EVALUASI AKHIR SEMESTER

Nama : Surya Abdillah

NRP : 5025201229

Kelas : Analisis Data Multivariat A (2023)

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

Source Code : https://drive.google.com/drive/folders/1-J_E4CeMXmqbJPhsQ8Na2MK9iNLBmukl?usp=sharing

LOGISTIC REGRESSION

Pertanyaan

Lakukan pemodelan regresi logistik dengan variabel dependent (class) adalah fitur ‘am’

Pengerjaan

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

1. Import library yang diperlukan
2. Load dataset, yakni data soal_logistic_regr.csv , terdapat 32 baris data dan 11 atribut. Dimana atribut am dan vs merupakan nominal, segingga akan dilakukan perubahan perubahan menjadi factor
3. Atribut model memiliki 32 unique value, maka bisa dikatakan model ini merupakan ID dalam data, sehingga keberadaannya tidak diperlukan dalam komputasi sehingga dapat di drop
4. Dilakukan pembuatan model logistic regression menggunakan fungsi glm dari glmnet

```
logistic <- glm(am ~ ., data = df_lr, family = "binomial")
tidy(logistic)
```

Didapatkan hasil sebagai berikut:

	term	estimate	std.error	statistic	p.value
	<chr>	<dbl>	<dbl>	<dbl>	<dbl>
1	(Intercept)	-2.56e+ 1	1514865.	-1.69e- 5	1.00
2	mpg	2.56e- 9	17182.	1.49e-13	1.00
3	cyl	-3.24e- 8	82581.	-3.93e-13	1.00
4	disp	4.60e-11	1471.	3.13e-14	1.00
5	hp	6.99e-10	1806.	3.87e-13	1.00
6	drat	-4.00e- 8	132999.	-3.01e-13	1.00
7	wt	-1.53e- 8	167920.	-9.08e-14	1.00
8	qsec	5.31e- 9	58136.	9.13e-14	1.00
9	vs	5.11e+ 1	167753.	3.05e- 4	1.00
10	gear	-1.10e- 8	117475.	-9.38e-14	1.00
11	carb	-9.54e- 9	67567.	-1.41e-13	1.00

Didapati bahwa nilai p.value pada semua prediktor bernilai 1, hal ini menandakan bahwa terima hipotesis null, sehingga tidak ada prediktor yang berpengaruh signifikan pada target

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
# 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(am ~ .)

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

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

folds <- vfold_cv(train, v = 3)

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

select_best(log_reg_tuned, metric = "roc_auc")
```

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 model adalah:

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

6. Pembuatan model dengan parameter hasil hyperparameter. Hasil dari evaluasi dengan akurasi, presisi, recall, dan confusion matriks, sebagai berikut:

   Akurasi

```
> mean(pred_class_mat==test$am)
[1] 0.8571429
```

   Presisi

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

Recall

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

Confusion matrix

```
          Truth
Prediction 0 1
         0 3 0
         1 1 3
```

Analisis dan Kesimpulan

- **Nilai significance predictor**
  Berdasarkan gambar dibawah didapati semua prediktor memiliki p.value yang sama, yakni 1 (> 0.05). Hal ini mengindikasikan bahwa tidak ada prediktor yang memiliki pengaruh/asosiasi yang signifikan dengan target. Namun, hal ini dapat diakibatkan oleh minim nya data yang ada (hanya 32 data). Sehingga kedepannya jumlah data bisa ditambahkan untuk bisa mengetahui prediktor mana yang memiliki pengaruh yang signifikan.

```
   term         estimate std.error statistic p.value
   <chr>           <dbl>     <dbl>     <dbl>   <dbl>
 1 (Intercept) -2.56e+ 1  1514865. -1.69e- 5    1.00
 2 mpg          2.56e- 9    17182.  1.49e-13    1.00
 3 cyl         -3.24e- 8    82581. -3.93e-13    1.00
 4 disp         4.60e-11     1471.  3.13e-14    1.00
 5 hp           6.99e-10     1806.  3.87e-13    1.00
 6 drat        -4.00e- 8   132999. -3.01e-13    1.00
 7 wt          -1.53e- 8   167920. -9.08e-14    1.00
 8 qsec         5.31e- 9    58136.  9.13e-14    1.00
 9 vs           5.11e+ 1   167753.  3.05e- 4    1.00
10 gear        -1.10e- 8   117475. -9.38e-14    1.00
11 carb        -9.54e- 9    67567. -1.41e-13    1.00
```

- **Pembuatan model**
  Sepertti yang dijelaskan pada bagian sebelumnya, bahwa data yang ada sangat sedikit, hal ini juga memengaruhi kualitas dari pembuatan model. Hal ini, dapat didukung dengan adanya warning pada R yang mengidikasikan bahwa jumlah distribusi label kurang dari 8 pada proses hyperparameter tuning.

```
+ )
→ A | warning: one multinomial or binomial class has fewer than 8  obs
ervations; dangerous ground
There were issues with some computations   A: x4
```

- **Hasil Pembuatan Model**
  Hasil evaluasi masing-masing scenario pengujian dapat dilihat pada table berikut:

Presisi	Recall	Akurasi	Confusion Matrix

1	0.75	0.8571429	Truth Prediction 0 1 0 3 0 1 1 3

Komponen penilaian utama merupakan akurasi karena tidak ada kasus yang menonjol (tidak seperti dalam kasus prediksi covid, akan lebih baik salah memprediksi orang sehat menjadi sakit daripada orang sakit terprediksi sehat). Dengan menimbang hal tersebut, maka dapat disimpulkan bahwa model logistic regression yang dibuat sudah sukses memprediksi atribut target (am) menggunakan prediktor yang ada.

LINEAR DISCRIMINANT ANALYSIS

Pertanyaan

Lakukan pemodelan Linear Discriminant Analysis dengan variabel dependent (Class) adalah kolom F.

Pengerjaan

Beberapa langkah yang dilakukan adalah:

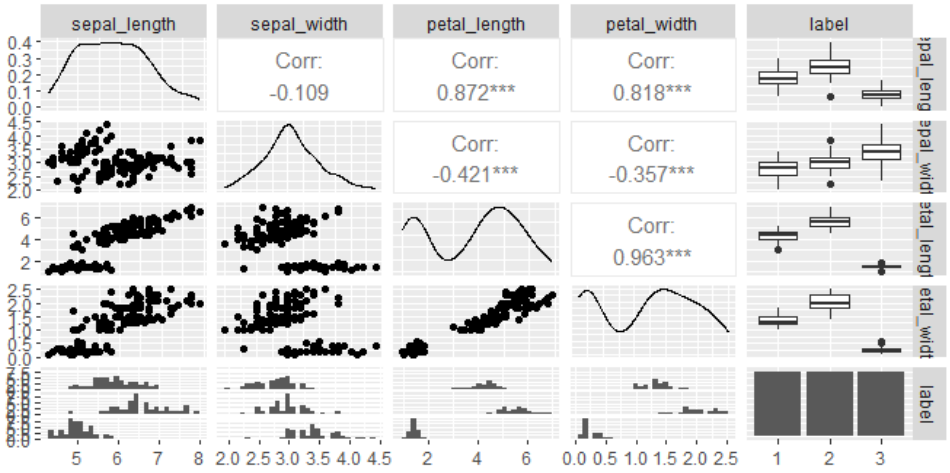
1. Import library yang diperlukan
2. Load dataset yang digunakan, yakni 'soal Discriinant Analysis.xlsx'. Dropping fitur Class karena sudah diwakili oleh label encoding (kolom G)
3. Berikut merupakan informasi yang didapat:
    - Dimensi: 150 baris data dengan 5 atribut
    - Skala data pada setiap atribut bersifat homogen
    - Terdapat 3 kelas target, yakni versicolor, virginica, dan setosa dengan distribusi yang sama, yakni 50 data per kelas
    - Korelasi antar variabel. Berdasarkan matriks korelasi dibawah terjadi multicolinearity, dimana antar independent variable memiliki korelasi yang sangat kuat, yakni pada sepal_length dengan petal_length dan petal_width, serta petal_length dengan petal_width.
      Pada umumnya akan dilakukan penanganan multicolinearity dengan droping fitur atau lainnya. Namun, dalam kondisi jumlah data dan atribut yang sudah sedikit, maka penanganan tersebut tidak akan dilakukan.

```
> cor_matrix
             sepal_length sepal_width petal_length petal_width
sepal_length         1.00       -0.11         0.87        0.82
sepal_width         -0.11        1.00        -0.42       -0.36
petal_length         0.87       -0.42         1.00        0.96
petal_width          0.82       -0.36         0.96        1.00
label               -0.46        0.61        -0.65       -0.58
             label
sepal_length -0.46
sepal_width   0.61
petal_length -0.65
petal_width  -0.58
label         1.00
```

Atribut prediktor sudah memiliki korelasi yang baik terhadap target

```
sepal_length  sepal_width petal_length  petal_width         label
       -0.46         0.61        -0.65        -0.58          1.00
```

- Pairing antar variabel



4. Pemenuhan asumsi MANOVA, yakni
   a. Asumsi normalitas multivariat
      - Sepal length: p-value < 0.05, reject H0
        Sepal length berdistribusi normal

```
        Shapiro-Wilk normality test

data:  Z
W = 0.97609, p-value = 0.01018
```

      - Sepal width: p-value > 0.05, not reject H0
        Sepal width tidak berdistribusi normal

```
        Shapiro-Wilk normality test

data:  Z
W = 0.98379, p-value = 0.07518
```

      - Petal length: p-value < 0.05, reject H0
        Petal length berdistribusi normal

```
        Shapiro-Wilk normality test

data:  Z
W = 0.87642, p-value = 7.545e-10
```

      - Petal width: p-value < 0.05, reject H0
        Petal width berdistribusi normal

```
        Shapiro-Wilk normality test

data:  Z
W = 0.90262, p-value = 1.865e-08
```

   b. Kesamaan matriks kovarian
      p-value < 0.05, reject H0
      terdapat perbedaan variance antar group

```
        Bartlett test of homogeneity of variances

data:  df_lda[, 0:4]
Bartlett's K-squared = 295.06, df = 3, p-value < 2.2e-16
```

5. Uji MANOVA
    a. Analisis secara simultan
        - Pillai, p-value < 0.05, reject H0
          Terdapat perbedaan nilai yang signifikan antar grup pada kombinasi variabel

```
              Df Pillai approx F num Df den Df    Pr(>F)
df_lda$label   2 1.1872   52.949      8    290 < 2.2e-16 ***
Residuals    147
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

        - Roy p-value < 0.05, reject H0
          Terdapat perbedaan nilai yang signifikan antar grup pada kombinasi variabel

```
              Df    Roy approx F num Df den Df    Pr(>F)
df_lda$label   2 32.272   1169.9      4    145 < 2.2e-16 ***
Residuals    147
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

        - Wilks Lambda p-value < 0.05, reject H0
          Terdapat perbedaan nilai yang signifikan antar grup pada kombinasi variabel

```
              Df    Wilks approx F num Df den Df    Pr(>F)
df_lda$label   2 0.023526   198.71      8    288 < 2.2e-16 ***
Residuals    147
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

        - Hotelling-Lawley p-value < 0.05, reject H0
          Terdapat perbedaan nilai yang signifikan antar grup pada kombinasi variabel

```
              Df Hotelling-Lawley approx F num Df den Df    Pr(>F)
df_lda$label   2            32.55   581.82      8    286 < 2.2e-16
Residuals    147

df_lda$label ***
Residuals
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

    b. Analisis secara parsial, p-value < 0.05 , reject H0
        - Sepal length berpengaruh signifikan pada label
        - Sepal width berpengaruh signifikan pada label
        - Petal length berpengaruh signifikan pada label
        - Petal width berpengaruh signifikan pada label

```
> summary.aov(manova_res)
 Response 1 :
             Df Sum Sq Mean Sq F value    Pr(>F)
df_lda$label   2 63.212  31.606  119.26 < 2.2e-16 ***
Residuals    147 38.956   0.265
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Response 2 :
             Df Sum Sq Mean Sq F value    Pr(>F)
df_lda$label   2 10.978  5.4888  47.364 < 2.2e-16 ***
Residuals    147 17.035  0.1159
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Response 3 :
             Df Sum Sq Mean Sq F value    Pr(>F)
df_lda$label   2 436.64 218.322    1179 < 2.2e-16 ***
Residuals    147  27.22   0.185
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Response 4 :
             Df Sum Sq Mean Sq F value    Pr(>F)
df_lda$label   2 80.604  40.302  959.32 < 2.2e-16 ***
Residuals    147  6.176   0.042
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

6. Spliting data set menjadi training dan testing, dengan perbandingan 70% : 30%
7. Pembuatan model LDA

```
Call:
lda(label ~ ., data = train)

Prior probabilities of groups:
        1         2         3
0.3333333 0.3333333 0.3333333

Group means:
  sepal_length sepal_width petal_length petal_width
1     5.911429    2.740000     4.185714   1.2914286
2     6.614286    2.960000     5.585714   2.0171429
3     5.011429    3.402857     1.442857   0.2457143

Coefficients of linear discriminants:
                    LD1        LD2
sepal_length -1.306494  0.5401259
sepal_width  -1.545135 -2.5091806
petal_length  2.655556  0.4561852
petal_width   3.010180 -2.3382506

Proportion of trace:
   LD1    LD2
0.9929 0.0071
```

8. Pengujian pada data testing

Analisis dan Kesimpulan

```
Call:
lda(label ~ ., data = train)

Prior probabilities of groups:
        1         2         3
0.3333333 0.3333333 0.3333333

Group means:
  sepal_length sepal_width petal_length petal_width
1     5.911429    2.740000     4.185714   1.2914286
2     6.614286    2.960000     5.585714   2.0171429
3     5.011429    3.402857     1.442857   0.2457143

Coefficients of linear discriminants:
                    LD1        LD2
sepal_length -1.306494  0.5401259
sepal_width  -1.545135 -2.5091806
petal_length  2.655556  0.4561852
petal_width   3.010180 -2.3382506

Proportion of trace:
   LD1    LD2
0.9929 0.0071
```

Model LDA yang terbuat seperti gambar di bawah. Dapat kita lihat bahwa secara garis besar nilai group means pada masing-masing prediktor dan kelas target memiliki perbedaan yang cukup berarti, kecuali pada prediktor sepal width kelas 1 dan 2, yakni 2.74 dan 2.96, yakni berbeda 0.22, dengan perbedaan yang kecil ini ada kemungkinan besar misklasifikasi akan terjadi.

Bagian Proportion of trace juga menunjukan bahwa persamaan LD1 mampu menampung 0.9929 variansi, sedangkan LD2 hanya menampung 0.0071 variansi. Sehingga, akan lebih baik menggunakan LD1 sebagai persamaan prediksi.

Persamaan LD1: sepal_length * (-1.306494) + sepal_width * (-1.545135) + petal_length * 2.655556 + petal_width * 3.010180

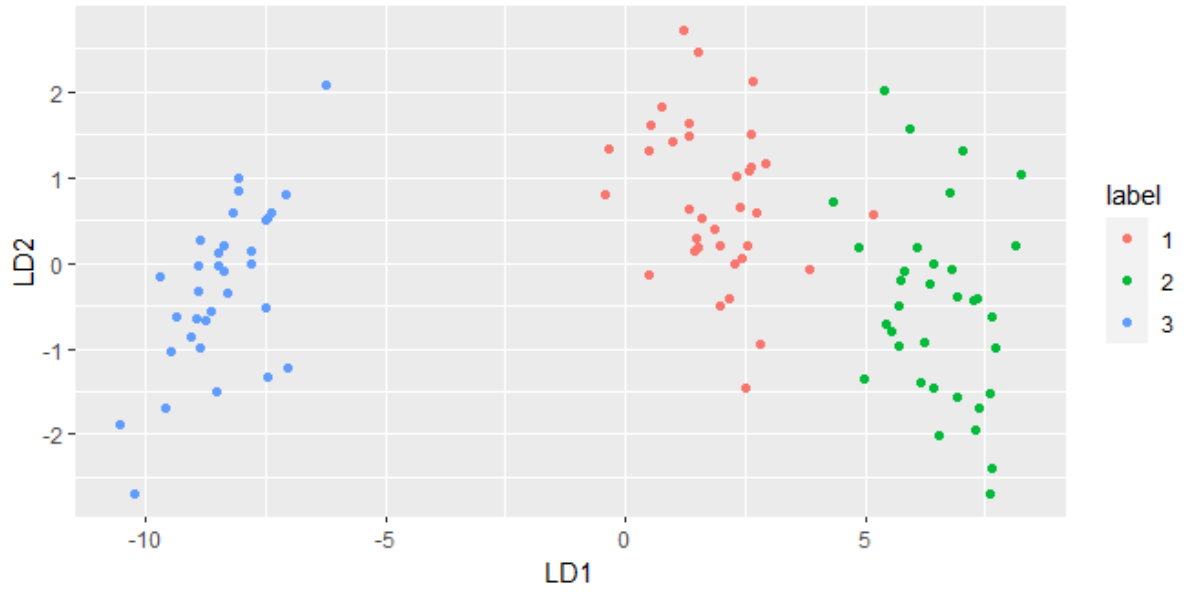
Dari pembuatan model didapatkan hasil confusion matrix, sebagai berikut:

```
          Truth
Prediction  1  2  3
         1 14  0  0
         2  1 15  0
         3  0  0 15
```

Adapun, nilai-nilai lain:

- Akurasi: 0.97778
- Presisi: 0.979
- Recall: 0.978

Hasil klasifikasi dari dataset dengan model LDA adalah sebagai berikut:



Keterangan label:

- 1: Versicolor
- 2: Virginica
- 3: Setosa

Berdasarkan hasil pengujian tersebut, maka dapat disimpulkan bahwa model LDA (parameter mixture=0 dan penalty = 0.0000000001) yang terbentuk telah sukses dalam melakukan klasifikasi pada dataset Iris dengan nilai akurasi 0.97778.

MULTIDIMENSIONAL SCALING (MDS)

Pertanyaan

Lakukan multidimensional scaling pada dataset ‘soal MDS’

Pengerjaan

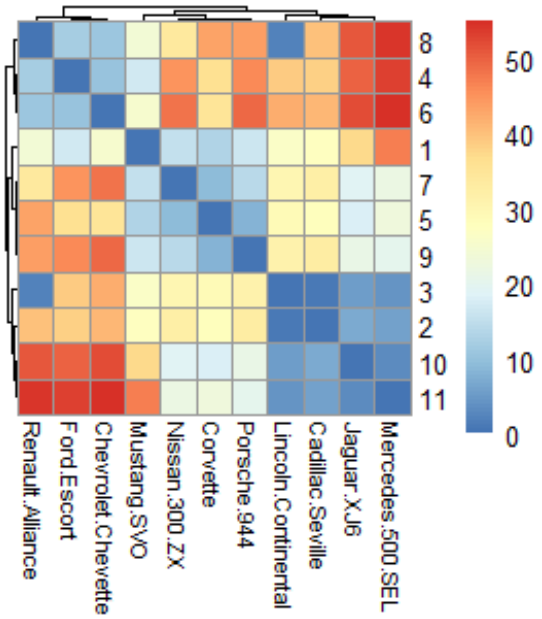
Beberapa langkah yang dilakukan adalah:

1. Import library yang diperlukan
2. Load dataset. Pembuatan dataframe dengan merubah _ menjadi NaN lalu, mengisi nilai pada index diagonal dengan 0 dan segitiga atas.

	...2	...3	...4	...5	...6	...7	...8	...9	...10	...11	...12
1	0	27	26	17	13	25	15	24	16	37	47
2	27	0	1	38	28	41	32	40	33	7	6
3	26	1	0	39	29	42	30	2	31	5	4
4	17	38	39	0	36	10	45	12	46	50	53
5	13	28	29	36	0	35	9	43	8	18	23
6	25	41	42	10	35	0	48	11	49	52	55
7	15	32	30	45	9	48	0	34	14	19	22
8	24	40	2	12	43	11	34	0	44	51	54
9	16	33	31	46	8	49	14	44	0	21	20
10	37	7	5	50	18	52	19	51	21	0	3
11	47	6	4	53	23	55	22	54	20	3	0

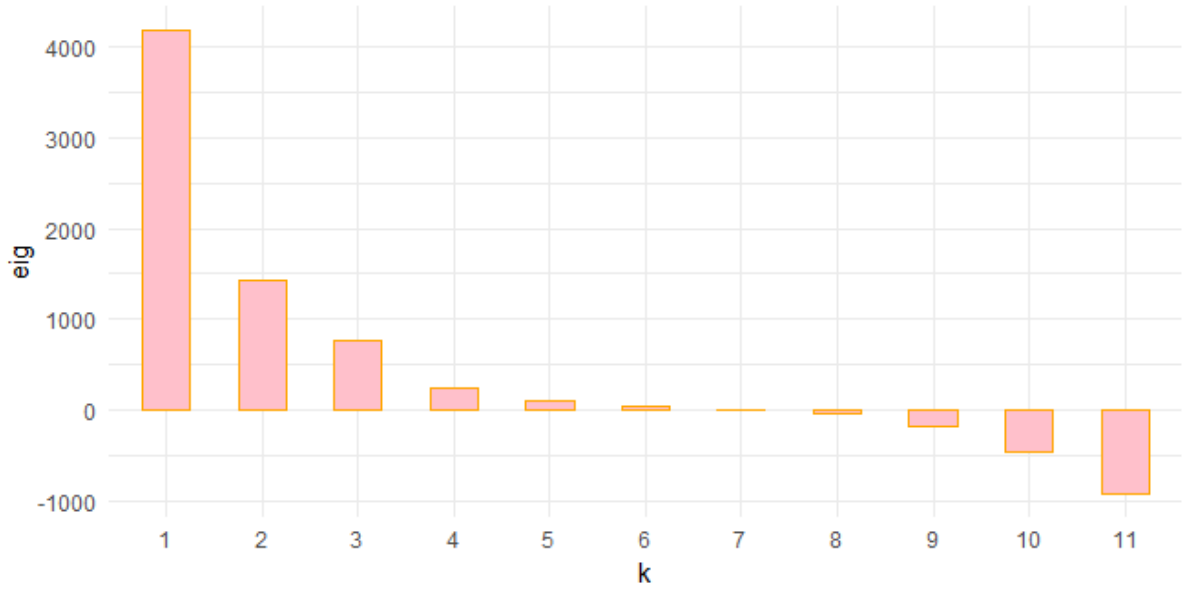
3. Pembuatan heatmap representasi lain dari distance matriks
4. Melakukan MDS dengan metric dimensional scaling
5. Bar plot menunjukan penyebaran variansi pada dimensi
6. Pembuatan scree plot

Analisis dan Kesimpulan

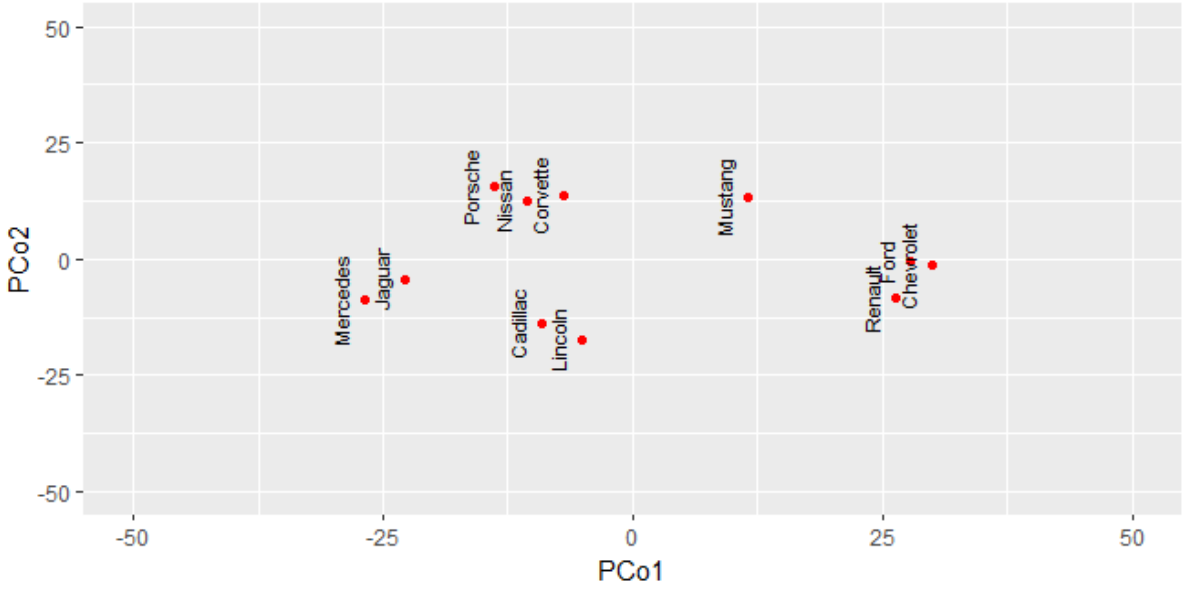


Pada heatmap diatas, dapat terlihat bahwa nilai jarak pada setiap kolom beragam. Warna yang semakin kebiruan menandakan bahwa jarak antara 2 variabel tersebut sangat dekat. Sedangkan, warna semakin kemerahan menandakan jarak semakin jauh. Dari heatmap tersebut sekilas dapat terbentuk 3 cluster, yakni dari persegi kiri atas, tengah, dan kanan bawah, dengan rincian berikut:

- Kiri atas: Renault Alliance, Ford Escort, dan Chevrolet Chevette
- Tengah: Mustang svo, nissan 300 ZX, corvette, dan Porsche 944
- Kanan bawah: Lincoln Cortinentral, Cadillac Seville, Jaguar XJ6, dan Mercedes 500 SEL



Pada barplot di atas dapat kita simpulkan bahwa penyebaran variansi berpusat pada variabel 1, 2, dan 3. Sehingga variabel yang cukup representatif adalah 3 variabel.



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

- Marcedes dan Jaguar
- Porsche, Nissan, dan Corvette
- Cadillac dan Lincoln
- Renault, Ford, dan Chevrolet

Titik yang memiliki karakteristik paling beda dari yang lainnya adalah mobil Mustang

Merek mobil yang berdekatan menandakan bahwa merek-merek mobil tersebut saling subtitut, sebagai contoh Marcedes dapat sebagai pilihan alternatif dari Jaguar, dan begitu sebaliknya. Hal ini berartian positif karena merek tersebut bisa menjadi alternatif dari produk lainnya.

Sedangkan, merek mobil Mustang tidak memiliki subtitusi. Kondisi Mustang ini dapat bernilai positif atau negatif, negatif dikarenakan tidak akan menjadi subtitu dari merek lainnya. Namun, dapat juga bernilai positif dengan artian tidak adanya kompetitor dalam kelas yang sama (memerlukan penelitian lebih lanjut dengan data tambahan terkait karakteristik mobil).