Clustering The Countries by Using K-Means for HELP International

By: Wilson Alfrido Situmorang

OUTLINE

1. **LATAR BELAKANG PERMASALAHAN**
2. **KONSEP SOLUSI PERMASALAHAN**
3. **ANALISIS DATA DAN PEMBAHASAN**
   Persiapan data, EDA, Clustering, dll
4. **HASIL DAN PEMBAHASAN**
5. **KESIMPULAN**
6. **REFERENSI**

01.

LATAR BELAKANG 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, seorang data scientist diminta mengkategorikan negara menggunakan beberapa faktor sosial ekonomi dan kesehatan yang menentukan perkembangan negara secara keseluruhan.

02.

KONSEP SOLUSI PERMASALAHAN

KONSEP SOLUSI PERMASALAHAN

Untuk menentukan negara yang layak menerima bantuan, digunakan metode K-Means Clustering. Variabel yang digunakan dalam metode K-Means dipilih berdasarkan variable yang paling mewakili factor ekonomi, social, dan kesehatan untuk perkembangan suatu negara. Selain itu, variable juga ditinjau menggunakan analisis Univariate, Bivariate, dan Multivariate. Setelah negara sudah dikategorikan dari hasil beberapa kali clustering dengan variable yang berbeda, maka dicari nilai irisan negara yang memiliki factor ekonomi, social, dan kesehatan terburuk sebab menggambarkan tingkat kemakmuran negara yang rendah.

EDA Analysis

- Cleaning data
- Univariate
- Bivariate
- Multivariate

K-Means Clustering

- Menentukan variable yang digunakan
- Mengkategorikan negara berdasarkan hasil klusterisasi

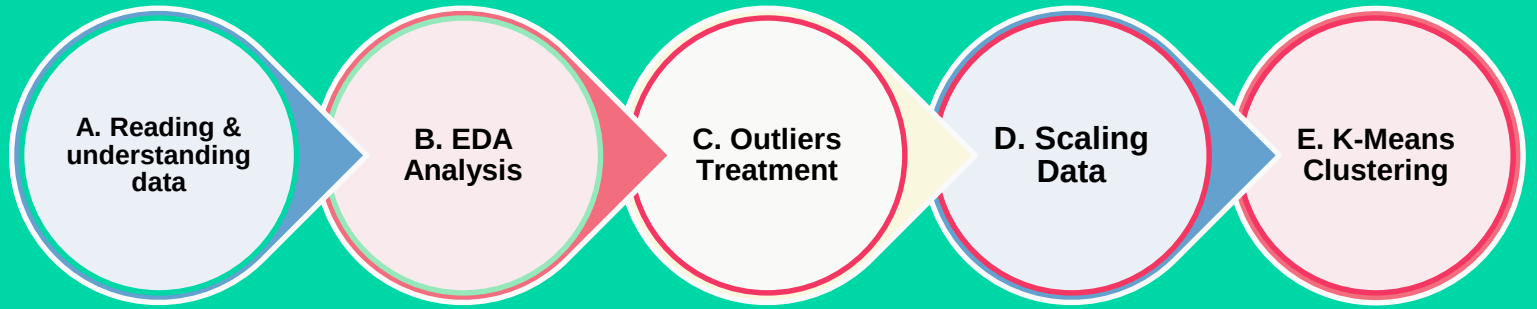
Hasil

Negara yang layak menerima bantuan ditentukan dari hasil irisan kelompok negara dari tiap hasil clustering dengan factor ekonomi, social, dan kesehatan terburuk

03.

ANALISIS DATA DAN CLUSTERING

STEPS


A. Reading & understanding data
B. EDA Analysis
C. Outliers Treatment
D. Scaling Data
E. K-Means Clustering

3A. 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

Dataset terdiri atas 10 kolom dan 167 baris , dengan berikut masing – masing deskripsi tiap kolom:

- **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.

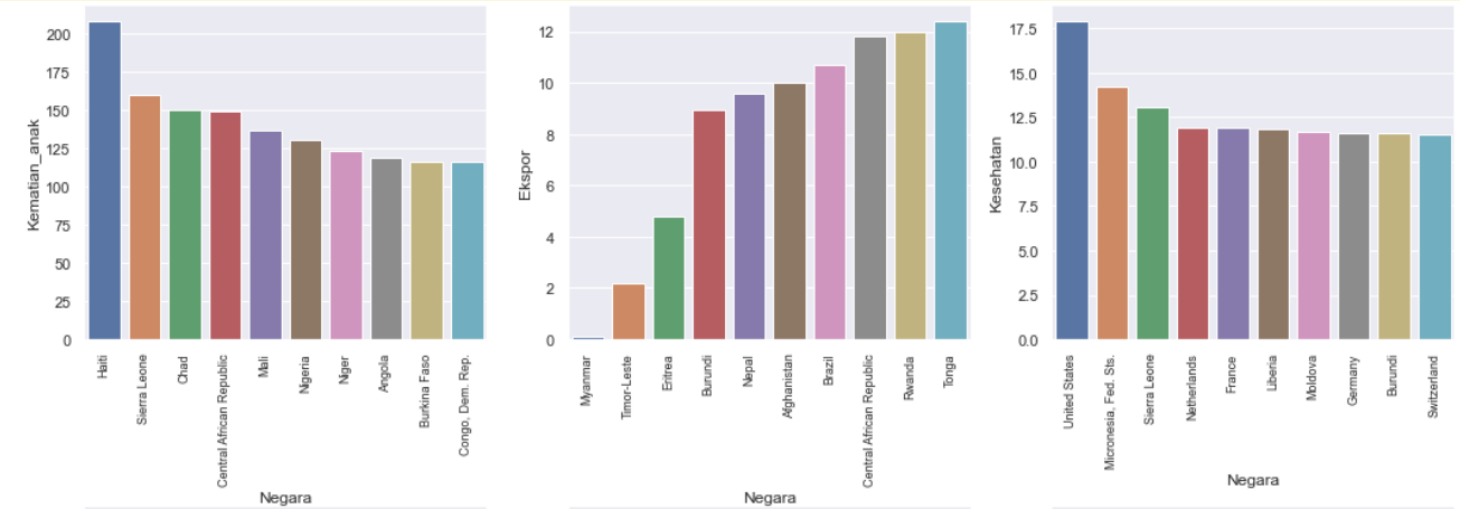
3B. EDA ANALYSIS (Cleaning Data)

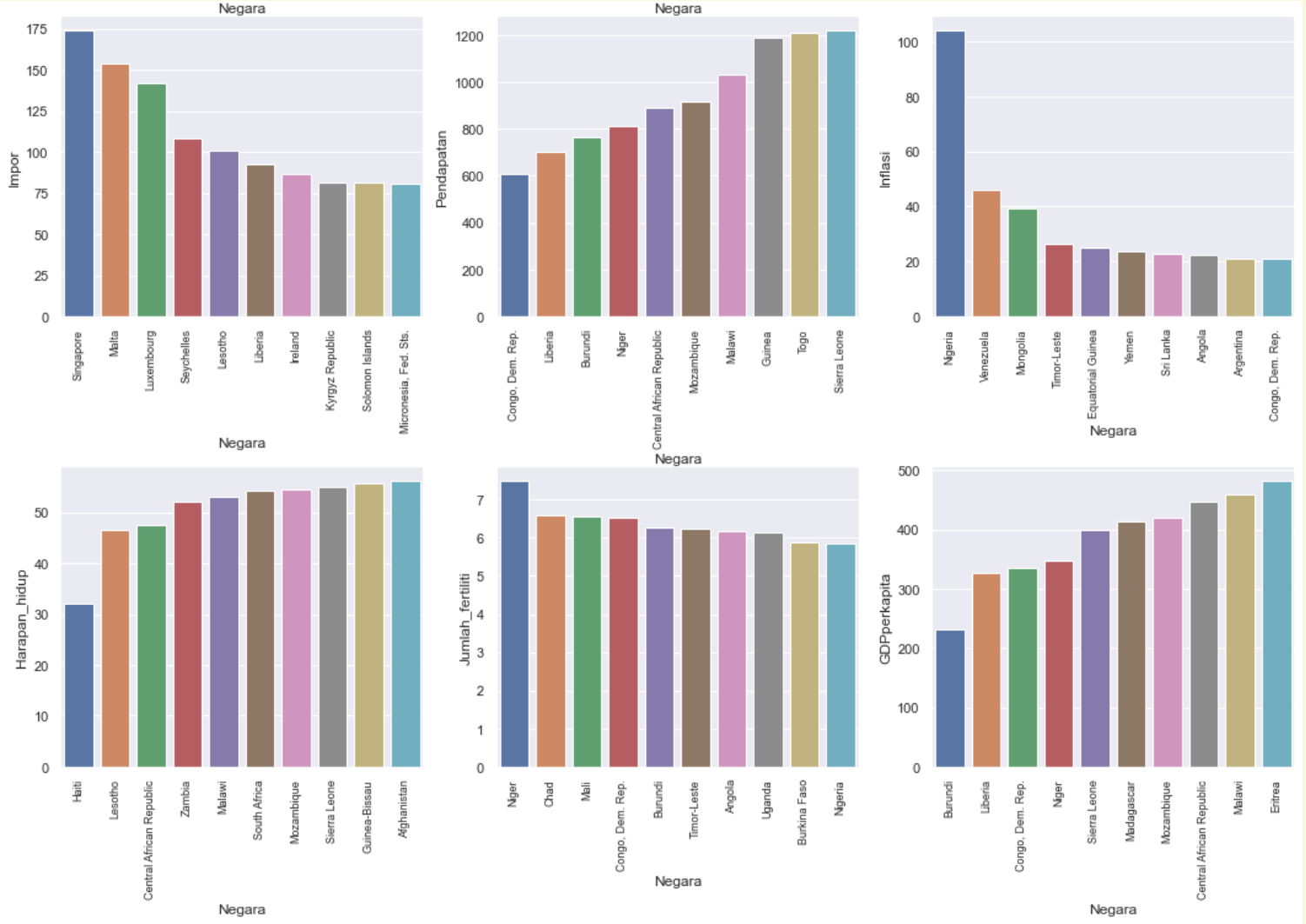
Tidak ada nilai kosong/ Nan pada tiap kolom dataset sehingga tidak perlu dilakukan proses *cleaning data*.

#	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

3B. EDA ANALYSIS (Univariate Analysis)

Analisis univariate dilakukan dengan cara membentuk ranking dari kolom - kolom data yang ada untuk merepresentasikan factor - faktor terburuk yang merepresentasikan tingkat kemakmuran suatu negara yang rendah. Untuk tiap factor/variable tersebut dipilih 10 negara. Contohnya adalah 10 negara dengan tingkat GDPperkapita terendah dan 10 negara dengan tingkat Kematian anak tertinggi.




Negara
Impor
175
150
125
100
75
50
25
0
Singapore
Malta
Luxembourg
Seychelles
Lesotho
Liberia
Ireland
Kyrgyz Republic
Solomon Islands
Micronesia, Fed. Sts.
Negara
Negara
Pendapatan
1200
1000
800
600
400
200
0
Congo, Dem. Rep.
Liberia
Burundi
Niger
Central African Republic
Mozambique
Malawi
Guinea
Togo
Sierra Leone
Negara
Negara
Inflasi
100
80
60
40
20
0
Nigeria
Venezuela
Mongolia
Timor-Leste
Equatorial Guinea
Yemen
Sri Lanka
Angola
Argentina
Congo, Dem. Rep.
Negara
Negara
Harapan_hidup
50
40
30
20
10
0
Haiti
Lesotho
Central African Republic
Zambia
Malawi
South Africa
Mozambique
Sierra Leone
Guinea-Bissau
Afghanistan
Negara
Negara
Jumlah_fertiliti
7
6
5
4
3
2
1
0
Niger
Chad
Mali
Congo, Dem. Rep.
Burundi
Timor-Leste
Angola
Uganda
Burkina Faso
Nigeria
Negara
Negara
GDPperkapita
500
400
300
200
100
0
Burundi
Liberia
Congo, Dem. Rep.
Niger
Sierra Leone
Madagascar
Mozambique
Central African Republic
Malawi
Eritrea
Negara

3B. EDA ANALYSIS (Univariate Analysis)

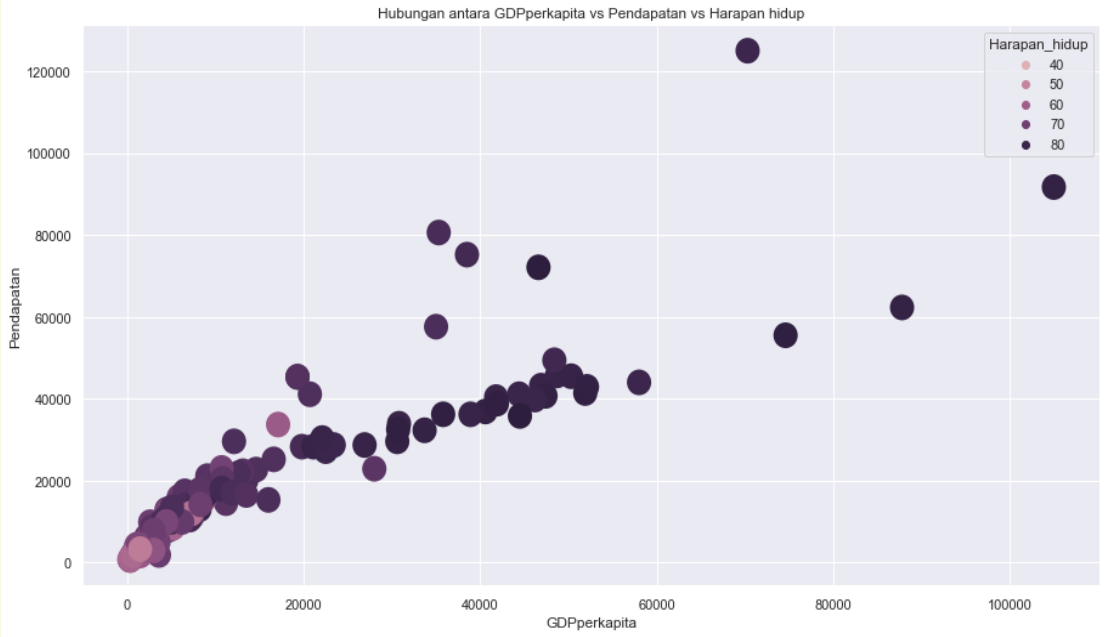
Insight analisis univariate :

- 10 negara dengan GDPperkapita terendah terdapat pada benua Afrika yang tersebar pada bagian barat,timur dan tengah. Faktor GDPP ini salah satu yang terpenting karena menunjukkan kesehatan ekonomi suatu negara. GDPP menjadi factor yang menentukan tingkat kemiskinan suatu negara.
- 10 negara dengan pendapatan per kapita terendah juga semuanya terdapat pada benua Afrika. 7 dari negara dengan pendapatan per kapitat terendah termasuk dalam 10 negara dengan GDPP terendah.
- 10 negara dengan tingkat fertilitas tertinggi juga sebagian besar berada pada benua Afrika dan 1 pada Timor-Leste. 3 di anataranya, yakni Brundi, Congo, dan Niger masuk dalam 10 negara dengan GDPP dan pendapatan per kapita terendah. Tingkat fertilitas yang tinggi dan pendapatan serta GDPP yang rendah sebagai salah satu factor kemiskinan di benua Afrika. Berdasarkan UNICEF, populasi Afrika pada 2050 akan mencapai 2 milyar orang, yang mana sangat tinggi dengan tingkat ekonomi yang rendah.
- 10 negara dengan tingkat harapan hidup terendah dengan 9 negara berada pada benua Afrika dan sisnya adalah Afghanistan.
- 10 negara dengan tingkat kematian anak tertinggi semuanya berada pada benua Afrika. Berdasarkan WHO, kematian yang terdapat di Afrika disebabkan karena kurangnya tingkat pengetahuan untuk merawat bayi dan fasilitas kesehatannya. Tingkat kematian anak tertinggi di Afrika juga tidak lepas karen tingkat fertilitas yang tinggi.

Faktor – factor di atas adalah factor yang terlihat polanya dan cukup mewakilkan suatu hal bahwa di Afrika tingkat kemakmuran negaranya masih sangat rendah

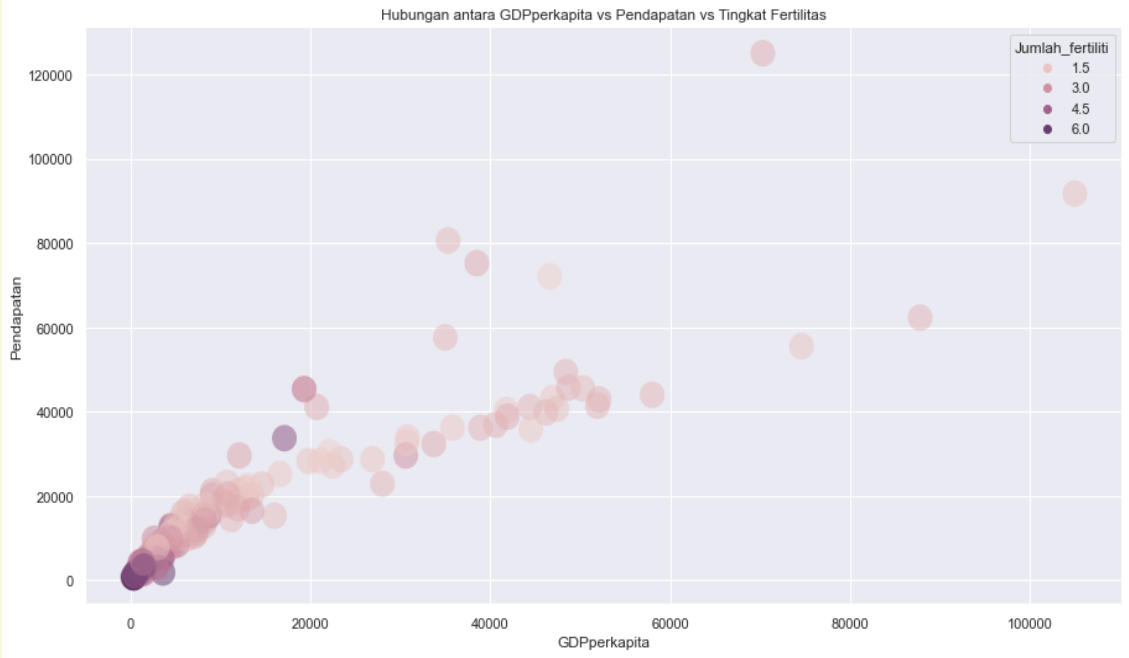
3B. EDA ANALYSIS (Bivariate Analysis)

Analisis bivariate untuk menggambarkan hubungan GDPP, pendapatan, dan harapan hidup. Terlihat melalui grafik disamping bahwa hubungan GDPP dan pendapatan per kapita cukup linear. Selain itu, terlihat bahwa harapan hidup seseorang akan semakin rendah saat nilai GDPP dan pendapatannya rendah. Hal ini dapat terjadi karena GDPP dan pendapatan perkapita sangat menentukan tingkat kemakmuran suatu negara. Saat hidup makmur, maka umur warga cenderung akan lebih lama.



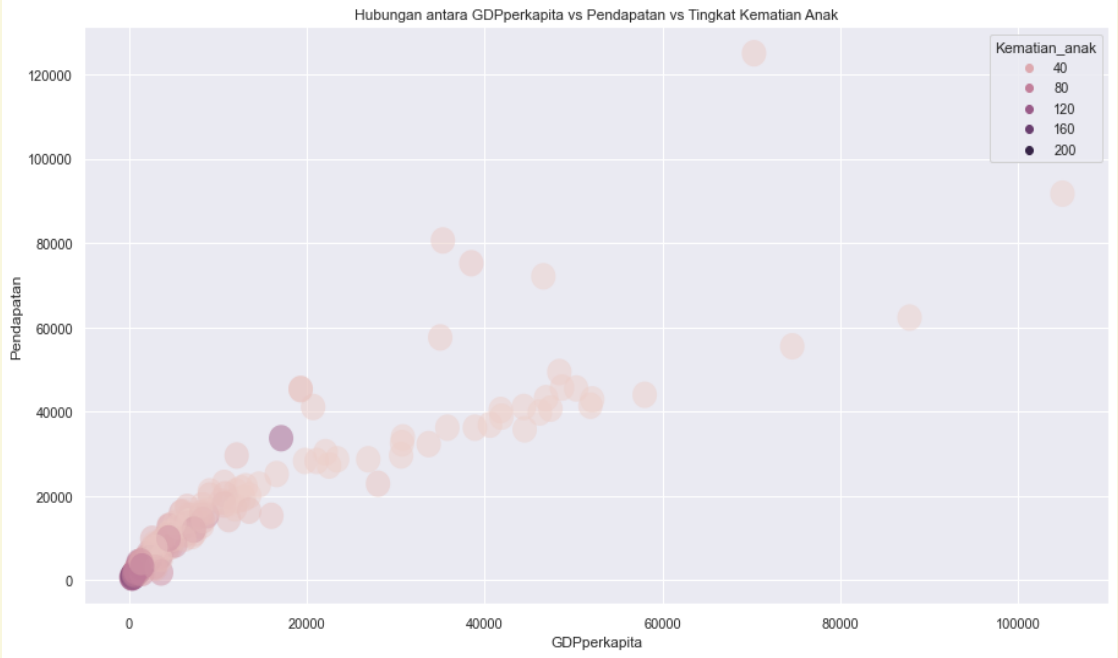
3B. EDA ANALYSIS (Bivariate Analysis)

Analisis bivariate untuk menggambarkan hubungan GDPP, pendapatan, dan jumlah fertiliti. Terlihat bahwa, negara dengan GDPP dan pendapatan rendah cenderung memiliki tingkat kelahiran anak yang lebih tinggi dibandingkan dengan negara yang memiliki GDPP dan pendapatan per kapita yang tinggi.



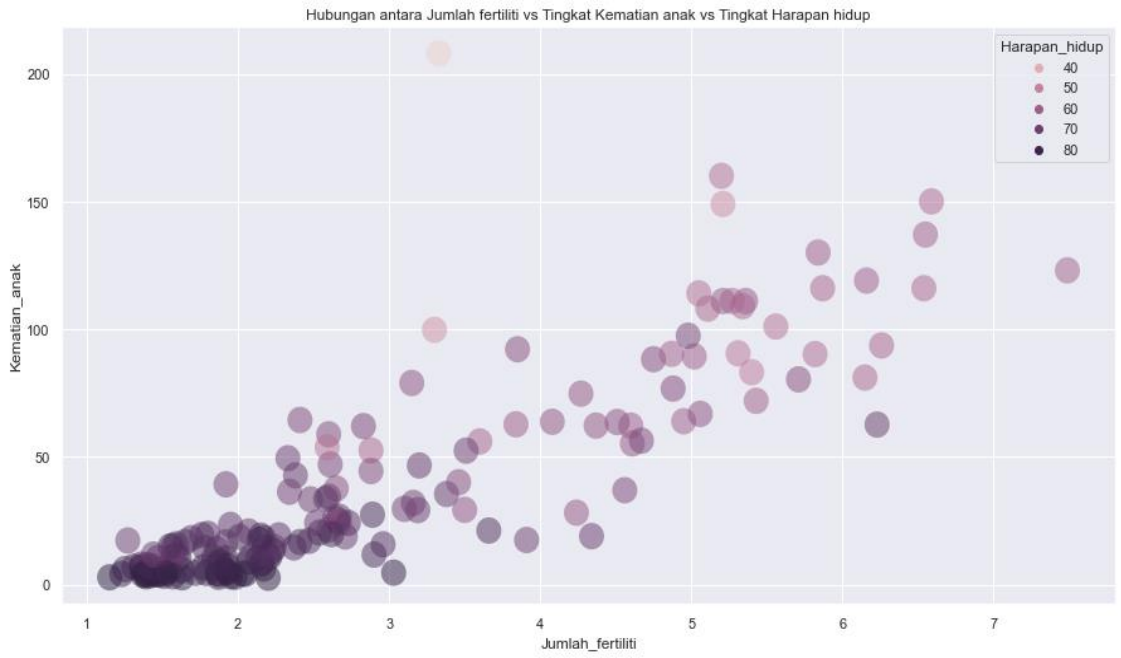
3B. EDA ANALYSIS (Bivariate Analysis)

Analisis bivariate untuk menggambarkan hubungan GDPP, pendapatan, dan Tingkat kematian anak. Terlihat bahwa, negara dengan GDPP dan pendapatan rendah cenderung memiliki tingkat kematian anak yang lebih tinggi dibandingkan dengan negara yang memiliki GDPP dan pendapatan per kapita yang tinggi.



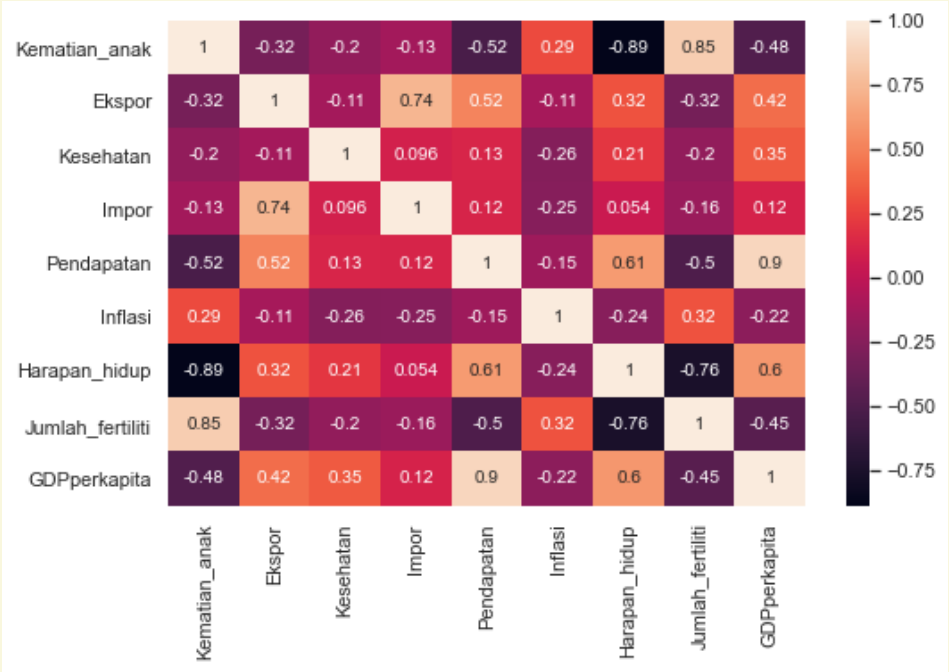
3B. EDA ANALYSIS (Bivariate Analysis)

Analisis bivariate untuk menggambarkan hubungan Jumlah vertiliti, Kematian anak, dan Harapan hidup. Terlihat melalui grafik disamping bahwa negara dengan tingkat fertilitas dan kematian anak yang rendah akan memiliki harapan hidup yang lebih tinggi. Hal tersebut dapat terjadi karena harapan hidup berkaitan erat dengan tingkat kematian sebab saat mati pada umur yang muda dapat membuat harapan hidup lebih rendah.



3B. EDA ANALYSIS (Multivariate Analysis)

Terlihat disamping grafik heatmap yang menunjukkan nilai korelasi dari tiap hubungan variable. Nilai korelasi tersebut menggambarkan kelinearitasan hubungan antar 2 variable. Terlihat bahwa nilai korelasi antara GDPP dan pendapatan perkapita cukup besar, yakni 0,9 sehingga grafik pada analisis bivariate sebelumnya cukup linear. Selain itu hubungan jumlah fertility dan kematian anak juga memiliki nilai korelasi yang cukup besar, yakni 0.85. Begitu juga hubungan harapan hidup dan kematian anak memiliki nilai -0.89.



3C. OUTLIER TREATMENT

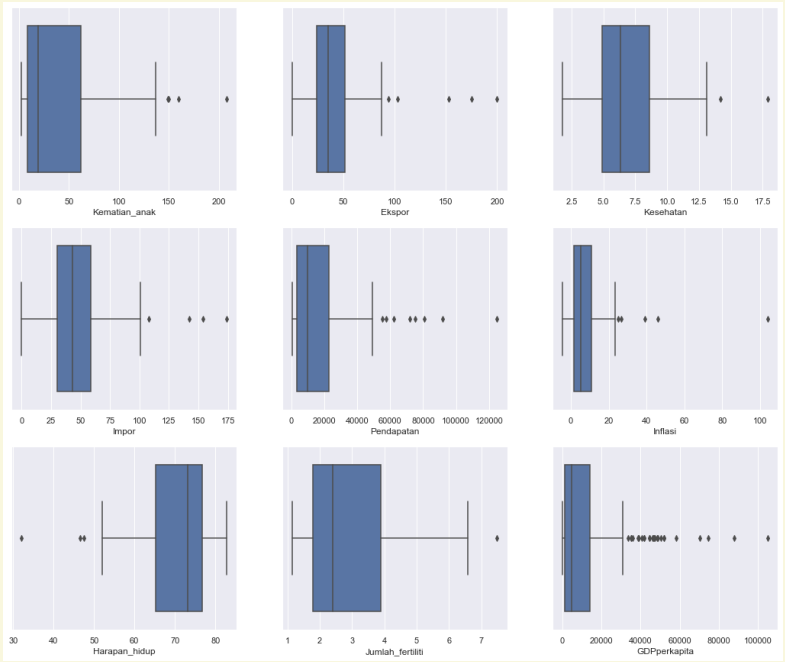
Outlier treatment dilakukan dengan membuang nilai outlier negara pada variable GDPP dan pendapatan perkapita. Hal tersebut dilakukan karena saat di aplikasikan kepada K-Means clustering membuat negara – negara yang notabene maju atau cukup makmur tidak termasuk sehingga dapat lebih focus pada clustering negara – negara miskin/tertinggal. Sesudah dilakukan treatmen, data rata – rata dan std variable GDPP dan pendapatan perkapita menjadi rendah. Sedangkan variable lain ada yang naik atau turun.

	Kematian_anak	Ekspor	Kesehatan	Impor	Pendapatan	Inflasi	Harapan_hidup	Jumlah_fertiliti	GDPperka
count	167.000000	167.000000	167.000000	167.000000	167.000000	167.000000	167.000000	167.000000	167.00
mean	38.270060	41.108976	6.815689	46.890215	17144.688623	7.781832	70.555689	2.947964	12964.15
std	40.328931	27.412010	2.746837	24.209589	19278.067698	10.570704	8.893172	1.513848	18328.70
min	2.600000	0.109000	1.810000	0.065900	609.000000	-4.210000	32.100000	1.150000	231.00
25%	8.250000	23.800000	4.920000	30.200000	3355.000000	1.810000	65.300000	1.795000	1330.00
50%	19.300000	35.000000	6.320000	43.300000	9960.000000	5.390000	73.100000	2.410000	4660.00
75%	62.100000	51.350000	8.600000	58.750000	22800.000000	10.750000	76.800000	3.880000	14050.00
max	208.000000	200.000000	17.900000	174.000000	125000.000000	104.000000	82.800000	7.490000	105000.00

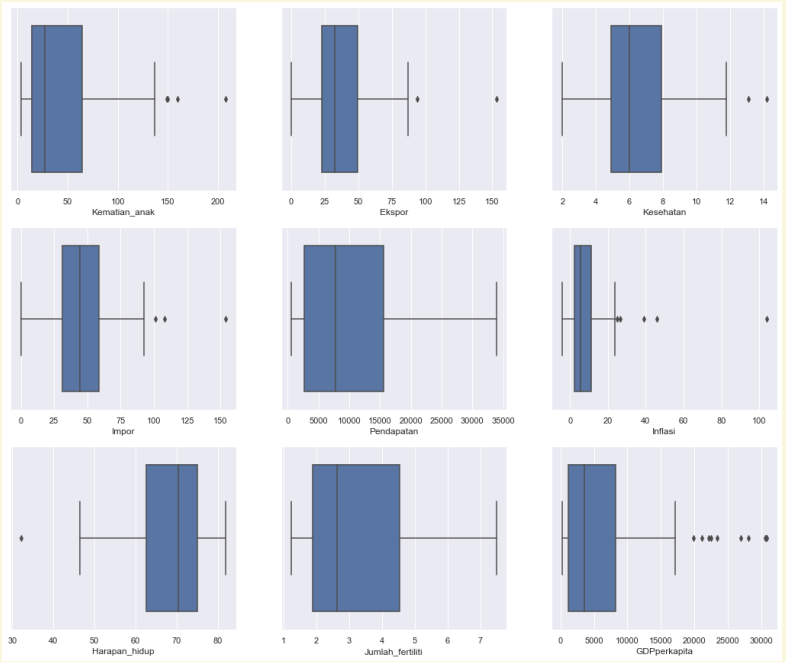
Sebelum treatment

Sesudah treatment

	Kematian_anak	Ekspor	Kesehatan	Impor	Pendapatan	Inflasi	Harapan_hidup	Jumlah_fertiliti	GDPperkap
count	139.000000	139.000000	139.000000	139.000000	139.000000	139.000000	139.000000	139.000000	139.0000
mean	44.805755	37.460424	6.476115	46.704071	10223.474820	8.495791	68.656835	3.159281	6051.1798
std	41.203406	21.755730	2.363593	21.331449	8704.818232	11.192298	8.519764	1.567514	6902.0749
min	3.200000	0.109000	1.970000	0.065900	609.000000	-4.210000	32.100000	1.230000	231.0000
25%	14.300000	22.750000	4.920000	31.100000	2675.000000	2.320000	62.500000	1.870000	1185.0000
50%	27.300000	32.700000	6.010000	44.500000	7820.000000	5.730000	70.400000	2.620000	3530.0000
75%	64.150000	49.800000	7.900000	58.900000	15650.000000	11.350000	75.050000	4.535000	8215.0000
max	208.000000	153.000000	14.200000	154.000000	33900.000000	104.000000	81.900000	7.490000	30800.0000



Sebelum treatment



Sesudah treatment

Data outlier yang dihapus pada kolom GDPperkapita dan pendapatan per kapita membuat variable lainnya juga berkurang nilai outliernya. Contohnya adalah seperti kolom impor dan ekspor.

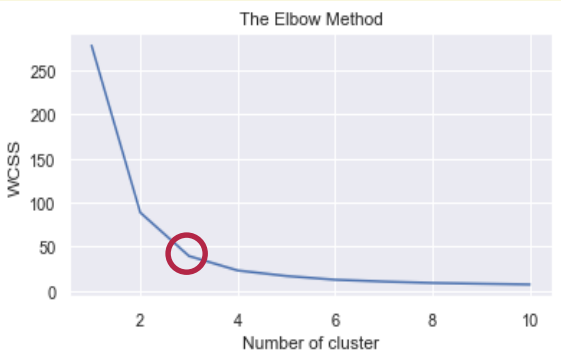
3D. SCALING DATA

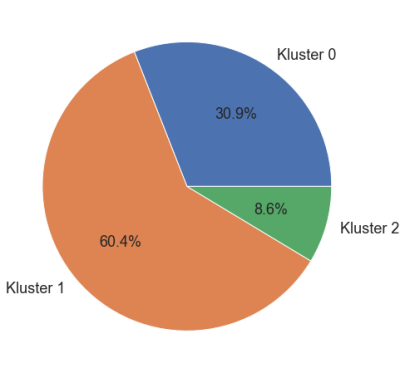
- Scaling data dilakukan karena antar variable memiliki batas – batas nilai yang berbeda sehingga perlu diubah agar membantu juga dalam proses visualisasi hubungan variable sehingga nilai – nilai dapat terlihat jelas.
- Scaling dilakukan dengan terlebih dahulu memilih pasangan variable yang ingin dibuat kluster dengan menggunakan K-Means. Adapun variable yang dipilih untuk dilakuk clustering, yakni GDPP perkapita vs Pendapatan per kapita dan Kematian anak vs Harapan hidup sehingga ada 2 kali proses clustrering. Alasan dari pemilihan variable – variable tersebut adalah GDPP dan pendapatan per kapita sangat mewakilkan factor ekonomi yang menentukan tingkat kemakmuran suatu negara. Semakin besar nilai kedua variable tersebut menunjukkan kegiatan perekonomian berjalan dengan baik. Selain itu, variable Kematian anak dan harapan hidup cukup mewakilkan factor social dan kesehatan, yang mana tingkata kematian anak yang tinggi menunjukkan suatu negara tidak mampu memberikan fasilitas kesehatan dan pengetahuan akan perawatan seorang anak. Semakin tinggi nilai harapan hidup maka menunjukkan tiap warganya hidup makmur serta menjadi indicator negara tersebut cukup maju.

3E. K-MEANS CLUSTERING

- GDPP vs Pendapatan per kapita

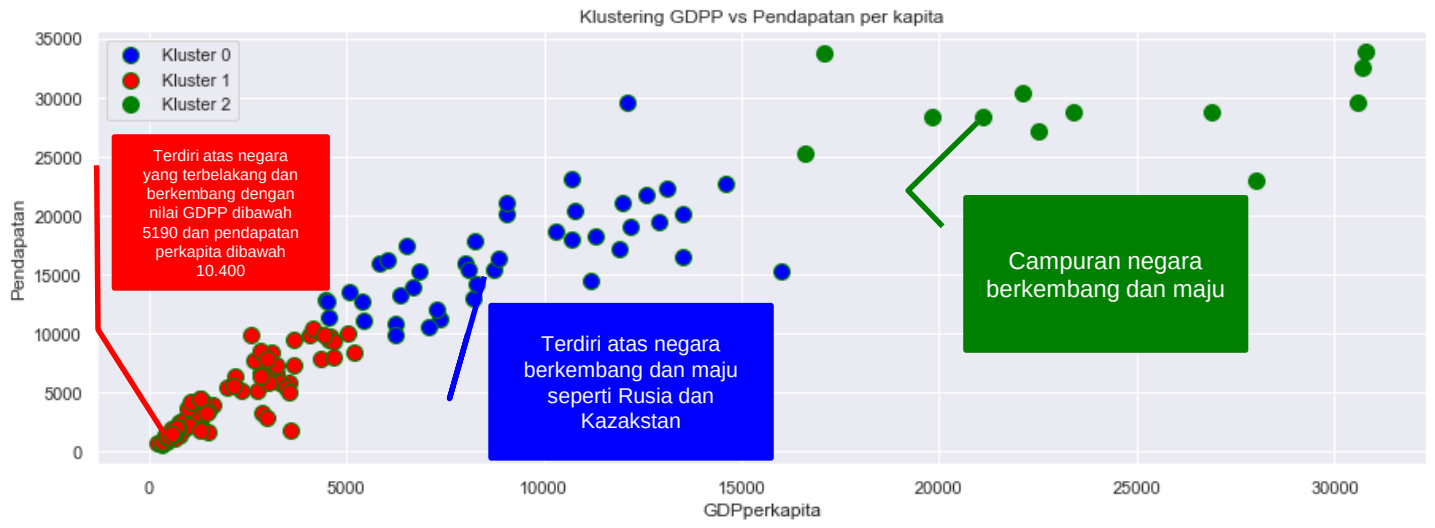
Jumlah kluster yang digunakan berjumlah 3 berdasarkan elbow method. Nilai skor silhouette yang diperoleh = 0.62. Berikut di bawah terdapat grafik pie ditribusi label. **Kluster 1 berisi negara yang ingin ditinjau** karena memiliki nilai GDPP dan Pendapatan per kapita yang cenderung lebih rendah dibanding label yang lain. Dalam kluster tersebut terdapat gabungan antara negeri berkembang seperti Indonesia dan negara terbelakang seperti negara yang ada di benua Afrika.





Kluster	Jumlah Negara
0	43
1	84
2	12

3E. K-MEANS CLUSTERING

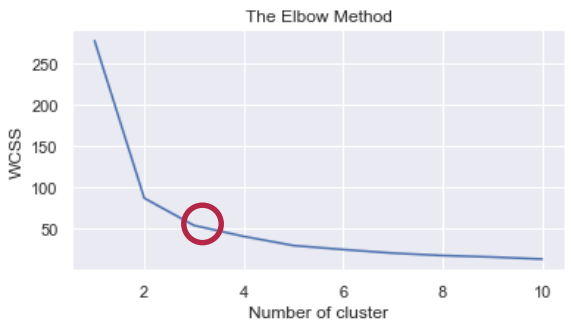


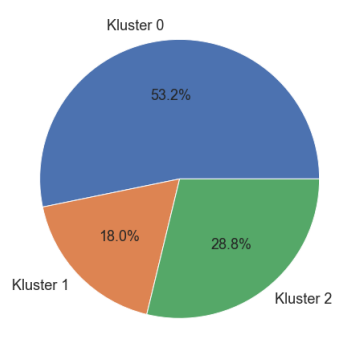
Negara – negara maju cenderung sudah sedikit karena telah dibuang saat proses outlier treatment. Efek outlier treatment membuat kluster 1 datanya berkurang dari 128 negara menjadi 84 negara dan data yang terbuang cenderung negara yang berkembang.

3E. K-MEANS CLUSTERING

- Kematian anak vs Harapan hidup

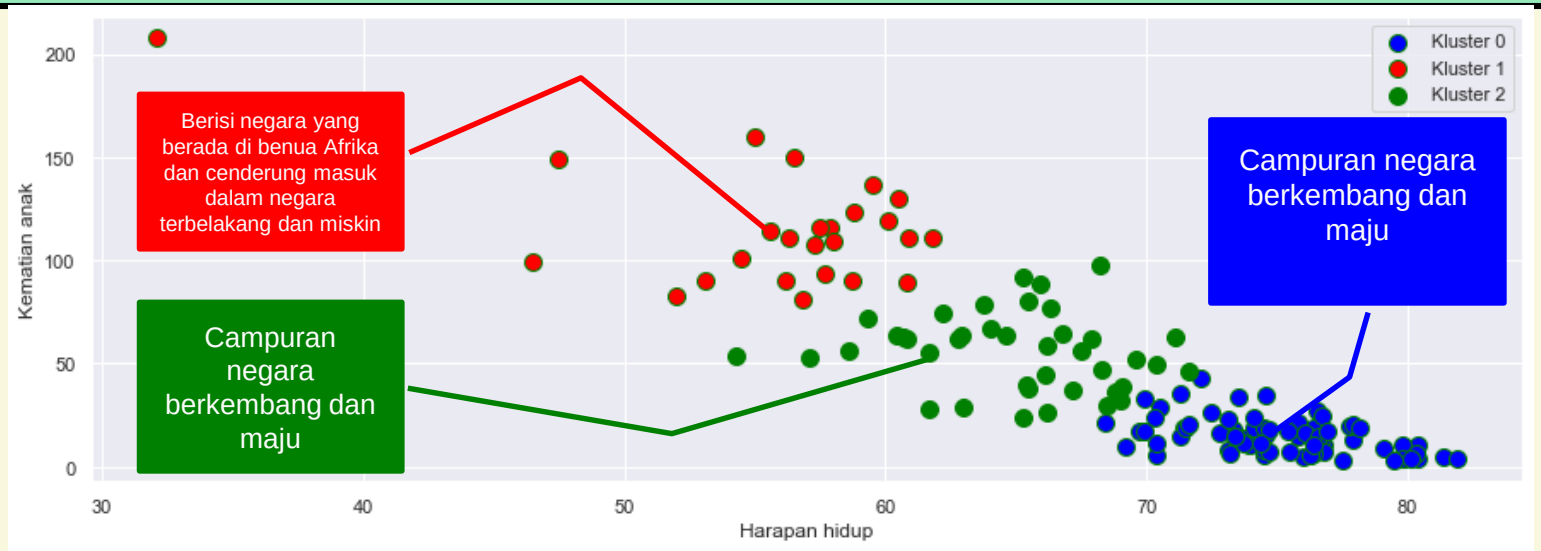
Jumlah kluster yang digunakan berjumlah 3 berdasarkan elbow method. Nilai skor silhouette yang diperoleh = 0.53. **Kluster yang ingin ditinjau adalah kluster 1** karena memiliki nilai harapan hidup lebih sedikit dan kematian anak yang lebih tinggi. Kluster 1 berisi 24 negara yang berada di benua Afrika dan sisanya adalah Afghanistan.





Kluster	Jumlah Negara
0	74
1	25
2	40

3E. K-MEANS CLUSTERING



04.

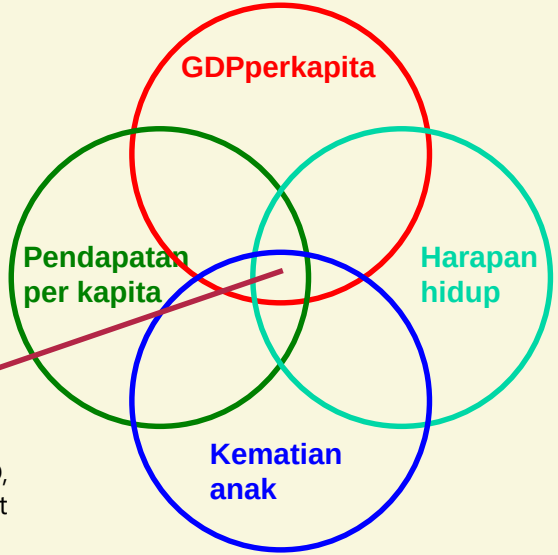
HASIL DAN PEMBAHASAN

4. HASIL DAN PEMBAHASAN

Negara yang berhak mendapatkan bantuan ditentukan dengan cara sebagai berikut:

1. Mengambil daftar 20 negara terburuk dari tiap variable berdasarkan masing – masing kluster yang ditinjau. Daftar tersebut adalah 20 negara terburuk nilai **GDPP** dari kluster 1, 20 negara terburuk nilai **Pendapatan per kapita** dari kluster 1, 20 negara dengan tingkat **kematian anak** tertinggi dari kluster 1, dan 20 negara dengan tingkat **harapan hidup** terendah dari kluster 1
2. Lalu, dari tiap variable tersebut dicari nilai irisan yang menunjukkan negara yang akan diberi bantuan

Terdapat 10 negara yang menerima bantuan = (Congo DR, Burundi, Niger, Central African, Mozambique, Guinea,Sierra Leone, Guinea-Bissau, Burkina Faso, Haiti, Malawi)



10 Negara tersebut merupakan negara pada benua Afrika. Berdasarkan UNCTAD, 8 negara tersebut termasuk dalam negara terbelakang. Negara – negara tersebut memiliki tingkat kemiskinan yang tinggi sehingga perlu mendapatkan bantuan dana jika terjadi bencana.

05.

KESIMPULAN

5. KESIMPULAN

- 10 Negara yang memperoleh bantuan dari HELP International adalah Congo DR, Burundi, Central African, Mozambique, Guinea,Sierra Leone, Guinea-Bissau, Burkina Faso, Malawi, Niger dan Haiti.
- Negara yang menerima bantuan termasuk dalam negara terbelakang
- Tingkat kemiskinan suatu negara dapat dilihat melalui GDPP dan Pendapatan per kapita.
- Tingkat kematian anak yang tinggi pada suatu negara menunjukkan kurangnya perhatian terhadap fasilitas dan pengetahuan dalam kesehatan anak.
- Tingkat harapan hidup lebih rendah pada negara terbelakang.

Referensi

- *Child mortality and causes of death*. Who.int. (2021). Retrieved 8 August 2021, from https://www.who.int/data/gho/data/themes/topics/topic-details/GHO/child-mortality-and-causes-of-death.
- Is GDP per Capita the Same as Income per Capita?. Forex Education. (2021). Retrieved 6 August 2021, from https://www.forex.in.rs/is-gdp-per-capita-the-same-as-average-income/.
- On the poorest continent, the plight of children is dramatic. SOS-US-EN. (2021). Retrieved 7 August 2021, from https://www.sos-usa.org/about-us/where-we-work/africa/poverty-in-africa.
- The Poorest Countries in the World (2019-2023). FocusEconomics | Economic Forecasts from the World's Leading Economists. (2021). Retrieved 7 August 2021, from https://www.focus-economics.com/blog/the-poorest-countries-in-the-world.
- *UN list of least developed countries | UNCTAD*. Unctad.org. (2021). Retrieved 7 August 2021, from https://unctad.org/topic/least-developed-countries/list.

Terima Kasih