



CLUSTERING THE COUNTRIES BY USING K- MEANS FOR HELP INTERNATIONAL

**FINAL
PROJECT**

PREPARED BY FATIH ASSIDHIQI

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, Tugas teman-teman adalah mengategorikan 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.



OBJECTIVE

Objective pada project ini adalah untuk mengategorikan negara menggunakan faktor sosial ekonomi dan kesehatan yang menentukan pembangunan negara secara keseluruhan.

READING AND UNDERSTANDING DATA

	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

UKURAN DATA

Kolom : 10
Baris : 167

INFORMASI DATA

```
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
```

INFORMASI NEGARA

Negara	
count	167
unique	167
top	Sri Lanka
freq	1

Pada tahap reading and understanding data ini menunjukan bahwa di dalam data tersebut memiliki 10 kolom (termasuk nama negara) dan 167 baris (unique of countries), berikut merupakan penjelasan dari setiap kolom pada dataset:

- **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 perorangan
- **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.

Dataset tersebut memiliki 1 kolom bertipe data object yaitu nama negara dan 9 kolom lainnya berupa karakteristik dari negara - negara tersebut (atau bisa disebut features), karakteristik dari setiap negara memiliki 2 jenis tipe data, yaitu 2 tipe data integer dan 7 tipe data float. Dataset yang berukuran kurang lebih 13.2 Kb tersebut tidak memiliki missing value atau bisa dibilang dataset yang kita miliki lengkap, sehingga tidak perlu dilakukan handling missing value

READING AND UNDERSTANDING DATA

STATISTIC SUMMARY

	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

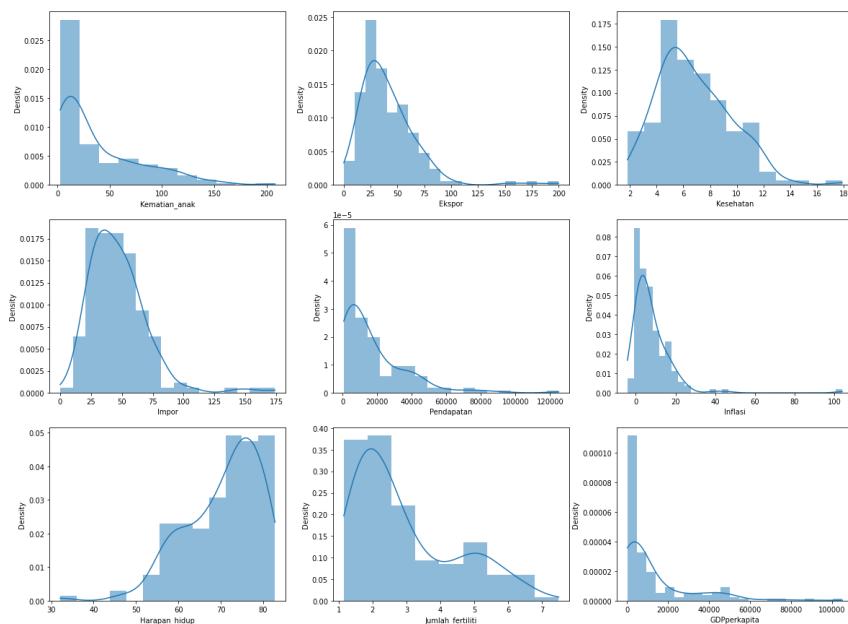
Berikut merupakan hasil ringkasan statistik sederhana dari karakteristik setiap negara:

- Rata - rata kematian anak dari seluruh negara sebesar 38,27 dengan kematian anak terbesar nya adalah 208
- Rata - rata harapan hidup dari seluruh negara sebesar 70,555689
- Jumlah fertiliti terbesar dari seluruh negara sebesar 7,49
- Total pengeluaran kesehatan diseluruh dunia memiliki rata - rata pengeluaran sebesar 6,815689
- Rata - rata pendapatan bersih perorangan dari seluruh negara yaitu 17144,688623
- Pengukuran tingkat pertumbuhan tahunan dari Total GDP (Inflasi) Terkecil dari seluruh negara yaitu -4,21
- GDP perkapita dari seluruh negara memiliki rata - rata sebesar 12964 dengan GDP terkecil yaitu sebesar 231
- Angka eksport terbesar yaitu 200 dengan rata - rata eksport seluruh negara 41,108976
- Sedangkan Angka impor terbesar yaitu 174 dengan rata - rata impor seluruh negara 46,890215

EXPLORATORY DATA ANALYSIS

Pada exploratory data analysis ini akan dilakukan 3 tahap analisis, yaitu univariate analysis, bivariate analysis, dan multivariate analysis.

UNIVARIATE ANALYSIS



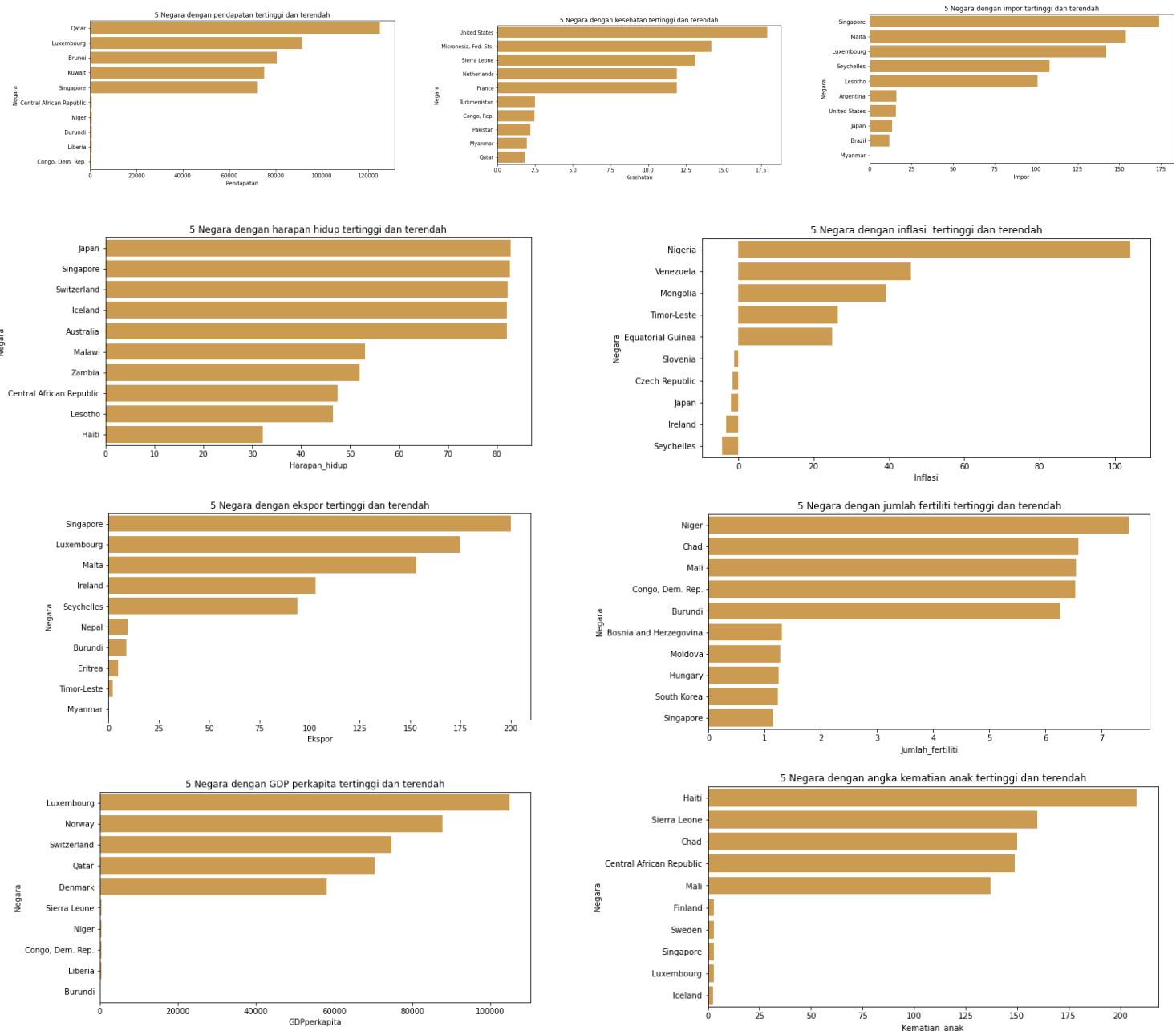
Jika kita melihat keseluruhan distribusi dari setiap features, maka kita akan langsung mengetahui bahwa features tersebut terindikasi memiliki outlier. Lalu grafik distribusi feature tersebut memiliki skewness baik itu positif maupun negatif.

Dari semua grafik diatas terindikasi memiliki outliers, outliers ini akan dianalisis lebih dalam lagi untuk dilakukan treatments outliers pada tahap pre-processing data

- Pada feature kematian anak memiliki skewness positif, bisa dilihat bahwa rata rata kematian anak lebih besar dari mayoritas angka kematian anak.
- Pada feature Ekspor memiliki skewness positif, mayoritas negara masih memiliki angka ekspor dibawah rata rata , dan terdapat negara yang memiliki tingkat ekspor yang tinggi dimana ini terlampau jauh dari pemusatan data
- Pada feature kesehatan memiliki skewness positif, yang mana mayoritas negara memiliki angka kesehatan yang sama dengan rata - rata negara.
- Pada feature impor hampir mendekati terdistribusi normal namun feature impor ini memiliki outlier sehingga distribusi datanya skewness positif, dan angka import mayoritas negara masih dibawah rata rata
- Pada feature pendapatan memiliki skewness positif, mayoritas negara masih memiliki pendapatan yang lebih kecil dari rata rata pendapatan,hanya sedikit negara - negara dengan tingkat pendapatan yang tinggi terlampau jauh dari pemusatan data.
- Pada feature inflasi memiliki skewness positif, yang mana angka inflasi mayoritas negara dibawah rata rata , dan terdapat sebuah negara yang memiliki angka inflasi yang sangat tinggi
- Pada feature harapan hidup memiliki skewness negatif dengan mayoritas negara memiliki angka harapan hidup lebih tinggi dari rata rata
- Pada feature jumlah fertiliti memiliki skewness positif, yang mayoritas negara memiliki angka fertiliti lebih kecil dari rata rata
- Pada feature GDP perkapita memiliki skewness positif yang tinggi, dimana pemusatan datanya berada hanya disekitar 0 - 10000 dan terdapat negara yang memiliki GDP perkapita yang sangat tinggi yaitu di angka 100000.

BIVARIATE ANALYSIS

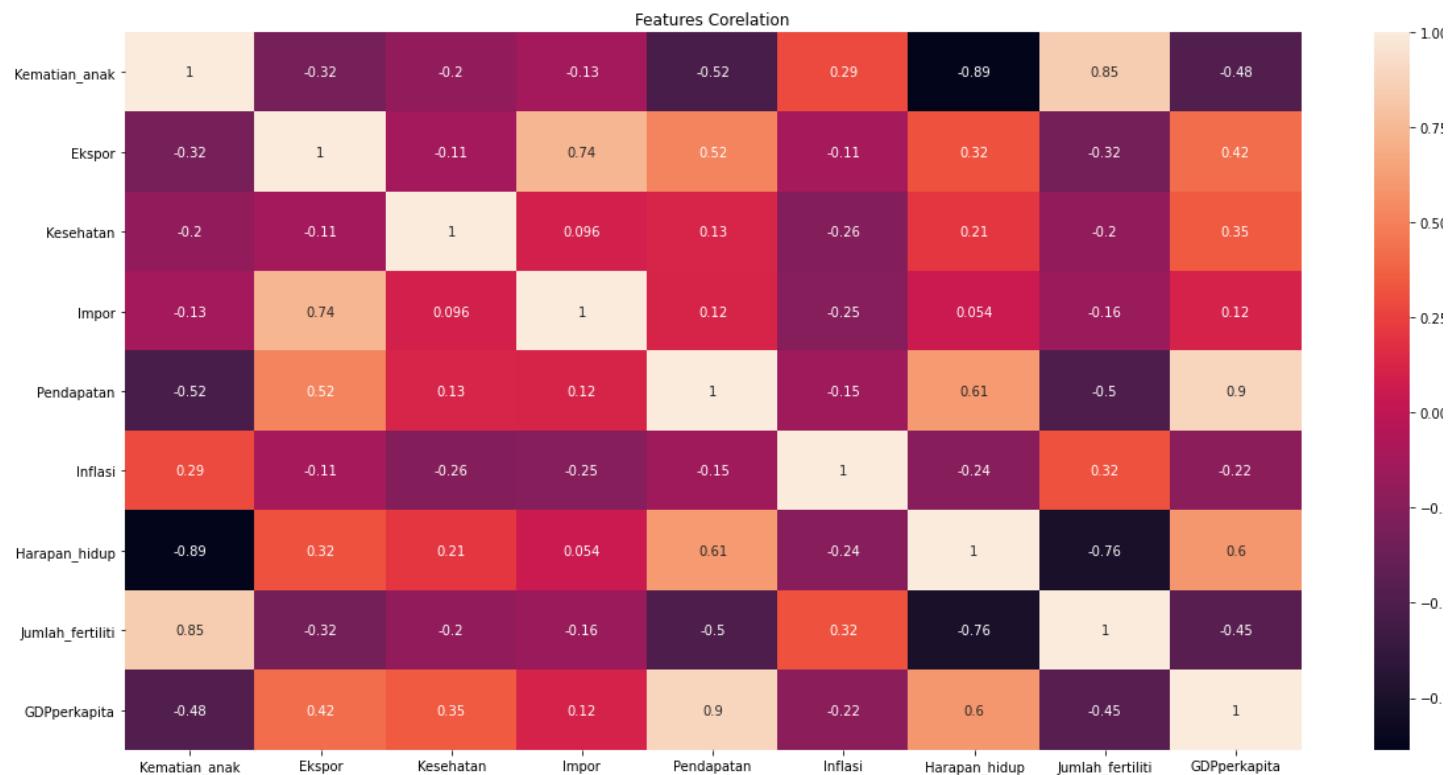
Pada bivariate analysis ini, akan dilakukan plotting untuk masing - masing features untuk menjawab pertanyaan sederhana yaitu "5 negara mana yang memiliki (features) tertinggi dan terendah". Pertanyaan tersebut dapat membantu untuk apa yang harus dilakukan ketika menemukan outliers dari suatu features.



Dari hasil ploting grafik diatas menunjukkan bahwa untuk setiap negara yang memiliki angka harapan hidup tinggi, GDP perkapita, Kesehatan, pendapatan, impor, dan eksport yang tinggi merupakan mayoritas negara dengan sektor sosial ekonomi dan kesehatan yang baik, jika sebaliknya maka negara tersebut merupakan negara dengan sektor sosial ekonomi dan kesehatan yang buruk . Sedangkan untuk setiap negara yang memiliki angka kematian anak, jumlah fertiliti, dan inflasi yang tinggi merupakan negara dengan sektor sosial ekonomi dan kesehatan yang buruk, dan juga sebaliknya maka negara tersebut merupakan negara dengan sektor sosial ekonomi dan kesehatan yang baik.

MULTIVARIATE ANALYSIS ANALYSIS

Pada multivariate analysis ini, akan dilakukan plotting untuk melihat hubungan korelasi antar features.



Berikut merupakan feature yang memiliki nilai korelasi yang kuat yaitu diatas 0,41 atau dibawah -0,41:

- kematian anak: berkorelasi dengan pendapatan, harapan_hidup, jumlah fertiliti, dan GDP perkapita
- Ekspor: berkorelasi dengan Impor, pendapatan, dan GDP perkapita
- Kesehatan: tidak memiliki korelasi yang kuat dengan feature manapun
- impor: hanya berkorelasi dengan ekspor
- Pendapatan: berkorelasi dengan kematian anak, ekspor, harapan hidup, jumlah fertiliti, dan GDP perkapita,
- Inflasi: tidak memiliki korelasi yang kuat dengan feature manapun
- harapan hidup: berkorelasi dengan kematian anak, pendapatan, jumlah fertiliti, dan GDP perkapita
- jumlah fertiliti: berkorelasi dengan kematian anak, pendapatan, harapan hidup, dan GDP perkapita
- GDP perkapita: berkorelasi dengan kematian anak, ekspor, pendapatan, harapan hidup, dan jumlah fertiliti

PREPROCESSING DATA

Pada tahap pre-processing data ini akan dilakukan beberapa tahapan yaitu:

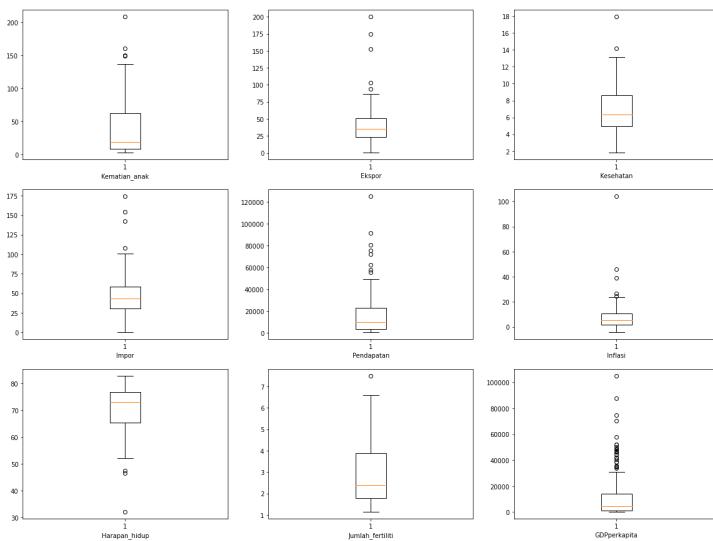
1. Handling outlier, untuk menghilangkan outlier pada sebuah feature yang mana ini berfungsi agar tidak terdapat noise ketika dilakukan clustering
2. Features scaling, yaitu mengubah range setiap feature agar memiliki range nilai yang sama.

HANDLING OUTLIER

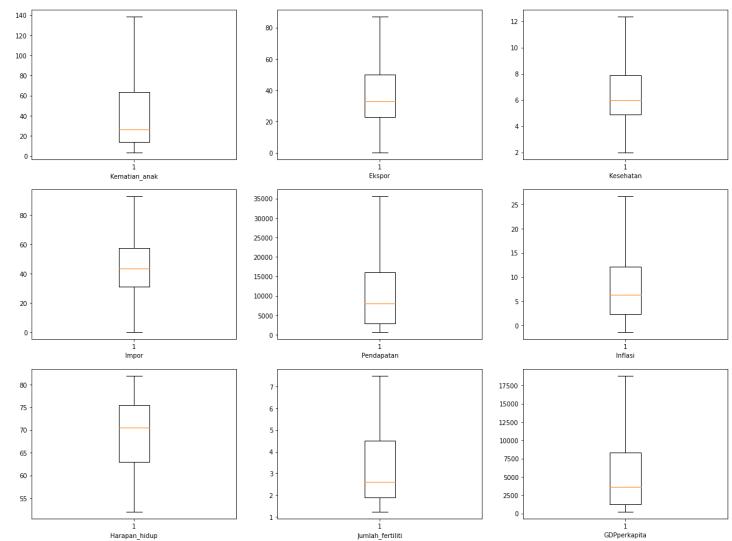
Untuk mendeteksi ada atau tidaknya outlier pada sebuah feature, yaitu dengan boxplot plotting pada setiap feature.

Mengingat tujuan dari project ini yaitu untuk mengkategorikan negara menggunakan faktor sosial ekonomi dan kesehatan yang menentukan pembangunan negara secara keseluruhan dan memilih prioritas negara mana yang paling membutuhkan bantuan. Pada tahap bivariate analysis kita juga menemukan bahwa untuk setiap negara yang memiliki angka harapan hidup tinggi, GDP perkapita, pendapatan, impor, dan ekspor yang tinggi merupakan negara sektor sosial ekonomi dan kesehatan yang baik dan untuk setiap negara yang memiliki angka kematian anak, jumlah fertiliti, dan inflasi yang tinggi merupakan negara sektor sosial ekonomi dan kesehatan yang buruk, untuk itu features harapan hidup tinggi, GDP perkapita, pendapatan, impor, dan ekspor akan dilakukan droping nilai outlier untuk mengatasi outliers jika masih terdapat nilai outlier maka akan dilakukan penggantian nilai oulier dengan nilai upper bound atau lower bound. Sedangkan untuk feature kematian anak, jumlah fertiliti, dan inflasi akan dilakukan perubahan nilai yaitu dengan mengganti nilai outlier tersebut dengan nilai upper bound atau lower bound.

BEFORE OUTLIER TREATMENT



AFTER OUTLIER TREATMENT



FEATURE SCALING

Perbedaan range nilai pada setiap features merupakan kendala yang signifikan karena beberapa algoritma pembelajaran mesin sangat sensitif terhadap fitur ini.

Untuk membuat performa terbaik pada model clustering, kita perlu melakukan features scaling. Feature scaling merupakan sebuah teknik untuk membuat semua feature pada dataset mempunyai rentang nilai yang sama. salah satu teknik feature scaling yang populer adalah standardization, atau membuat rata-rata 0 dan standard deviasi 1

STANDARDIZATION

$$X' = \frac{X - \mu}{\sigma}$$

BEFORE STANDARDIZATION

	Negara	Kematlan_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

AFTER STANDARDIZATION

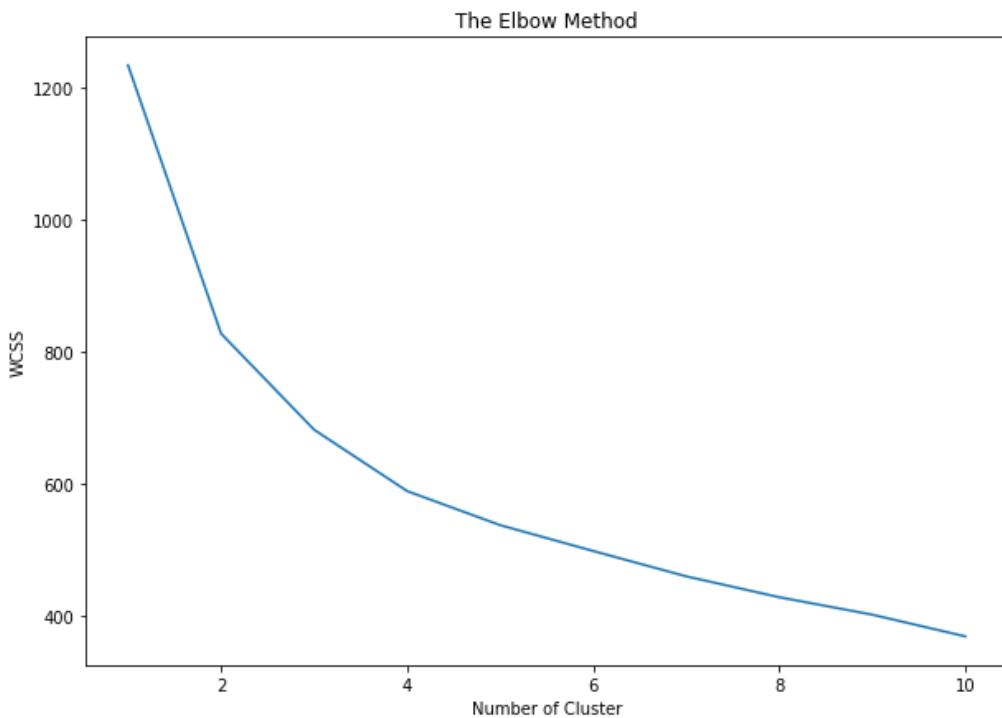
	Kematlan_anak	Ekspor	Kesehatan	Impor	Pendapatan	Inflasi	Harapan_hidup	Jumlah_fertiliti	GDPperkapita
0	1.305830	-1.422781	0.520962	-0.004687	-0.990697	0.201424	-1.746064	1.714783	-0.926821
1	-0.687742	-0.476690	0.069741	0.202540	-0.090010	-0.508477	0.938097	-0.965119	-0.304920
2	-0.397916	0.069941	-0.972887	-0.760787	0.231509	1.156564	0.964805	-0.168218	-0.239864
3	2.085923	1.326140	-1.551152	-0.116702	-0.526280	2.060075	-1.225257	1.933288	-0.403384
4	-0.858387	0.443121	-0.158060	0.779416	0.902693	-0.945892	1.004867	-0.656641	1.121039

CREATING K MEANS CLUSTERING AND VISUALIZING CLUSTERS FORMED

Setelah melakukan pre-processing data, kini dataset yang kita punya sudah siap untuk dilakukan proses clustering. Salah satu algoritma untuk clustering data adalah k-means, yang mana k-means merupakan salah satu "Unsupervised machine learning algorithms" yang paling sederhana dan populer.

Untuk menggunakan kmeans kita perlu menentukan jumlah cluster terlebih dahulu, tetapi kita tidak tahu berapa jumlah cluster yang paling optimal. Untuk menentukan jumlah cluster yang optimal kita akan menggunakan elbow method dan silhouette method.

ELBOW METHOD



Hasil dari elbow method yaitu menyarankan menggunakan 3 n cluster, tetapi untuk memastikan bahwa jumlah cluster yang paling optimal, kita akan menggunakan silhouette score

SILHOUETTE SCORE

Silhouette score adalah metrics yang digunakan untuk menghitung performa sebuah cluster. Nilainya berkisar dari -1 hingga 1.

- 1: Berarti cluster terpisah satu sama lain dan dibedakan dengan jelas
- 0: Berarti clusters indifferent atau bisa dikatakan jarak antar cluster tidak signifikan
- -1: Berarti cluster ditugaskan dengan cara yang salah

Berikut merupakan hasil dari perhitungan silhouette score yang terdiri dari 2 sampai 5 cluster :

```
2 ncluster = 0.278729857160389
3 ncluster = 0.23218497336089056
4 ncluster = 0.2213171486541378
5 ncluster = 0.2217849097456095
```

Dari hasil silhouette score tersebut menunjukkan jumlah cluster yang paling optimal adalah 2 cluster. Namun nilai pada silhouette score masih sangat kecil, itu dikarenakan kita masih menggunakan high-dimension features atau menggunakan feature yang banyak, untuk itu kita akan mencoba mengurangi jumlah feature dari 9 feature menjadi 3 feature saja

FEATURE SELECTION

Dalam memilih feature mana yang memiliki nilai silhouette terbaik, kita akan mencoba melakukan semua kombinasi 3 feature dan n_cluster untuk mendapatkan nilai silhouette score terbaik.

	feature1	feature2	feature3	n_cluster	score
0	Kematian_anak	Pendapatan	GDPperkapita	3	0.510469
1	Pendapatan	Jumlah_fertiliti	GDPperkapita	3	0.508082
2	Kematian_anak	Jumlah_fertiliti	GDPperkapita	3	0.501719
3	Kematian_anak	Harapan_hidup	Jumlah_fertiliti	3	0.492470
4	Kematian_anak	Harapan_hidup	GDPperkapita	3	0.478385
...
79	Ekspor	Kesehatan	Harapan_hidup	4	0.312711
80	Kesehatan	Impor	GDPperkapita	5	0.308597
81	Impor	Pendapatan	Inflasi	5	0.308140
82	Kesehatan	Impor	Inflasi	3	0.303605
83	Kesehatan	Impor	Pendapatan	4	0.286627

0,510469

The best of silhouette score

setelah melakukan pengurangan jumlah feature dan mencari kombinasi feature, kita mendapatkan nilai silhouette score terbaik yaitu 0,510469 yang mana nilai terbilang bagus untuk performa k-means dengan menggunakan 3 cluster.

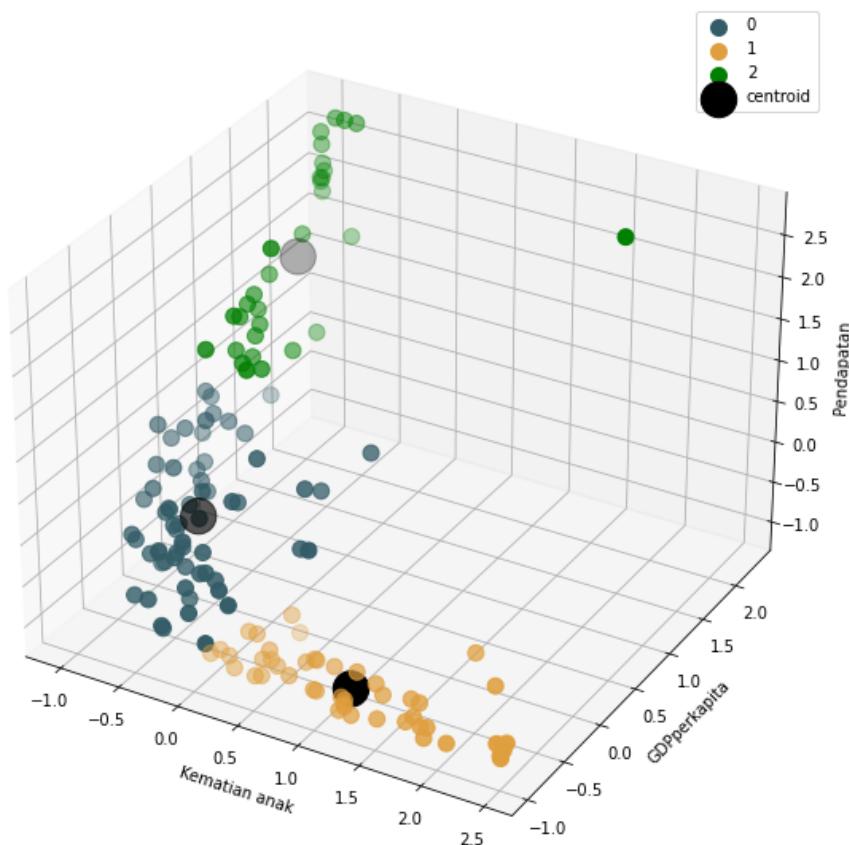
Kita akan menggunakan kolom Kematian_anak, GDPperkapita, dan Pendapatan sebagai features serta 3 n_cluster untuk model kmeans

USING 3 N_CLUSTERS TO KMEANS MODEL

Dengan menggunakan 3 jumlah cluster dan menggunakan feature Kematian anak, GDPperkapita, dan Pendapatan kita mendapatkan silhouette score yaitu 0,510469

Untuk melihat hasil dari clustering, kita akan memplotting klaster - klaster tersebut menggunakan scatter 3d plot

PLOTING CLUSTERING RESULT



Berdasarkan hasil plotting kita bisa menyimpulkannya sebagai berikut:

- Pada klaster 0 merupakan negara - negara yang memiliki jumlah Kematian anak yang rendah, GDPperkapita yang tinggi, dan Pendapatan yang tinggi.
- Pada klaster 1 merupakan negara - negara yang memiliki Kematian anak yang tinggi, GDPperkapita yang rendah, dan Pendapatan yang rendah.
- Pada klaster 2 merupakan negara - negara yang memiliki Kematian anak yang rendah, GDPperkapita yang menengah, dan Pendapatan yang menengah.

REPORT COUNTRIES

Sesuai dengan tujuan dari project ini yaitu untuk menentukan prioritas negara mana saja yang layak mendapatkan bantuan. pada tahap ini kita akan menentukan kluster mana yang diprioritaskan mendapat bantuan.

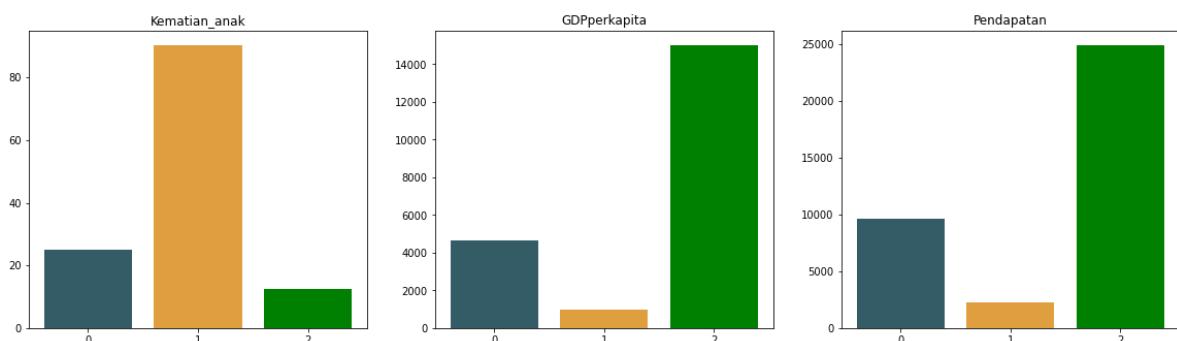
COUNTRIES CLUSTERED

	Negara	Kematian_anak	GDPperkapita	Pendapatan	label
0	Afghanistan	90.2	553.0	1610.0	1
1	Albania	16.6	4090.0	9930.0	0
2	Algeria	27.3	4460.0	12900.0	0
3	Angola	119.0	3530.0	5900.0	1
4	Antigua and Barbuda	10.3	12200.0	19100.0	2

STATISTIC SUMMARY BY CLUSTER

label	Kematian_anak	GDPperkapita	Pendapatan
0	25.011290	4624.290323	9664.838710
1	90.389130	991.478261	2269.630435
2	12.409677	15027.741935	24935.080645

PLOTTING FOR EACH FEATURES BY CLUSTER



Mengingat hasil dari analisis bivariate yaitu negera yang memiliki Kematian_anak yang tinggi sebagian besar merupakan negara berkembang, serta negara yang memiliki GDPperkapita dan Pendapatan yang tinggi sebagian besar merupakan negara - negara maju

Jika kita melihat dari hasil summary statistic dan plotting feature berdasarkan tiap klasternya, maka klaster 1 lah kelompok negara yang diprioritaskan untuk mendapatkan bantuan, karena pada klaster 1 memiliki Kematian_anak yang tinggi, GDPperkapita yang rendah, dan Pendapatan rendah.

CLUSTER 1 ALREADY SORTED

Berikut merupakan daftar negara yang mendapatkan prioritas untuk bantuan yang sudah diurutkan berdasarkan kematian anak (descending), jumlah fertiliti (descending), dan harapan hidup (ascending)

	Negara	Kematian_anak	GDPperkapita	Pendapatan	label
0	Sierra Leone	139.375	399.0	1220.0	1
1	Central African Republic	139.375	446.0	888.0	1
2	Haiti	139.375	662.0	1500.0	1
3	Chad	139.375	897.0	1930.0	1
4	Mali	137.000	708.0	1870.0	1
5	Nigeria	130.000	2330.0	5150.0	1
6	Niger	123.000	348.0	814.0	1
7	Angola	119.000	3530.0	5900.0	1
8	Congo, Dem. Rep.	116.000	334.0	609.0	1
9	Burkina Faso	116.000	575.0	1430.0	1
10	Guinea-Bissau	114.000	547.0	1390.0	1
11	Benin	111.000	758.0	1820.0	1
12	Cote d'Ivoire	111.000	1220.0	2690.0	1
13	Guinea	109.000	648.0	1190.0	1
14	Cameroon	108.000	1310.0	2660.0	1
15	Mozambique	101.000	419.0	918.0	1
16	Lesotho	99.700	1170.0	2380.0	1
17	Mauritania	97.400	1200.0	3320.0	1
18	Burundi	93.600	231.0	764.0	1
19	Pakistan	92.100	1040.0	4280.0	1
20	Malawi	90.500	459.0	1030.0	1
21	Togo	90.300	488.0	1210.0	1

22	Afghanistan	90.200	553.0	1610.0	1
23	Liberia	89.300	327.0	700.0	1
24	Comoros	88.200	769.0	1410.0	1
25	Zambia	83.100	1460.0	3280.0	1
26	Uganda	81.000	595.0	1540.0	1
27	Gambia	80.300	562.0	1660.0	1
28	Lao	78.900	1140.0	3980.0	1
29	Sudan	76.700	1480.0	3370.0	1
30	Ghana	74.700	1310.0	3060.0	1
31	Tanzania	71.900	702.0	2090.0	1
32	Senegal	66.800	1000.0	2180.0	1
33	Myanmar	64.400	988.0	3720.0	1
34	Congo, Rep.	63.900	2740.0	5190.0	1
35	Rwanda	63.600	563.0	1350.0	1
36	Kiribati	62.700	1490.0	1730.0	1
37	Timor-Leste	62.600	3600.0	1850.0	1
38	Madagascar	62.200	413.0	1390.0	1
39	Kenya	62.200	967.0	2480.0	1
40	India	58.800	1350.0	4410.0	1
41	Yemen	56.300	1310.0	4480.0	1
42	Eritrea	55.200	482.0	1420.0	1
43	Tajikistan	52.400	738.0	2110.0	1
44	Bangladesh	49.400	758.0	2440.0	1
45	Nepal	47.000	592.0	1990.0	1