Split data into train and test
set.seed(42) # random_state

split <- initial_split(df_classification, prop = 0.8, strata = X4)
train <- split %>%
  training()
test <- split %>%
  testing()

split2 <- initial_split(df_classification2, prop = 0.8, strata = X4)
train2 <- split2 %>%
  training()
test2 <- split2 %>%
  testing()

split3 <- initial_split(df_classification3, prop = 0.8, strata = X4)
train3 <- split3 %>%
  training()
test3 <- split3 %>%
  testing()

folds <- vfold_cv(train, v = 5)
folds2 <- vfold_cv(train2, v = 5)
folds3 <- vfold_cv(train3, v = 5)

# mencari parameter per grid
log_reg_tuned <- tune_grid(
  log_reg_wf,
  resamples = folds,
  grid = grid,
  control = control_grid(save_pred = TRUE)
)
log_reg_tuned2 <- tune_grid(
  log_reg_wf,
  resamples = folds2,
  grid = grid,
  control = control_grid(save_pred = TRUE)
)
log_reg_tuned3 <- tune_grid(
  log_reg_wf,
  resamples = folds3,
  grid = grid,
  control = control_grid(save_pred = TRUE)
)
```

Metrik yang digunakan untuk menilai model adalah ROC-AUC, yakni Receiver Operating Characteristics-Area Under the Curve. Dari hasil pencarian tersebut, didapatkan nilai parameter terbaik untuk masing-masing skema pengujian, yaitu:

- Skenario pengujian 1:

```
       penalty mixture .config
         <dbl>   <dbl> <chr>
1 0.0000000001   0.667 Preprocessor1_Model07
```

- Skenario pengujian 2:

```
       penalty mixture .config
         <dbl>   <dbl> <chr>
1 0.0000000001       0 Preprocessor1_Model01
```

- Scenario pengujian 3:

```
       penalty mixture .config
         <dbl>   <dbl> <chr>
1 0.0000000001   0.333 Preprocessor1_Model04
```

6. Pembuatan model dengan parameter hasil hyperparameter. Hasil dari evaluasi dengan akurasi, recall, dan confusion matriks, sebagai berikut:
   - Skenario pengujian 1:
     Akurasi

```
> mean(pred_class==test$X4)
[1] 0.8095238
```

     Recall

```
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 recall  binary          0.75
```

     Confusion matrix

```
          Truth
Prediction  0  1
         0  6  2
         1  2 11
```

   - Skenario pengujian 2:
     Akurasi

```
> mean(pred_class2==test2$X4)
[1] 0.8571429
```

     Recall

```
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 recall  binary         0.875
```

     Confusion matrix

```
          Truth
Prediction  0  1
         0  7  2
         1  1 11
```

   - Skenario pengujian 3:
     Akurasi

```
> mean(pred_class3==test3$X4)
[1] 1
```

     Recall

```
  .metric .estimator .estimate
  <chr>   <chr>           <dbl>
1 recall  binary              1
```

Confusion matrix

```
           Truth
Prediction  0  1
         0  8  0
         1  0 13
```

Analisis dan Kesimpulan

- **Nilai significance predictor**
  Berdasarkan gambar di bawah, didapati bahwa predictor dengan yang memiliki nilai p.value kurang dari 0.05 adalah X7, X10, dan X12. Sehingga dapat ditarik kesimpulan bahwa

a. fitur X7 (E-commerce activities) memiliki asosiasi yang signifikan terhadap X4 (Region)
b. fitur X10 (Advertising) memiliki asosiasi yang signifikan terhadap X4 (Region)
c. fitur X12 (Salesfoce Image) memiliki asosiasi yang signifikan terhadap X4 (Region)

```
> tidy(logistic)
# A tibble: 14 × 5
   term        estimate std.error statistic p.value
   <chr>          <dbl>     <dbl>     <dbl>   <dbl>
 1 (Intercept) -23.8       17.1      -1.40   0.163
 2 X6           -1.07       0.646    -1.66   0.0968
 3 X7           -5.49       2.27     -2.42   0.0156
 4 X8            0.268      0.666     0.403  0.687
 5 X9            0.134      0.824     0.162  0.871
 6 X10          -1.91       0.898    -2.13   0.0332
 7 X11           2.64       5.50      0.480  0.631
 8 X12           8.20       3.11      2.64   0.00835
 9 X13           0.0711     0.532     0.134  0.894
10 X14           0.367      1.17      0.312  0.755
11 X15           0.364      0.355     1.03   0.305
12 X16           1.45       1.08      1.35   0.177
13 X17           8.88       6.41      1.39   0.166
14 X18         -12.2       11.3      -1.08   0.280
```

- **Hasil Skenario Pengujian**
  Hasil evaluasi masing-masing scenario pengujian dapat dilihat pada table berikut:

Aspek	Skenario Pengujian		
	1	2	3
Akurasi	0.8095238	0.8571429	1
Presisi	0.75	0.778	1
Recall	0.75	0.875	1

Confusion Matriks	Truth Prediction 0 1 0 6 2 1 2 11	Truth Prediction 0 1 0 7 2 1 1 11	Truth Prediction 0 1 0 8 0 1 0 13

Model dengan akurasi tertinggi: skenario pengujian 3, 2, 1

Model dengan presisi tertinggi: skenario pengujian 3, 2, 1

Model dengan recall tertinggi: skenario pengujian 3, 2, 1

Dalam hal ini model dengan skenario pengujian 3 (pengambilan fitur dengan nilai significanse yang cukup tinggi) berhasil melakukan prediksi dengan benar semua. Sehingga, model ini dapat dikatakan terbaik.

Adapun, nilai presisi dapat digunakan sebagai pertimbangan dikarenakan kondisi kesalahan prediksi 0 (dalam NA) akan lebih baik salah prediksi menjadi 1 (luar NA) daripada prediksi 1 (luar NA) menjadi 0 (dalam NA) sehingga budgeting promosi dan sejenisnya tidak akan kurang.

MULTIDIMENSIONAL SCALING (MDS)

Pertanyaan

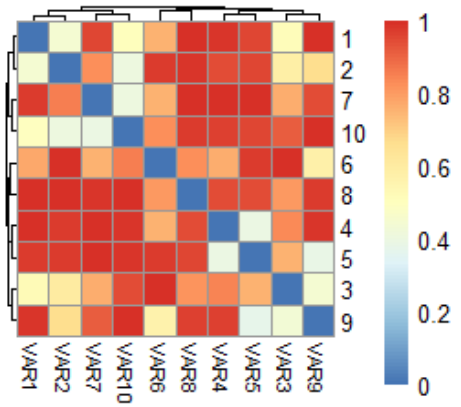
Lakukan multidimensional scaling pada dataset HBAT_MDS

Pengerjaan

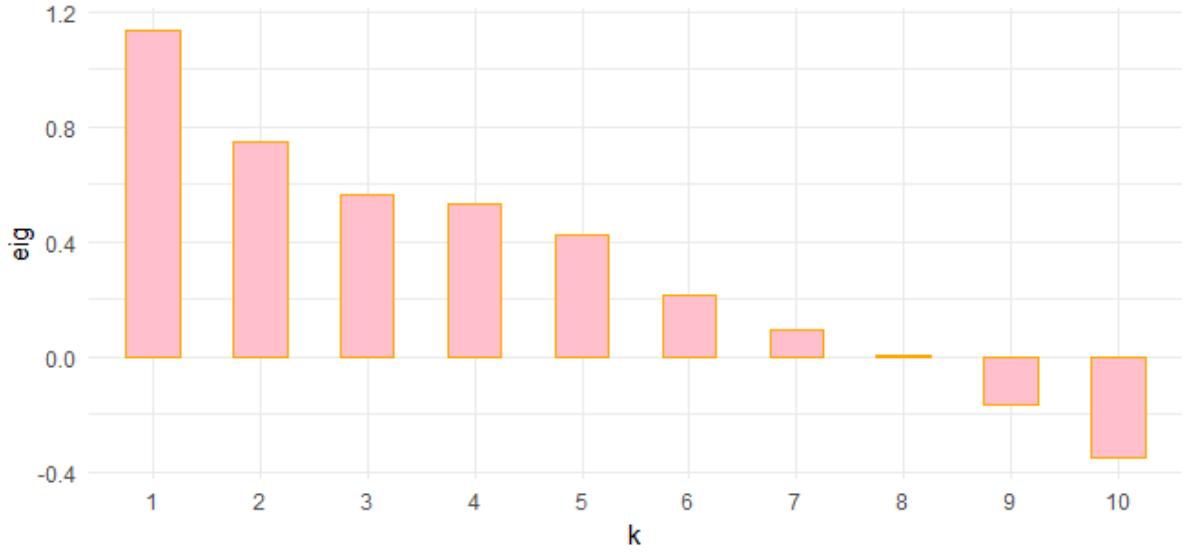
Beberapa langkah yang dilakukan adalah:

1. Import library yang diperlukan
2. Load datase HBAT_MDS, yakni terdiri dari 18 *distance matrix*. Sehingga, akan dilakukan penjumlahan seluruh nilai pada indeks yang sesuai. Dilanjutkan dengan pengisian segitigas atas data frame
3. Normalisasi dengan min max (tidak harus dilakukan, hanya untuk keperluan mempermudah screeplot)
4. Pembuatan heatmap representasi lain dari distance matriks
5. Melakukan MDS dengan metric dimensional scaling
6. Bar plot menunjukan penyebaran variansi pada dimensi
7. Pembuatan scree plot

Analisis dan Kesimpulan

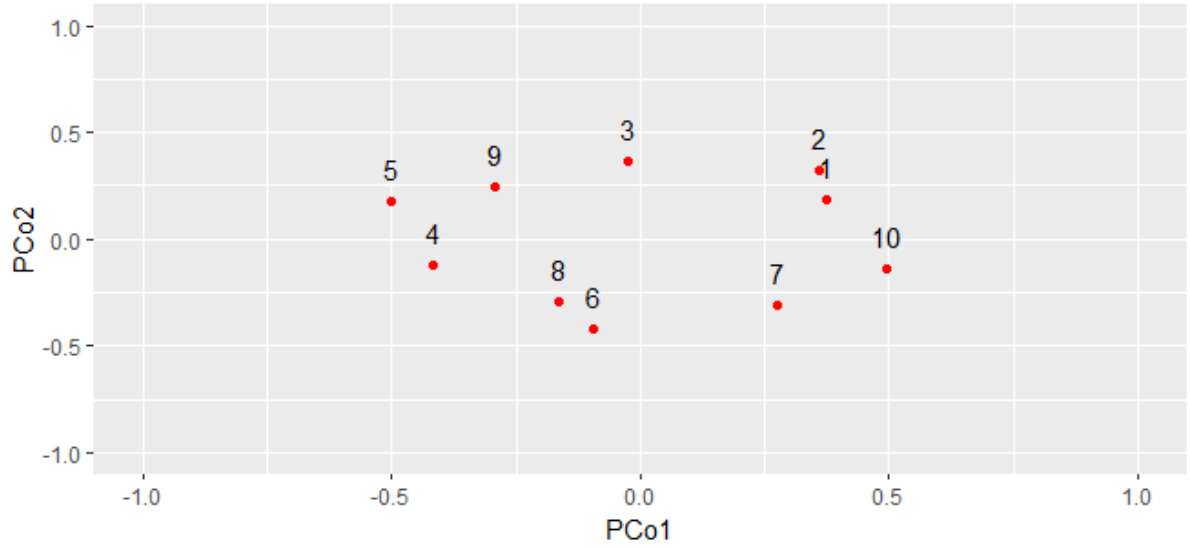


Pada heatmap diatas, dapat terlihat bahwa nilai jarak pada setiap kolom hampir memiliki jarak yang cukup jauh, kecuali pada beberapa bagian. Antara lain, jarak antara V10 dengan VAR2 dan VAR7, VAR4 dengan VAR5, VAR9 dengan VAR3 dan VAR5.



Pada barplot di atas dapat kita simpulkan bahwa penyebaran variansi cukup beragam, dimana nilai variansi termasuk rendah pada posisi variabel 8 saja. Sehingga, apabila ingin didapatkan informasi yang cukup representatif, masih diperlukan cukup banyak variabel, dalam kasus ini

dapat 1-5 variabel



Screeplot di atas menggambarkan jarak antar titik. Dari hasil tersebut terdapata beberapa titik yang berdekatan yang kedepannya dapat digunakan sebagai bahan clusterisasi atau pemrosesan lainnya, antara lain:

- Titik 8 dan 6
- Titik 2 dan 1

Titik yang memiliki karakteristik paling beda dari yang lainnya adalah titik 3