PREDICTIVE ANALYTICS AND MACHINE LEARNING

ANALISIS CLUSTERING MODEL PEMBERIAN BANTUAN PADA SUATU NEGARA BERDASARKAN FAKTOR SOSIAL, EKONOMI DAN KESEHATAN

Mochammad Reyhan Mauluddi (6032222003)

Dosen Mata Kuliah:

Shofi Andari, S.Stat., M.Si., Ph.D.

Departemen Manajemen Teknologi

Fakultas Sekolah Interdisiplin Manajemen dan Teknologi

Institut Teknologi Sepuluh Nopember

2024

I. LATAR BELAKANG

Organisasi non-pemerintah kemanusiaan internasional telah berhasil mengumpulkan sekitar $10 juta melalui program pendanaan saat ini. Saat ini permasalahannya adalah tugas CEO dari NGO untuk memutuskan cara terbaik dalam menggunakan dana ini secara strategis dan efektif. Untuk membuat keputusan yang tepat, CEO menyadari bahwa penting untuk mempertimbangkan negara-negara yang membutuhkan bantuan kemanusiaan dengan mendalam. Oleh karena itu, dibutuhkan analisis yang komprehensif untuk mengklasifikasikan negara-negara berdasarkan faktor-faktor sosial-ekonomi dan kesehatan yang mempengaruhi perkembangan negara secara keseluruhan.

Pendekatan yang dapat digunakan dalam memetakan pengelompokkan negara berdasarkan faktor-faktor tertentu adalah clustering. Clustering adalah metode yang digunakan untuk mengelompokkan data menjadi kelompok-kelompok yang serupa berdasarkan kemiripan fitur atau karakteristik yang dimiliki oleh setiap data. Dalam konteks ini, data yang akan dianalisis adalah faktor-faktor sosial-ekonomi dan kesehatan yang dapat digunakan sebagai indikator perkembangan negara.

Dengan menggunakan pendekatan clustering, akan diidentifikasi pola dan hubungan antara negara-negara berdasarkan faktor-faktor tersebut. Faktor-faktor sosial-ekonomi seperti pendapatan per kapita, tingkat pengangguran, dan indeks pembangunan manusia dapat menjadi fitur yang relevan. Selain itu, faktor kesehatan seperti angka harapan hidup, tingkat kematian bayi, dan akses terhadap layanan kesehatan juga akan menjadi pertimbangan penting dalam analisis.

Dengan mengelompokkan negara-negara berdasarkan kesamaan karakteristik sosial-ekonomi dan kesehatan, peneliti/analis dapat memberikan rekomendasi kepada CEO mengenai negara-negara yang perlu difokuskan dan diberi prioritas tertinggi. Misalnya, negara-negara yang termasuk dalam kelompok dengan tingkat perkembangan yang rendah atau mengalami krisis kesehatan mungkin membutuhkan bantuan mendesak dan harus menjadi prioritas utama dalam penggunaan dana tersebut.

Dengan pendekatan clustering, analisis data akan membantu CEO dalam pengambilan keputusan yang lebih terinformasi dan memastikan bahwa dana yang tersedia digunakan secara efektif dan secara strategis di negara-negara yang membutuhkan bantuan kemanusiaan. Klasifikasi negara-negara berdasarkan faktor-faktor sosial-ekonomi dan kesehatan akan memberikan wawasan yang lebih mendalam tentang kondisi masing-masing negara dan membantu dalam menentukan negara-negara yang paling membutuhkan bantuan serta harus diberikan prioritas tertinggi.

MDS adalah teknik statistik multivariat yang digunakan untuk menganalisis kesamaan atau perbedaan antara objek-objek berdasarkan matriks jarak atau kesamaan antara objek-objek tersebut. Dengan menggunakan MDS, CEO NGO akan

mendapatkan pemahaman visual tentang hubungan antara negara-negara berdasarkan faktor-faktor sosial-ekonomi dan kesehatan yang relevan. Hal ini akan membantu dalam mengidentifikasi negara-negara yang membutuhkan bantuan mendesak dan menentukan prioritas penggunaan dana secara efektif dan strategis.

II. RUMUSAN MASALAH

Penelitian ini diharapkan dapat menjawab permasalahan sebagai berikut:

1. Bagaimana membangun model clustering untuk melihat karakteristik dari masing-masing kelompok negara terbentuk.
2. Bagaimana pola visual hubungan antara negara-negara yang terbentuk berdasarkan Multidimensional Scaling?
3. Negara-negara prioritas mana yang perlu didahulukan berdasarkan hasil analisa clustering untuk diberi bantuan terlebih dahulu?

III. TUJUAN ANALISA

Tujuan dari penelitian ini adalah sebagai berikut:

1. Untuk membuat model dalam menelaah lebih jauh karakteristik dari masing-masing kelompok negara yang terbentuk.
2. Untuk membuat visualisasi pola dan hubungan antara masing-masing kelompok negara yang terbentuk.
3. Untuk membantu CEO NGO dalam memilih negara prioritas dalam pemberian bantuan berdasarkan hasil analisa clustering dan Multidimensional Scaling.

IV. PENGOLAHAN DAN ANALISA DATA

Data yang digunakan dalam penelitian ini adalah data negara-negara dengan variabel observasi yang berbeda. Total data sebanyak 167 pengamatan dengan detail data yang termuat dijabarkan sebagai berikut:

a. Nama masing-masing negara (country).
b. X1 (child_mort) = Kematian anak di bawah usia lima tahun per 1000 kelahiran.
c. X2 (exports) = Tingkat ekspor barang dan jasa. Ekspor barang dan jasa dinyatakan sebagai persentase dari Total Produk Domestic Bruto.
d. X3 (health) = Tingkat kesehatan
e. X4 (import) = Tingkat impor barang dan jasa. Dinyatakan sebagai persentase dari Total Produk Domestic Bruto.
f. X5 (income_std) = Pendapatan bersih per orang (nilai sudah distandarkan).
g. X6 (inflation) = Tingkat inflasi.
h. X7 (life_expect) = Usia harapan hidup
i. X8 (total_fertility) = Total kelahiran.

j. X9 (gdpp_std) = Gross Domestic Product/ Produk Domestik Bruto (nilai sudah distandarkan).

Mengubah nama kolom atau variable agar mudah pada pembacaan grafik

```
> df <- df %>%
+   rename(
+     x1 = child_mort,
+     x2 = exports,
+     x3 = health,
+     x4 = imports,
+     x5 = `income std`, # karena ada spasi, gunakan backticks
+     x6 = inflation,
+     x7 = life_expect,
+     x8 = total_fertility,
+     x9 = `gdpp std` # karena ada spasi, gunakan backticks
+   )
```

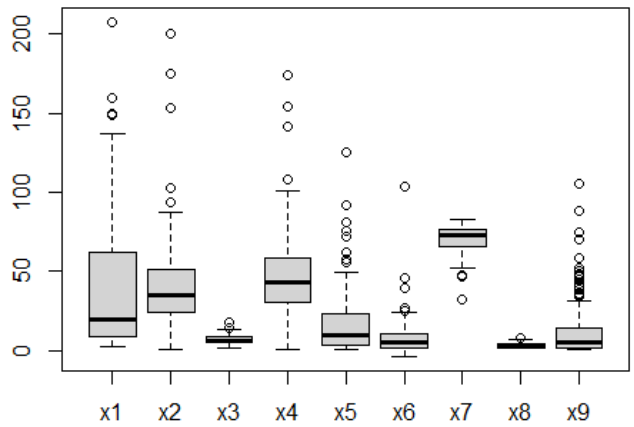
Menampilkan 5 observasi pertama dari dataframe (df)

```
> head(df, 5)
# A tibble: 5 x 11
     No country                x1    x2    x3    x4    x5    x6    x7    x8    x9
  <dbl> <chr>                <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1     1 Afghanistan           90.2  10    7.58  44.9  1.61  9.44  56.2  5.82  0.553
2     2 Albania               16.6  28    6.55  48.6  9.93  4.49  76.3  1.65  4.09
3     3 Algeria               27.3  38.4  4.17  31.4 12.9  16.1   76.5  2.89  4.46
4     4 Angola               119    62.3  2.85  42.9  5.9  22.4   60.1  6.16  3.53
5     5 Antigua and Barbuda   10.3  45.5  6.03  58.9 19.1   1.44  76.8  2.13 12.2
```

Data Exploration: Ukuran Pemusatan Data, Box Plot dan Scatter Plot

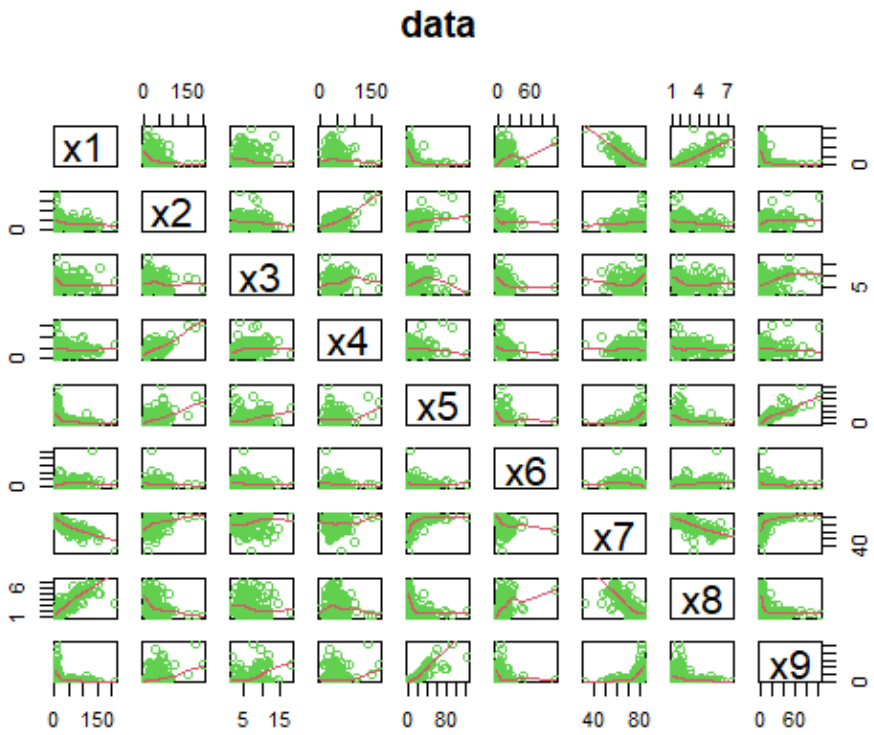
```
> summary(df)
       No          country                x1               x2               x3
 Min.   :  1.0   Length:167         Min.   :  2.60   Min.   :  0.109   Min.   : 1.810
 1st Qu.: 42.5   Class :character   1st Qu.:  8.25   1st Qu.: 23.800   1st Qu.: 4.920
 Median : 84.0   Mode  :character   Median : 19.30   Median : 35.000   Median : 6.320
 Mean   : 84.0                      Mean   : 38.27   Mean   : 41.109   Mean   : 6.816
 3rd Qu.:125.5                      3rd Qu.: 62.10   3rd Qu.: 51.350   3rd Qu.: 8.600
 Max.   :167.0                      Max.   :208.00   Max.   :200.000   Max.   :17.900
       x4                x5                x6                x7               x8
 Min.   :  0.0659   Min.   :  0.609   Min.   : -4.210   Min.   :32.10   Min.   :1.150
 1st Qu.: 30.2000   1st Qu.:  3.355   1st Qu.:  1.810   1st Qu.:65.30   1st Qu.:1.795
 Median : 43.3000   Median :  9.960   Median :  5.390   Median :73.10   Median :2.410
 Mean   : 46.8902   Mean   : 17.145   Mean   :  7.782   Mean   :70.56   Mean   :2.948
 3rd Qu.: 58.7500   3rd Qu.: 22.800   3rd Qu.: 10.750   3rd Qu.:76.80   3rd Qu.:3.880
 Max.   :174.0000   Max.   :125.000   Max.   :104.000   Max.   :82.80   Max.   :7.490
```

```
> df_subset <- df[, !(names(df) %in% c("country"))]
> boxplot(df_subset)
```



Dari boxplot terlihat ukuran numerik data dengan rentang 0.2 hingga 208. Masing-masing variable memiliki pemusatan data berbeda namun dengan rentang yang dekat. Hal ini menunjukkan data sudah memiliki standarisasi yang baik.

```
pairs(df_subset, panel = panel.smooth, main = "data", # fancy scatterplots
      col = 3)
```



Grafik di atas adalah grafik dua dimensi berpasangan dari x1 hingga x9

Missing Value

```
> sum(is.na(df))
[1] 0
```

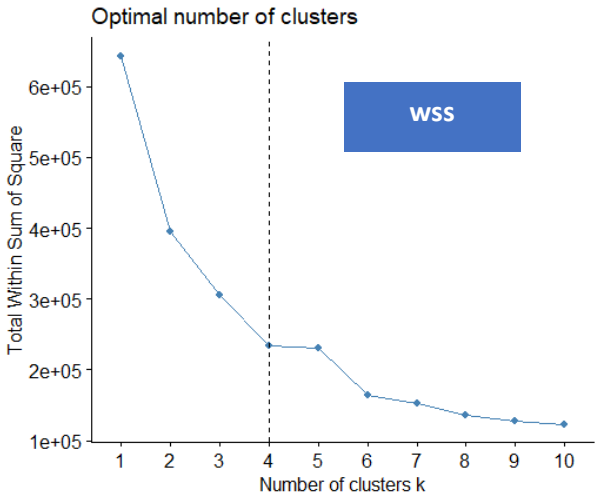
Dari output di samping terlihat tidak ada missing value atau semua observasi memiliki nilai yang tidak kosong

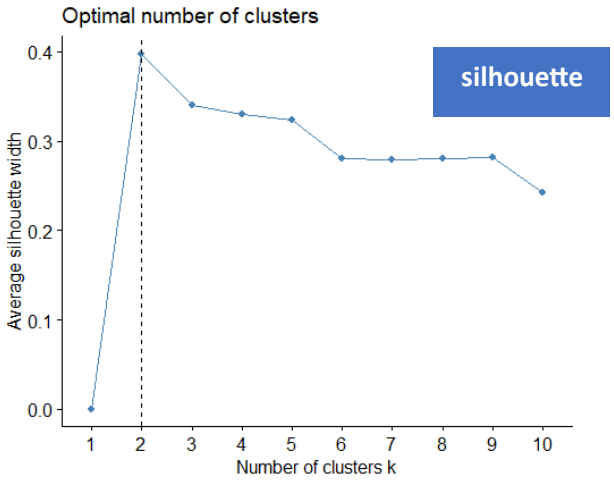
Analisis Clustering

Untuk menghasilkan cluster yang tepat maka dari 167 negara perlu disederhanakan pengelompokkannya berdasarkan dimensi faktor ekonomi, sosial dan kesehatan. Untuk itu jumlah cluster yang terbentuk sebaiknya sederhana. Selain itu karena aspek efisiensi agar mudah dipahami dan diimplementasikan pada kasus dan skalibilitas terhadap ukuran data maka dipilih k-means cluster.

Menentukan Jumlah Cluster Optimal

```
> fviz_nbclust(df_subset, FUNcluster = kmeans, method = "wss", k.max = 10)+
+   geom_vline(xintercept = 4, linetype = 2)

> fviz_nbclust(df_subset, FUNcluster = kmeans, method = "silhouette", k.max = 10)+
+   geom_vline(xintercept = 2, linetype = 2)
```





Pemilihan jumlah cluster yang terbentuk berdasarkan metode wss adalah 4 dan silhouette adalah 2. Untuk memberikan informasi yang lebih lengkap terhadap hasil clustering maka digunakan variasi dari cluster yang lebih tinggi sehingga dipilih jumlah cluster ideal terbentuk adalah berdasarkan Within-Cluster Sum of Squares sebanyak 4 cluster.

Melakukan Clustering dengan K-Means

```
> df_km <- kmeans(df_subset, 4, nstart = 25)
> print(df_km)
```

```
K-means clustering with 4 clusters of sizes 34, 87, 43, 3

Cluster means:
          x1        x2       x3        x4       x5        x6       x7       x8        x9
1  5.529412  47.81176 8.675588  40.59118 44.63529  3.561059 79.88235 1.864412 40.561765
2 23.190805  41.30207 6.398391  48.58391 11.68011  7.699667 72.10115 2.386667  6.297632
3 97.048837  26.00742 6.190930  40.78525  3.19286 11.656163 59.29535 5.049767  1.519070
4  4.133333 176.00000 6.793333 156.66667 64.03333  2.468000 81.43333 1.380000 57.566667

Clustering vector:
  [1] 3 2 2 3 2 2 2 1 1 2 2 1 2 2 2 1 2 3 2 2 2 2 2 1 2 3 3 2 3 1 2 3 3 2 2 2 3 3 2 2 3 2 1 1 1
 [46] 2 2 2 2 3 3 2 2 1 1 3 3 2 1 3 1 2 2 3 3 2 3 2 1 3 2 2 2 1 1 1 2 1 2 2 3 3 1 2 3 2 2 3 3 2
 [91] 2 4 2 3 3 2 2 3 4 3 2 2 2 2 2 2 3 3 2 2 1 1 3 3 1 1 3 2 2 2 2 2 1 1 2 2 3 2 1 3 2 2 3 4 2
[136] 1 2 2 1 1 2 2 3 2 1 1 2 3 2 3 3 2 2 2 2 3 2 1 1 1 2 2 2 2 2 2 3

Within cluster sum of squares by cluster:
[1] 46364.43 95626.13 84412.74  7471.68
 (between_SS / total_SS =  63.6 %)
```

Analisa hasil K-Means Clustering:

- Dalam pengelompokan K-means ini, data telah dikelompokkan menjadi empat cluster dengan ukuran masing-masing cluster adalah cluster pertama dengan 34 negara, cluster kedua dengan 87 negara, cluster ketiga dengan 43 negara, dan cluster keempat dengan 3 negara.
- Berdasarkan pengelompokan k-means cluster yang telah dilakukan, interpretasi dari pengertian variabel yang diberikan adalah sebagai berikut:

- Cluster 1: Kematian anak di bawah usia lima tahun (child_mort) relatif rendah, pendapatan bersih per orang (income_std) cenderung tinggi, tingkat kesehatan (health) dan usia harapan hidup (life_expect) relatif tinggi, inflasi (inflation) rendah, dan GDP per kapita (gdpp_std) tinggi. Cluster ini mencerminkan negara-negara yang memiliki tingkat kesejahteraan dan kesehatan yang baik serta kestabilan ekonomi. Cluster ini beranggotakan negara seperti: Australia, Austria, Bahrain, Belgium, Denmark, France, Germany, Jepang, dsb.
- Cluster 2: Kematian anak di bawah usia lima tahun (child_mort) sedang, tingkat ekspor (exports) dan impor (imports) cenderung tinggi, pendapatan bersih per orang (income_std) rendah, inflasi (inflation) sedang, dan GDP per kapita (gdpp_std) sedang. Ini mungkin mencerminkan negara-negara yang memiliki tingkat perdagangan internasional yang tinggi tetapi tingkat kesejahteraan masyarakat yang lebih rendah. Cluster ini beranggotakan negara seperti: Albania, Argentina, Brazil, Chile, Indonesia, Iran, Lebanon, dsb.
- Cluster 3: Kematian anak di bawah usia lima tahun (child_mort) cenderung tinggi, tingkat kesehatan (health) rendah, dan GDP per kapita (gdpp_std) rendah. Ini mungkin mencerminkan negara-negara yang mengalami masalah kesehatan masyarakat dan kemiskinan. Cluster ini beranggotakan negara seperti: Afghanistan, Angola, Congo, Kenya, Liberia, Myanmar, Timor Leste, dsb.
- Cluster 4: Dalam cluster ini, hanya ada 3 observasi. Karena ukuran cluster yang sangat kecil, sulit memberikan interpretasi yang akurat. Namun, berdasarkan nilai-nilai rata-rata variabel, terdapat kecenderungan tingkat exports (exports) tinggi, tingkat import (imports) tinggi, usia harapan hidup (life expect) tinggi, total kelahiran (total fertility) rendah. Hal ini mungkin mencerminkan negara-negara yang aktif pada perdagangan internasional, yang masyarakatnya memiliki usia harapan hidup tinggi namun memiliki tantangan dalam total kelahiran yang rendah. Cluster ini beranggotakan negara: Luxemburg, Malta, Singapore.
- Cluster 1 memiliki WSS sebesar 46.364,43, Cluster 2 memiliki WSS sebesar 95.626,1, Cluster 3 memiliki WSS sebesar 84.412,74, Cluster 4 memiliki WSS sebesar 7.471,68. Semakin rendah nilai WSS, semakin padat dan homogen cluster tersebut. Artinya, titik-titik dalam cluster tersebut cenderung berada lebih dekat satu sama lain dan lebih dekat dengan centroidnya. Semakin tinggi nilai WSS, semakin tidak homogen cluster tersebut. Ini bisa disebabkan oleh titik-titik yang tersebar luas di sekitar centroid atau memiliki jarak yang lebih jauh dari centroidnya. Cluster 4 memiliki WSS yang paling rendah, menunjukkan bahwa cluster ini mungkin lebih padat dan homogen dibandingkan dengan cluster lainnya. Cluster 2 memiliki WSS yang tertinggi, menunjukkan bahwa cluster ini mungkin lebih tersebar atau tidak homogen dibandingkan dengan yang lain.
- Nilai antara_SS / total_SS sebesar 63.6% menunjukkan seberapa besar variabilitas antara cluster dibandingkan dengan total variabilitas dalam dataset. Nilai ini

menunjukkan seberapa baik cluster-cluster tersebut memisahkan data dan seberapa baik model k-means menangkap struktur dalam data. Semakin tinggi nilai ini, semakin baik model k-means memisahkan cluster. Dalam hal ini, nilai 63.6% menunjukkan bahwa sebagian besar variabilitas dalam data berhasil dijelaskan oleh pembagian menjadi cluster-cluster yang dibuat oleh model k-means.

Analisis Multidimensional Scaling

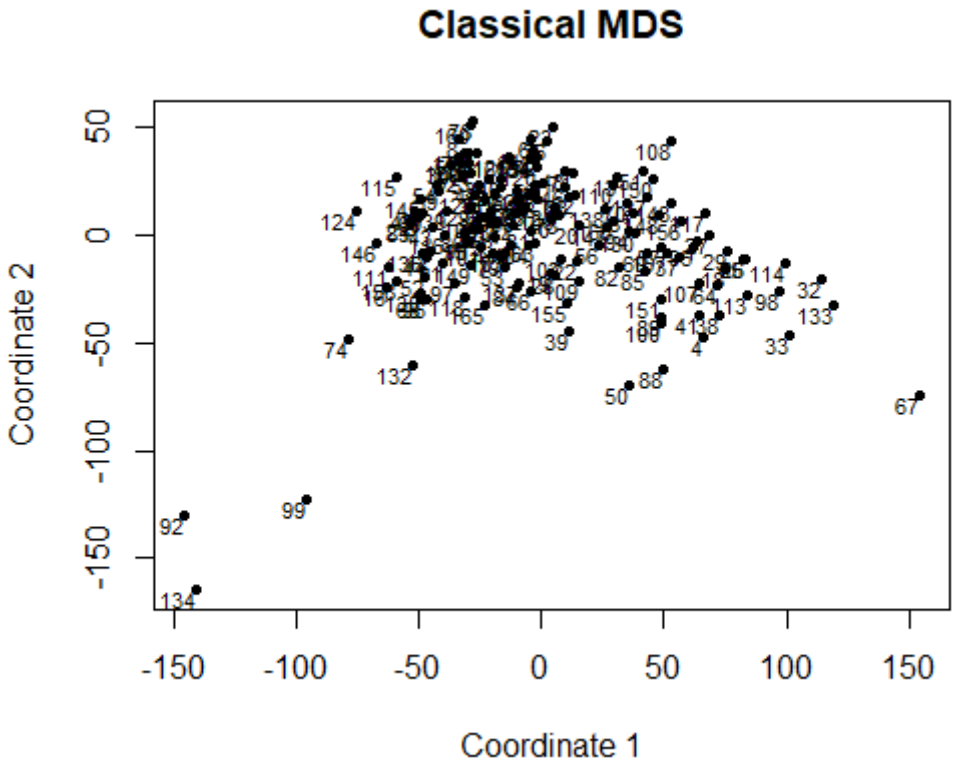
```
> # Melakukan analisa MDS
> df_subset_dis <- dist(df_subset) # mendapatkan jarak matriks
> fitdis <- cmdscale(df_subset_dis,eig=T, k=2) # k is the number of dim
> fitdis

$points
               [,1]         [,2]
  [1,]   63.6350695   -3.1965621
  [2,]   -9.3506292   17.7007173
  [3,]   -1.4290201   17.0494308
  [4,]   65.6894945  -47.6284624
  [5,]  -28.0922316    2.7389309
  [6,]   -4.1283247   44.5200153
  [7,]   -2.8763463   23.4475823
  [8,]  -33.3676992   44.7780630
  [9,]  -49.8998486    8.1744538
 [10,]    3.9459793    7.6447792
```

Sintaks di atas menjelaskan pembuatan sebuah matriks jarak (df_subset_dis) dari dataset df_subset. Matriks jarak ini mengukur jarak antara setiap pasangan observasi dalam dataset. Kemudian, fungsi cmdscale() digunakan untuk melakukan fitting MDS pada matriks jarak (df_subset_dis). Parameter eig=T digunakan untuk mendapatkan nilai eigen (eigenvalues) dari solusi MDS, dan k=2 digunakan untuk menentukan bahwa kita ingin mereduksi dimensi menjadi dua dimensi. Hasil fitting disimpan dalam fitdis.

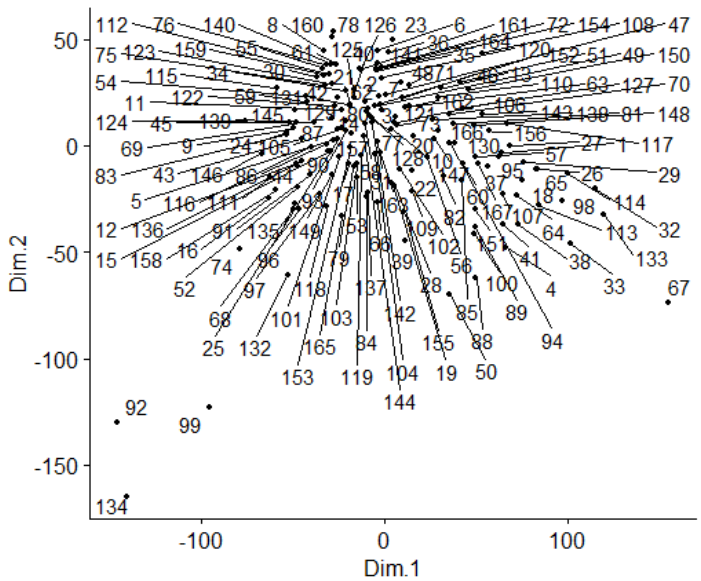
```
> # plot solution
> x <- fitdis$points[,1]
> y <- fitdis$points[,2]
> plot(x, y, xlab="Coordinate 1", ylab="Coordinate 2",
+      main="Classical MDS", pch = 20)
> text(x, y, labels = row.names(df), cex=.7)
```

Selanjutnya, hasil dari fitting MDS (fitdis) digunakan untuk mengambil koordinat dari setiap observasi dalam ruang dua dimensi. Koordinat ini disimpan dalam variabel x dan y. Terakhir, hasil koordinat dari setiap observasi direpresentasikan dalam sebuah

plot dua dimensi. Label dari setiap observasi ditambahkan ke plot tersebut sesuai dataframe (df).

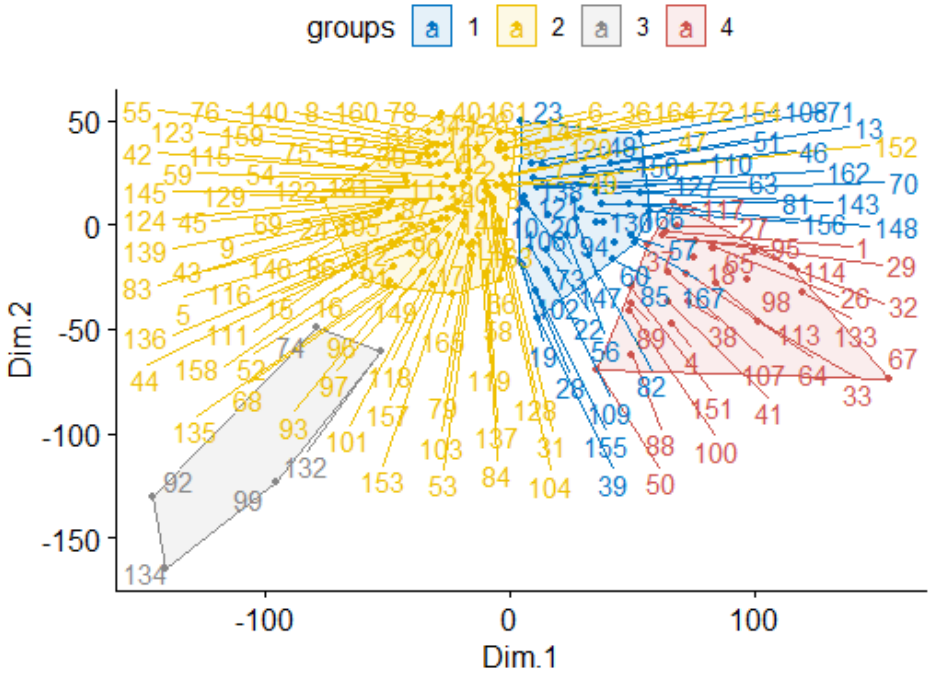



```
> # Compute MDS
> mds <- df_subset %>%
+   dist() %>%      # obtain distance matrix
+   cmdscale() %>%
+   as_tibble()
> colnames(mds) <- c("Dim.1", "Dim.2")
> # Plot MDS
> ggscatter(mds, x = "Dim.1", y = "Dim.2",
+           label = rownames(df),
+           size = 1,
+           repel = TRUE)
```



Sintaks di atas menghitung MDS dari df_subset yang menghasilkan plot scatter dua dimensi dengan menggunakan ggplot2, dan menambahkan label yang mewakili baris dari dataset awal df. Dalam menginterpretasikan gambar di atas titik-titik yang lebih dekat dalam plot menunjukkan data yang lebih mirip atau serupa dalam struktur atau atributnya seperti negara 78 (Jepang) dan160 (United States) yang memilik cluster sejenis yaitu cluster 1.

```
> # K-means clustering
> clust <- kmeans(mds, 4)$cluster %>%
+   as.factor()
> mds_cl <- mds %>%
+   mutate(groups = clust)
> # Plot and color by groups
> ggscatter(mds_cl, x = "Dim.1", y = "Dim.2",
+           label = rownames(df),
+           color = "groups",
+           palette = "jco",
+           size = 1,
+           ellipse = TRUE,
+           ellipse.type = "convex",
+           repel = TRUE)
```



Sintaks di atas menghasilkan plot yang menunjukkan titik-titik data yang dikelompokkan ke dalam 4 klaster menggunakan metode K-Means, dengan masing-masing cluster ditampilkan dalam warna yang berbeda. Penggunaan ellips yang mengelilingi tiap kelompok membantu dalam memvisualisasikan pola atau struktur cluster tersebut. Dalam menginterpretasikan gambar di atas, menunjukkan hasil cluster yang mirip dengan sebelumnya. Perbedaan berada pada cluster 4 sebelumnya

yang beranggotakan Luxemburg, Malta dan Singapore, pada clustering DMS bertambah negara Irlandia dan Seychelles. Sedangkan cluster lain seperti Jepang dan United States, kemudian Nigeria dengan Guinea, Nepal dengan Dominican Republic tetap berada pada cluster yang sama. Hal ini menunjukkan clustering dengan DMS bisa memperbaiki atau meningkatkan hasil clustering k-means sebelumnya tanpa menggunakan DMS.

V. KESIMPULAN

1. Model clustering yang terbentuk dengan menggunakan K-Means menghasilkan sebanyak empat cluster dengan karakteristik masing-masing cluster, mulai dari cluster 1 mencerminkan negara-negara yang memiliki tingkat kesejahteraan dan kesehatan yang baik serta kestabilan ekonomi, cluster 2 mencerminkan negara-negara yang memiliki tingkat perdagangan internasional yang tinggi tetapi tingkat kesejahteraan masyarakat yang lebih rendah, cluster 3 mencerminkan negara-negara yang mengalami masalah kesehatan masyarakat dan kemiskinan, dan cluster 4 mencerminkan negara-negara yang aktif pada perdagangan internasional, yang masyarakatnya memiliki usia harapan hidup tinggi namun memiliki tantangan dalam total kelahiran yang rendah. Dengan masing-masing cluster berjumlah cluster 1 sebanyak 34 negara, cluster 2 sebanyak 87 negara, cluster 3 sebanyak 43 negara dan cluster 4 sebanyak 3 negara.
2. Hasil dari clustering dengan Multidimensional Scaling memperbaiki dan memperjelas kedekatan antara masing-masing kelompok negara. Pada cluster 4 yang sebelumnya beranggotakan Luxemburg, Malta dan Singapore, menurut hasil clustering MDS bertambahkan anggota dari negara Irlandia dan Seychelles. Pada cluster lainnya anggota negara yang masuk cenderung sama.
3. CEO NGO bisa memfokuskan prioritas bantuan utama kepada negara-negara yang masuk ke dalam cluster 3 terlebih dahulu karena negara pada cluster ini memiliki kondisi kesehatan masyarakat dan kemiskinan yang buruk. Lalu, bisa memprioritaskan negara pada cluster 2 karena negara pada cluster ini tingkat kesejahteraan masyarakatnya masih rendah dari rata-rata negara maju. Kemudian, mengikuti cluster 4 dan cluster 1.