FINAL PROJECT

PYTHON DATA SCIENCE SANBERCODE - BATCH 25

CLUSTERING THE COUNTRIES USING K-MEANS ALGORITHM FOR HELP INTERNATIONAL

JULY 2021

NADIA RIZKY HAIRUNNISA

HELP

LATAR BELAKANG

HELP International adalah LSM kemanusiaan internasional yang berkomitmen untuk memerangi kemiskinan dan menyediakan fasilitas dan bantuan dasar bagi masyarakat di negara-negara terbelakang saat terjadi bencana dan bencana alam. 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, diperlukan pengkategorian negara menggunakan beberapa faktor sosial ekonomi dan kesehatan yang menentukan perkembangan negara secara keseluruhan. Hasil akhir proyek ini adalah saran negara mana saja yang akan difokuskan untuk menerima bantuan.

TUJUAN

Tujuan *project* ini adalah untuk mengkategorikan berbagai negara berdasarkan faktor sosial, ekonomi, dan kesehatan yang menentukan pembangunan negara secara keseluruhan agar keputusan pemberian dana bersifat objektif dan berbasis data.

READING AND UNDERSTANDING DATA

Data yang digunakan merupakan data negara beserta faktor-faktor yang akan menentukan tingkat kelayakan negara tersebut untuk mendapatkan bantuan dana dari LSM Help International. Faktor-faktor tersebut meliputi tingkat kematian anak, ekspor, kesehatan, impor, pendapatan, inflasi, harapan hidup, jumlah fertility, dan GDP per kapita. Berikut lima data teratas dari dataset yang telah dikumpulkan oleh LSM Help International.

	Negara	Kematian_anak	Ekspor	Kesehatan	Impor	Pendapatan	Inflasi	Harapan_hidup	Jumlah_fertiliti	GDPperkapita
0	Afghanistan	90.2	10.0	7.58	44.9	1610	9.44	56.2	5.82	553
1	Albania	16.6	28.0	6.55	48.6	9930	4.49	76.3	1.65	4090
2	Algeria	27.3	38.4	4.17	31.4	12900	16.10	76.5	2.89	4460
3	Angola	119.0	62.3	2.85	42.9	5900	22.40	60.1	6.16	3530
4	Antigua and Barbuda	10.3	45.5	6.03	58.9	19100	1.44	76.8	2.13	12200

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 167 entries, 0 to 166
Data columns (total 10 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   Negara            167 non-null    object
 1   Kematian_anak     167 non-null    float64
 2   Ekspor            167 non-null    float64
 3   Kesehatan         167 non-null    float64
 4   Impor             167 non-null    float64
 5   Pendapatan        167 non-null    int64
 6   Inflasi           167 non-null    float64
 7   Harapan_hidup     167 non-null    float64
 8   Jumlah_fertiliti  167 non-null    float64
 9   GDPperkapita      167 non-null    int64
dtypes: float64(7), int64(2), object(1)
memory usage: 13.2+ KB
```

Dataset memiliki 10 kolom dan 167 baris, satu kolom bertipe data *object* berupa nama negara, sembilan kolom lainnya merupakan data numerik dengan tipe data *integer* dan *float*. Berdasarkan informasi data, setiap kolom memiliki 167 data non-null sehingga tidak ada *missing value* pada dataset. Format dataset berupa CSV dan berukuran sekitar 13.2 KB. Berikut deskripsi singkat dari kolom pada dataset.

Index	Fitur	Deskripsi
0.	Negara	Nama negara
1.	Kematian_anak	Kematian anak di bawah usia 5 tahun per 1000 kelahiran
2.	Ekspor	Ekspor barang dan jasa per kapita dalam per 1000 USD
3.	Kesehatan	Total pengeluaran per kapita pada sektor kesehatan dalam per 1000 USD
4.	Impor	Impor barang dan jasa per kapita dalam per 1000 USD
5.	Pendapatan	Penghasilan bersih rata-rata per individu dalam USD
6.	Inflasi	Pengukuran tingkat pertumbuhan tahunan dari total GDP
7.	Harapan_hidup	Jumlah tahun rata-rata seorang anak akan hidup jika pola kematian saat ini tetap sama
8.	Jumlah_feritliti	Jumlah anak yang akan lahir per wanita jika tingkat kesuburan usia saat ini tetap sama
9.	GDPperkapita	GDP per kapita, dihutung sebagai total GDP dibagi dengan total populasi

	Kematian_anak	Ekspor	Kesehatan	Impor	Pendapatan	Inflasi	Harapan_hidup	Jumlah_fertiliti	GDPperkapita
count	167.000000	167.000000	167.000000	167.000000	167.000000	167.000000	167.000000	167.000000	167.000000
mean	38.270060	41.108976	6.815689	46.890215	17144.688623	7.781832	70.555689	2.947964	12964.155689
std	40.328931	27.412010	2.746837	24.209589	19278.067698	10.570704	8.893172	1.513848	18328.704809
min	2.600000	0.109000	1.810000	0.065900	609.000000	-4.210000	32.100000	1.150000	231.000000
25%	8.250000	23.800000	4.920000	30.200000	3355.000000	1.810000	65.300000	1.795000	1330.000000
50%	19.300000	35.000000	6.320000	43.300000	9960.000000	5.390000	73.100000	2.410000	4660.000000
75%	62.100000	51.350000	8.600000	58.750000	22800.000000	10.750000	76.800000	3.880000	14050.000000
max	208.000000	200.000000	17.900000	174.000000	125000.000000	104.000000	82.800000	7.490000	105000.000000

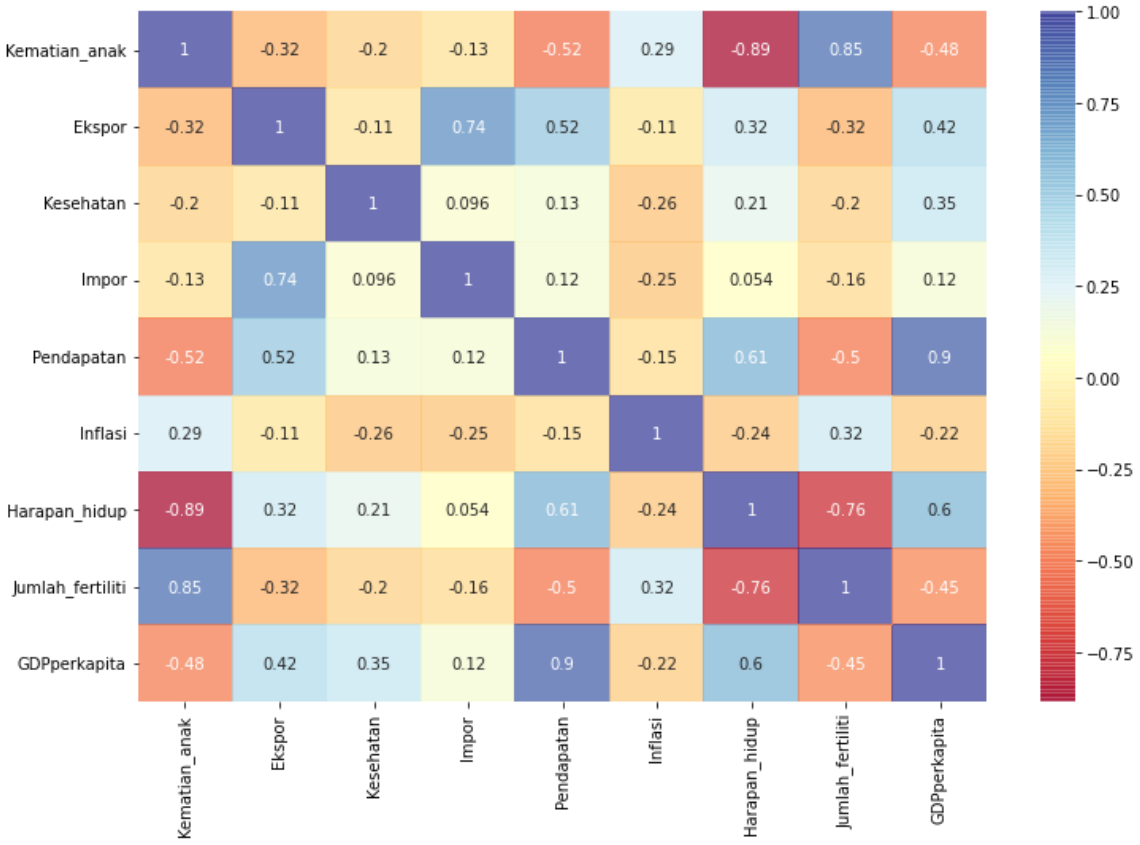
Tabel di atas merupakan ringkasan statistik dari fitur-fitur pada dataset. Berikut kesimpulan yang dapat diambil dari informasi tersebut.

- Rata-rata kematian anak di bawah lima tahun dari seluruh negara adalah 38.27 per 1000 kelahiran anak, dengan kematian tertinggi sebesar 208 dan kematian terendah sebesar 2.6.
- Sekitar 2.94 atau 3 anak lahir per wanita dari keseluruhan negara, dengan jumlah fertiliti tertinggi sebesar 7.49 dan terendah sebesar 1.51.
- Harapan hidup seluruh dunia memiliki rata-rata sebesar 70.5 tahun, dengan harapan hidup tertinggi adalah 82 tahun sedangkan yang terendah adalah 32 tahun.
- Rata-rata GDP per kapita dari seluruh negara adalah USD 12,964. GDP terbesar adalah USD 105,000 dan terendah adalah USD 231 per total populasi.
- Pendapatan bersih rata-rata tertinggi berada pada angka USD 125,000 per individu, sedangkan yang terendah sebesar USD 608 per individu.
- Rata-rata pengeluaran per kapita pada sektor kesehatan dari seluruh negara adalah adalah USD 6,815, sedangkan pengeluaran tertinggi dan terendah masing-masing sebesar USD 17,900 dan USD 1,810 per total penduduk.
- Tingkat inflasi tertinggi berada pada angka 104%, sedangkan yang terendah sebesar -4.21%. Inflasi rata-rata pada keseluruhan negara adalah 7.78%
- Rata-rata ekspor dunia sebesar USD 41,108 per kapita, sedangkan rata-rata impor dunia adalah USD 46,890.

EXPLORATORY DATA ANALYSIS

| MULTIVARIATE ANALYSIS

Analisis multivariat dilakukan untuk melihat korelasi dari keseluruhan fitur, pendekatan yang digunakan adalah menganalisis grafik *heatmap* untuk mencari fitur yang mana yang memiliki hubungan terkuat. Indikator warna menunjukkan bahwa jika semakin dingin warna dan semakin besar nilai mutlak dari angka yang terdapat pada *heatmap*, maka semakin kuat korelasi antar fitur yang bersangkutan dan begitu juga sebaliknya.



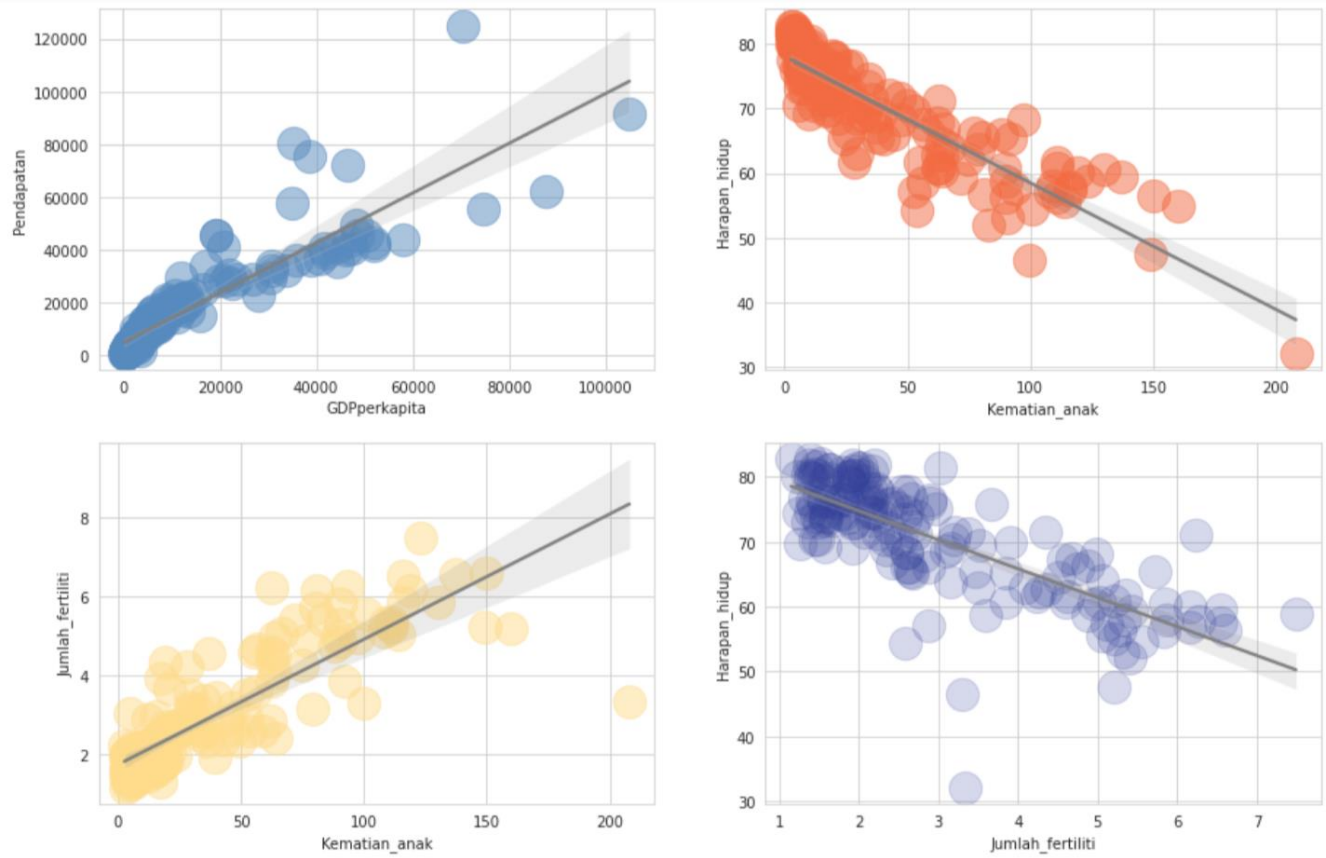
Berdasarkan grafik *heatmap* di atas, fitur yang saling berkorelasi kuat memiliki warna biru-keunguan dan jingga-kemerahan atau memiliki nilai mendekati satu dan negatif satu (atau mendekati satu jika nilainya dimutlakkan). Berikut fitur-fitur dengan urutan korelasi terendah hingga tertinggi.

- Pendapatan – GDP Per Kapita dengan nilai 0.9.
- Kematian Anak – Harapan Hidup dengan nilai -0.89.

- Kematian Anak – Jumlah Fertiliti dengan nilai 0.85.
- Jumlah Fertiliti – Harapan Hidup dengan nilai -0.76
- Ekspor – Impor dengan nilai 0.74

| BIVARIATE ANALYSIS

Analisis bivariat dilakukan untuk melihat korelasi antar dua fitur, pendekatan yang digunakan adalah menganalisis *scatter plot* yang dikombinasikan dengan *regplot* untuk mencari fitur yang mana yang memiliki hubungan terkuat. Analisis ini hanya dilakukan pada fitur empat teratas dengan korelasi terkuat atau yang memiliki nilai >0.75 atau nilai>-0.75 saja.



Plot berwarna biru, merah, dan kuning memiliki *skewness* atau tingkat kemiringan paling curam. Semakin curam grafis, maka semakin tinggi korelasinya. Berdasarkan analisis multivariat dan bivariat, maka fitur yang dipilih adalah GDP Per Kapita-Pendapatan dan Kematian Anak-Harapan Hidup.

| UNIVARIATE ANALYSIS & FEATURE SELECTION

Analisis univariat dilakukan untuk mengeksplorasi lebih mendalam masing-masing fitur, pendekatan yang digunakan adalah menganalisis histogram untuk melihat persebaran nilai pada tiap-tiap fitur. Analisis hanya dilakukan pada fitur-fitur yang telah dipilih pada tahap analisis sebelumnya, yaitu GDP Per Kapita, Pendapatan, Kematian Anak, dan Harapan Hidup.

Berdasarkan histogram di atas, dapat dilihat bahwa persebaran data tidak merata karena masih terdapat data pencilan atau *outlier* pada tiap fitur. Pada grafik GDP Per Kapita dan Pendapatan, dapat dilihat bahwa masih terdapat beberapa negara dengan GDP per kapita dan pendapatan terlampau tinggi. Hal ini akan mengganggu proses *modelling* yang akan dilakukan pada tahap selanjutnya karena tujuan kita adalah mencari negara yang berhak diberi bantuan dana sehingga data negara dengan keadaan finansial yang baik tidak diperlukan.

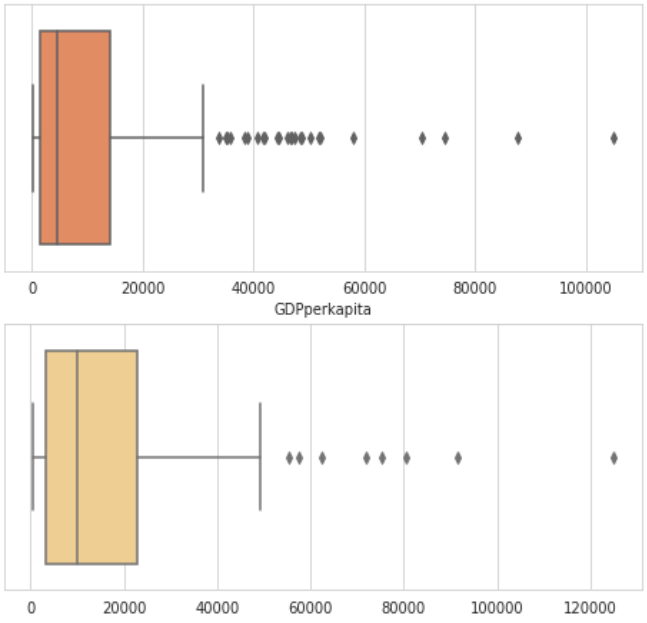
Sedangkan untuk grafik Kematian Anak dan Harapan Hidup, *outlier* merupakan data negara dengan tingkat kematian anak yang sangat tinggi dan harapan hidup yang sangat rendah. Hal ini menyebabkan *outlier* menjadi data yang sangat penting karena negara yang membutuhkan bantuan dana berkemungkinan besar memiliki taraf kematian anak yang tinggi dengan angka harapan hidup rendah pula. Sehingga untuk grafik Kematian Anak dan Harapan Hidup, persebaran data yang tidak merata dan *outlier* tidak menjadi masalah.

PRE-PROCESSING DATA

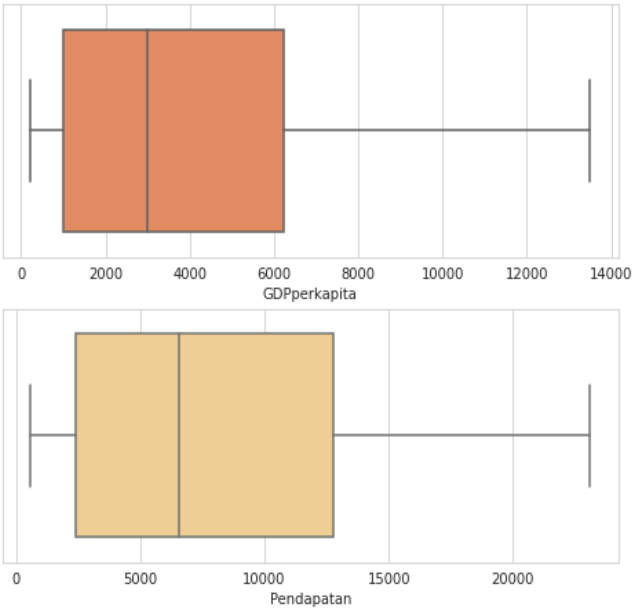
| OUTLIER TREATMENT

Berdasarkan analisis univariat yang telah dilakukan pada tahap sebelumnya, terdapat *outliers* pada masing-masing fitur. *Outliers* pada fitur GDP Per Kapita dan Pendapatan memiliki angka yang tinggi sehingga mengindikasikan negara-negara yang memiliki keadaan finansial yang sangat baik. *Outliers* pada kedua fitur ini perlu dihilangkan karena dapat menganggu model yang bertujuan untuk mengelompokkan negara yang perlu diberi dana bantuan.

Sedangkan *outliers* pada fitur Kematian Anak memiliki angka yang tinggi dan pada fitur Harapan Hidup memiliki angka yang rendah sehingga mengindikasikan negara dengan tingkat kesejahteraan yang mengkhawatirkan. *Outliers* pada kedua grafik ini tidak disarankan untuk dihilangkan karena data tersebut sangat dibutuhkan dan sejalan dengan target *cluster* yang akan dibuat oleh model. Digunakan *boxplot* untuk memetakan posisi *outliers.*



Grafik boxplot sebelum *outlier* dihilangkan



Grafik boxplot setelah *outlier* dihilangkan

| FEATURE SCALING

Perbedaan rentang nilai antara tiap fitur dapat menyebabkan proses pelatihan model terganggu karena *gap* nilai yang terlalu jauh. Digunakan pendekatan standarisasi yang telah disediakan oleh pustaka Scikit-Learn yang akan mengubah rentang nilai seluruh fitur berada di antara -1 dan 1.

Fitur GDP Per Kapita dan Pendapatan memiliki nilai ratusan hingga ratusan ribu, sedangkan fitur Kematian Anak dan Harapan Hidup didominasi oleh nilai puluhan dan ratusan saja. Jika data fitur tidak distandarisasi, maka hasil *modelling* akan sangat dipengaruhi oleh fitur GDP Per Kapita dan Pendapatan saja, sedangkan fitur Kematian Anak dan Harapan Hidup akan memiliki peran yang sangat sedikit pada model karena nilainya insignifikan dibandingkan kedua fitur sebelumnya.

	GDP Per Kapita	Pendapatan	Kematian Anak	Harapan Hidup
0	553.0	1610.0	90.2	56.2
1	4090.0	9930.0	16.6	76.3
2	4460.0	12900.0	27.3	76.5
3	3530.0	5900.0	119.0	60.1
4	12200.0	19100.0	10.3	76.8

Data *features* sebelum proses standarisasi

	GDP Per Kapita	Pendapatan	Kematian Anak	Harapan Hidup
0	-0.965151	-1.027871	1.291532	-1.619092
1	-0.003626	0.290383	-0.538949	0.647866
2	0.096958	0.760962	-0.272833	0.670423
3	-0.155861	-0.348146	2.007808	-1.179234
4	2.201058	1.743315	-0.695634	0.704258

Data *features* sebelum proses standarisasi

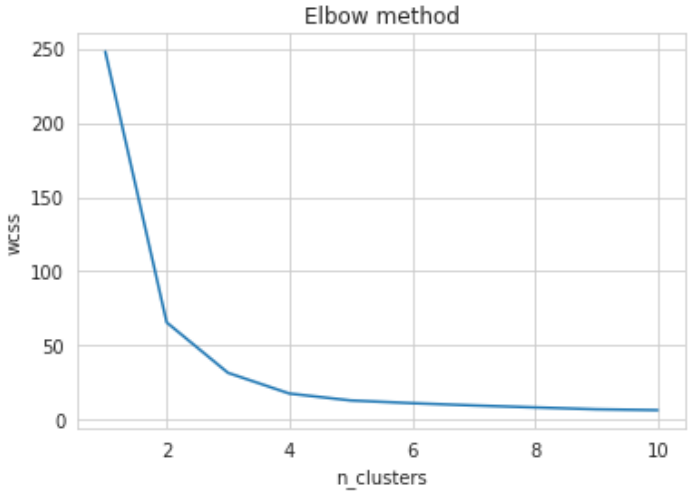
CREATING K-MEANS CLUSTERING MODEL

Data yang telah melewati tahap *pre-processing* akan dilatih menjadi suatu model yang nantinya akan dipakai untuk mengelompokkan kandidat negara untuk mendapatkan dana bantuan. Algoritma yang dipakai untuk proses *modelling* adalah K-Means Clustering, salah satu algoritma *clustering* paling populer, sederhana, dan efisien karena waktu komputasinya yang cepat. K-Means merupakan algoritma yang berbasis sentroid sehingga bersifat sensitif terhadap *outliers* dan skala nilai pada dataset, tetapi telah dilakukan usaha untuk menangani masalah tersebut pada tahap sebelumnya.

Berdasarkan hasil analisis multivariat, didapatkan hasil bahwa fitur yang memiliki korelasi terkuat adalah GDP Per Kapita-Pendapatan dan Kematian Anak-Harapan Hidup, sehingga menggabungkan keempat fitur ini pada satu proses *clustering* yang sama secara sekaligus akan berdampak buruk pada hasil klasifikasi. Oleh karena itu, proses *clustering* akan dilakukan dua kali dengan *clustering* pertama dilakukan pada fitur GDP Perkapita dan Pendapatan, lalu *clustering* terakhir pada fitur yang tersisa. Label yang dihasilkan pada kedua proses *clustering* ini akan dikombinasikan agar didapatkan hasil klasifikasi yang optimal.

| CLUSTERING PERTAMA

Tahap awal untuk mengimplementasikan algoritma K-Means adalah menentukan jumlah *cluster* awal yang akan dibentuk. Salah satu cara untuk mengetahui jumlah *cluster* yang optimal adalah menggunakan pendekatan Elbow Method dan Silhouette Score.



| ELBOW METHOD

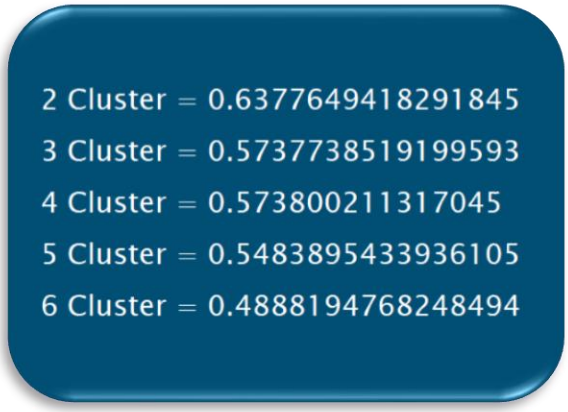
Berdasarkan grafik implementasi Elbow Method, potensi jumlah *cluster* optimal adalah 2 dan 3 *clusters* karena mulai dari garis grafik *cluster* keempat, perubahan yang dialami bersifat insignifikan.

| SILHOUETTE SCORE

Untuk mendukung hasil analisis tersebut, dilakukan perhitungan Silhouette Score sebagai metrik untuk mengukur kinerja tiap *cluster*. Rentang nilai koefisien Silhouette berada dari -1 hingga 1 dengan penjelasan sebagai berikut.

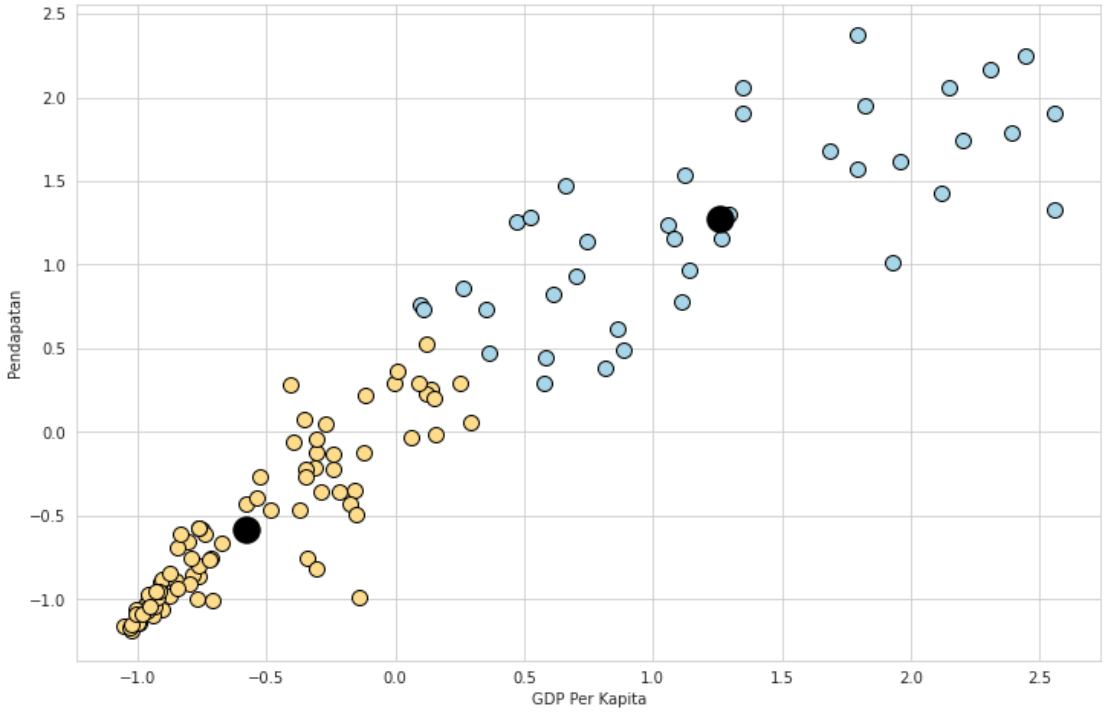
- 1: *cluster* terpisah dan dapat dibedakan secara jelas dengan *cluster* lainnya.
- 0: *cluster* tidak dapat dibedakan dengan *cluster* lainnya atau bisa disimpulkan jarak antar *cluster* sangat tidak signifikan.
- -1: *cluster* dihasilkan melalui proses yang tidak tepat.

Berdasarkan hasil perhitungan Silhouette Score, koefisien tidak lagi mengalami perubahan yang berarti setelah *cluster* ketiga dan menurun secara siginfikan pada *cluster* keenam. Koefisien Silhouette terbaik dihasilkan oleh dua *clusters*, hal ini sesuai dengan analisis grafik Elbow Method yang mengindikasikan dua sebagai jumlah *cluster* yang optimal.



| CLUSTERING VISUALIZATION

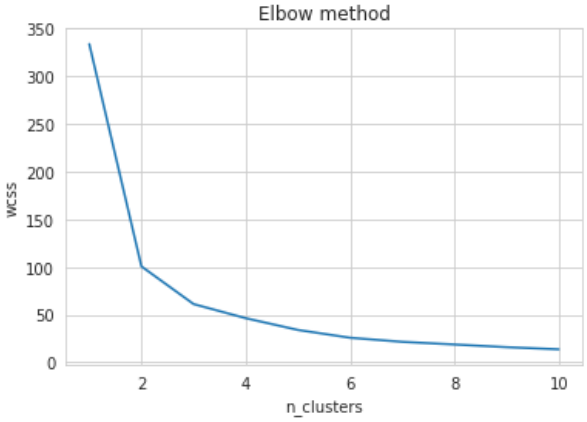
Untuk *clustering* pertama, yaitu menggunakan fitur GDP Per Kapita dan Pendapatan, digunakan *scatter* plot dua dimensi untuk melihat persebaran data negara-negara berdasarkan kedua fitur. Dapat disimpulkan bahwa *cluster* berwarna biru merupakan *cluster* negara yang telah memiliki keadaan finansial yang baik karena memiliki angka GDP per kapita dan pendapatan yang tinggi. Sedangkan *cluster* berwarna kuning yang berada di sudut kiri bawah merupakan negara yang berpotensi membutuhkan dana bantuan karena memiliki nilai GDP per kapita dan pendapatan yang rendah. Clustering berwarna kuning diberi label '0' dan cluster biru diberi label '1'.



Visualisasi hasil *clustering* berdasarkan fitur GDP Per Kapita dan Pendapatan

| CLUSTERING KEDUA

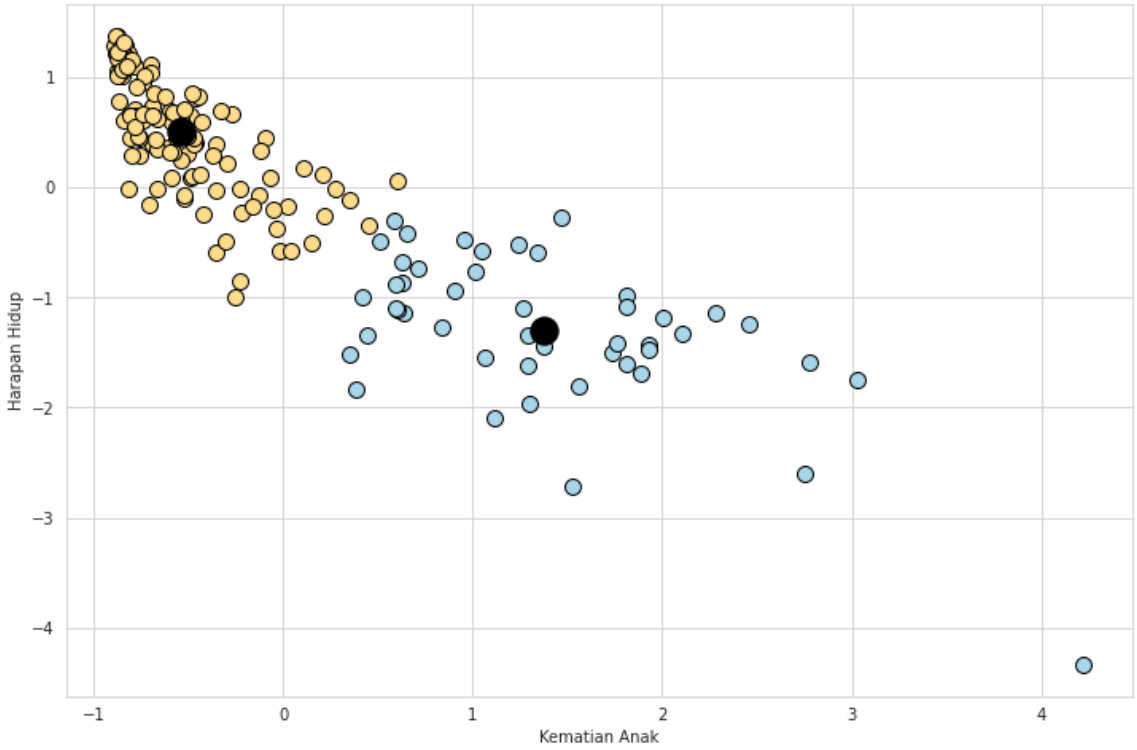
Proses yang dilakukan untuk *clustering* kedua sama dengan *clustering* pertama. Clustering dilakukan pada fitur Kematian Anak dan Harapan Hidup. Seperti *clustering* pertama, dilakukan implementasi Elbow Method dan Silhouette Score untuk mendapatkan jumlah *cluster* yang optimal. Berdasarkan hasil kedua metode ini dapat disimpulkan bahwa jumlah kluster yang optimal adalah dua.



2 Clusters = 0.644009650033211
3 Clusters = 0.5588335433575887
4 Clusters = 0.4689268230881661
5 Clusters = 0.4454933413352794
6 Clusters = 0.4608076762151421

| CLUSTERING VISUALIZATION

Berdasarkan visualisasi pada fitur Kematian Anak dan Harapan Hidup, dapat disimpulkan bahwa *cluster* berwarna kuning merupakan *cluster* negara yang telah memiliki tingkat kesejahteraan yang baik karena memiliki tingkat kematian anak yang rendah dan level harapan hidup yang tinggi. Sedangkan *cluster* berwarna biru merupakan negara yang berpotensi membutuhkan dana bantuan karena memiliki angka kematian anak yang tinggi dan harapan hidup yang rendah. Clustering berwarna kuning diberi label '0' dan cluster biru diberi label '1'.



FINAL RESULT

Berdasarkan hasil kedua *clustering* yang telah dilakukan, dilakukan *filtering* pada dataset negara yang memiliki label '0' pada *clustering* pertama, yaitu *cluster* kuning pada *scatter* plot yang merupakan negara dengan angka GDP per kapita dan pendapatan yang rendah dan label '1' untuk *clustering* kedua, yaitu *cluster* biru pada *scatter* plot yang merupakan negara dengan tingkat kematian anak yang tinggi dan harapan hidup yang rendah. Berikut daftar negara yang memenuhi kriteria tersebut dan akan menjadi pertimbangan untuk CEO LSM Help International dalam memberikan dana bantuan.

index	Negara	GDP Per Kapita	Pendapatan	Kematian Anak	Harapan Hidup	1st Clustering	2nd Clustering
0	Afghanistan	553	1610	90.2	56.2	0	1
3	Angola	3530	5900	119.0	60.1	0	1
21	Botswana	6350	13300	52.5	57.1	0	1
26	Burundi	231	764	93.6	57.7	0	1
28	Cameroon	1310	2660	108.0	57.3	0	1
32	Chad	897	1930	150.0	56.5	0	1
36	Comoros	769	1410	88.2	65.9	0	1
37	Congo, Dem. Rep.	334	609	116.0	57.5	0	1
38	Congo, Rep.	2740	5190	63.9	60.4	0	1
49	Equatorial Guinea	17100	33700	111.0	60.9	0	1
55	Gabon	8750	15400	63.7	62.9	0	1
56	Gambia	562	1660	80.3	65.5	0	1
59	Ghana	1310	3060	74.7	62.2	0	1
64	Guinea-Bissau	547	1390	114.0	55.6	0	1
69	India	1350	4410	58.8	66.2	0	1
80	Kenya	967	2480	62.2	62.8	0	1
81	Kiribati	1490	1730	62.7	60.7	0	1
84	Lao	1140	3980	78.9	63.8	0	1
88	Liberia	327	700	89.3	60.8	0	1
94	Malawi	459	1030	90.5	53.1	0	1
99	Mauritania	1200	3320	97.4	68.2	0	1
106	Mozambique	419	918	101.0	54.5	0	1
107	Myanmar	988	3720	64.4	66.8	0	1
112	Niger	348	814	123.0	58.8	0	1
116	Pakistan	1040	4280	92.1	65.3	0	1