Klasterisasi Negara Menggunakan K-Means

Sayid Ghufron

Permasalahan

HELP International telah berhasil mengumpulkan sekitar $ 10 juta. Saat ini, CEO LSM perlu memutuskan bagaimana menggunakan uang ini secara strategis dan efektif. Jadi, CEO harus mengambil keputusan untuk memilih negara yang paling membutuhkan bantuan. Oleh karena itu, Tugas teman-teman adalah mengkategorikan negara menggunakan beberapa faktor sosial ekonomi dan kesehatan yang menentukan perkembangan negara secara keseluruhan. Kemudian kalian perlu menyarankan negara mana saja yang paling perlu menjadi fokus CEO.

Metode Penyelesaian

Metode yang digunakan untuk memecahkan permasalahan diatas ialah dengan menggunakan algoritma Unsupervised Learning yaitu K-Means dengan dataset yang dimiliki oleh HELP International.

Pada dataset yang dimiliki oleh HELP International terdapat kolom-kolom dengan penjelasan sebagai berikut :

- Negara : Nama negara
- Kematian_anak: Kematian anak di bawah usia 5 tahun per 1000 kelahiran
- Ekspor : Ekspor barang dan jasa perkapita
- Kesehatan: Total pengeluaran kesehatan perkapita
- Impor: Impor barang dan jasa perkapita
- Pendapatan: Penghasilan bersih perorang
- Inflasi: Pengukuran tingkat pertumbuhan tahunan dari Total GDP
- Harapan_hidup: Jumlah tahun rata-rata seorang anak yang baru lahir akan hidup jika pola kematian saat ini tetap sama
- Jumlah_fertiliti: Jumlah anak yang akan lahir dari setiap wanita jika tingkat kesuburan usia saat ini tetap sama
- GDPperkapita: GDP per kapita. Dihitung sebagai Total GDP dibagi dengan total populasi.

Untuk library yang saya gunakan antara lain adalah :

- Pandas
- Sklearn
- Seaborn
- Numpy
- Matplotlib

Untuk kolom yang saya gunakan adalah kolom GDPperkapita dan kolom Pendapatan.

1. Melakukan pembacaan dataset, lalu melakukan data menggunakan rumus Normalisasi dan juga pembersihan data yang kosong.

$$Xnorm = \frac{(X - Xmin)}{(Xmax - Xmin)}$$

Rumus Normalisasi

```
Kematian_anak        0
Ekspor               0
Kesehatan            0
Impor                0
Pendapatan           0
Inflasi              0
Harapan_hidup        0
Jumlah_fertiliti     0
GDPperkapita         0
dtype: int64
```

Tidak terdapat data yang kosong

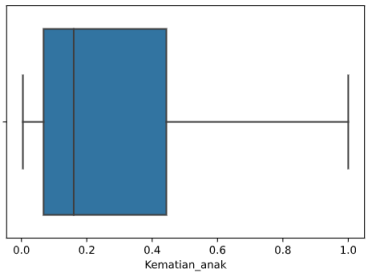
	Kematian_anak	Ekspor	Kesehatan	Impor	Pendapatan	Inflasi	Harapan_hidup	Jumlah_fertiliti	GDPperkapita
0	0.651786	0.113963	0.511072	0.444192	0.020516	0.490831	0.136364	0.858456	0.010534
1	0.104167	0.321358	0.419841	0.480849	0.191039	0.312837	0.788961	0.091912	0.126239
2	0.183780	0.441186	0.209035	0.310441	0.251911	0.730313	0.795455	0.319853	0.138343
3	0.866071	0.716560	0.092117	0.424377	0.108442	0.956850	0.262987	0.920956	0.107920
4	0.057292	0.522992	0.373782	0.582896	0.378984	0.203164	0.805195	0.180147	0.391540

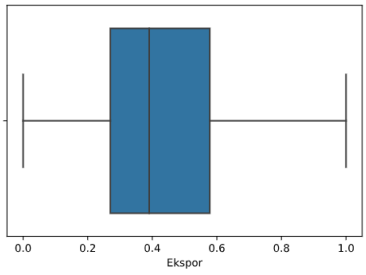
Hasil data yang telah dinormalisasi

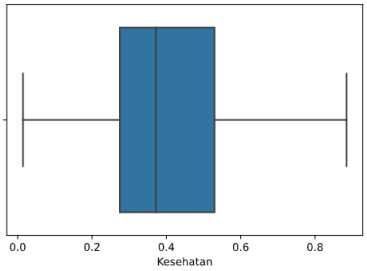
2. Lalu dilakukan pembersihan data Outlier menggunakan rumus *Interquartile Range.*

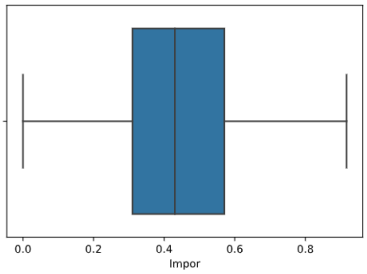
$$IQR = Q3 - Q1$$

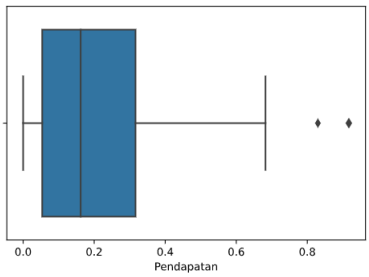
Setelah dilakukan pembersihan data Outlier menggunakan rumus *Interquartile Range.*

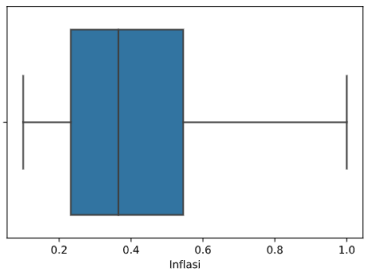


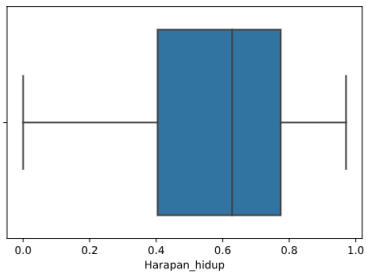


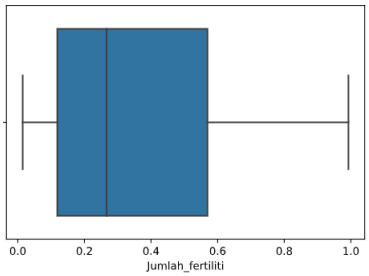


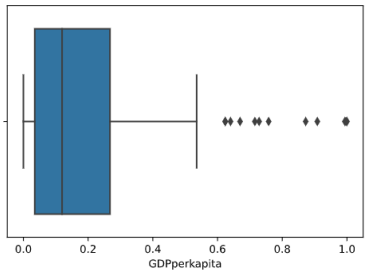




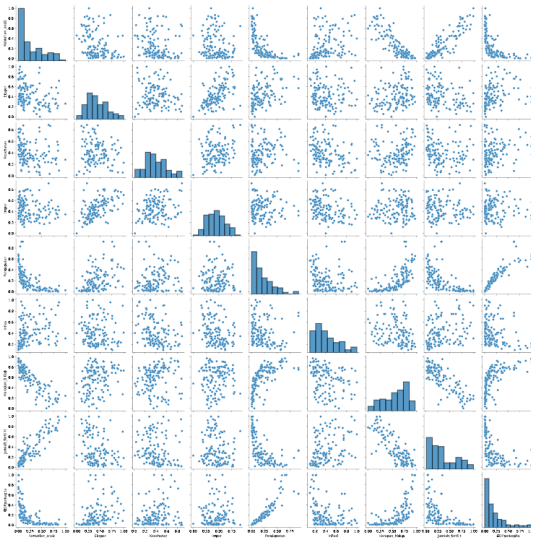




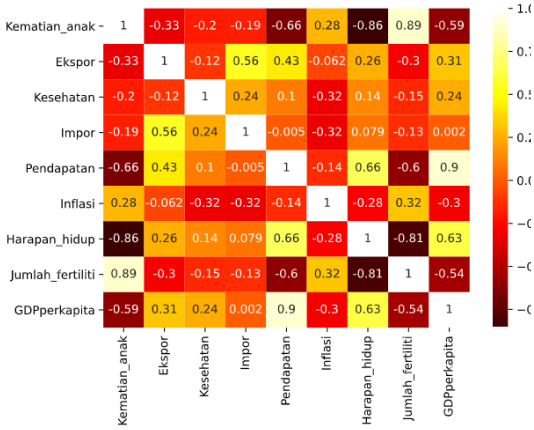




Boxplot

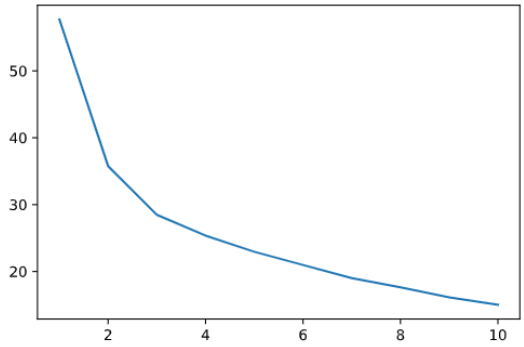


Pairplot

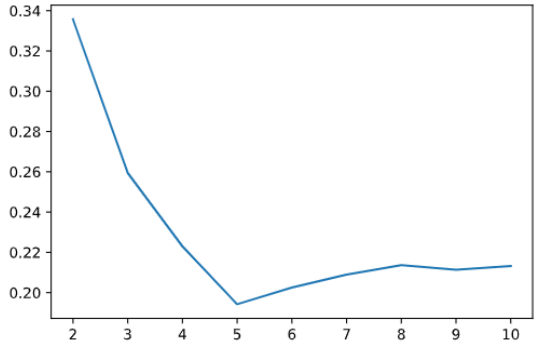


Heatmap

3. Setelah dilakukan normalisasi dan juga handling outlier, lalu dilakukan pencarian nilai elbow dan juga silhouette dengan rentang nilai K = 2, 3, 4, 5, 6, 7, 8, 9, 10



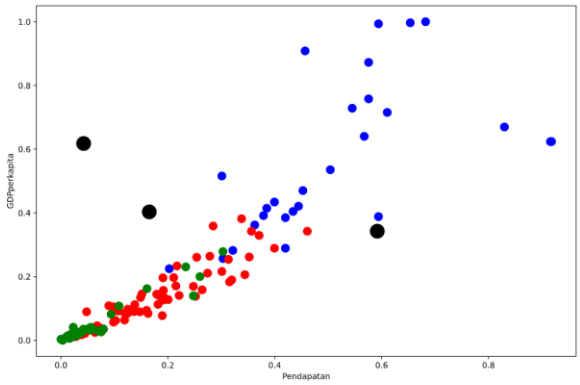
Elbow Method



Silhouette Score

Dari kedua graph diatas, saya memutuskan K optimal yaitu K = 3, dengan random state = 42.

4. Lalu dilakukanlah Clusterisasi dengan K = 3, serta data dari kolom Pendapatan dan kolom GDPperkapita.



Klasterisasi data Pendapatan terhadap GDPperkapita.

5. Berdasarkan hasil klusterisasi diatas didapatkan negara-negara mana saja yang layak diberi bantuan oleh HELP International. Terpilihlah negara-negara sebagai berikut.

	Negara
0	Burundi
1	Liberia
2	Congo, Dem. Rep.
3	Madagascar
4	Mozambique
5	Malawi
6	Eritrea
7	Togo
8	Guinea-Bissau
9	Afghanistan