LAPORAN FINAL PROJECT DATA SCIENCE

BOOTCAMP SANBERCODE

Clustering Countries by Using K-Means for HELP International

Dibuat oleh:

Raihan Astrada Fathurrahman

Latar Belakang

Gambar 1. HELP International

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,

Tujuan

Tujuan dari project ini adalah untuk mengkategorikan negara menggunakan beberapa faktor sosial ekonomi dan kesehatan yang menentukan perkembangan negara secara keseluruhan. Berdasarkan kategori tersebut kemudian menyarankan negara mana saja yang paling perlu menjadi fokus CEO.

Pembahasan

Dalam membuat project ini, hal yang dilakukan pertama kali adalah mengimpor modul-modul yang akan digunakan dalam project. Berikut merupakan beberapa modul yang digunakan pada pengerjaan project ini.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score
```

Gambar 2. Modul yang digunakan

Setelah seluruh modul telah diimpor, tahapan selanjutnya adalah memuat dataset dari file asal yang digunakan. Setelah dataset berhasil dimuat, selanjutnya yaitu memahami dataset yang akan digunakan.

	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
...	...	...	...	...	...	...	...	...	...	...
162	Vanuatu	29.2	46.6	5.25	52.7	2950	2.62	63.0	3.50	2970
163	Venezuela	17.1	28.5	4.91	17.6	16500	45.90	75.4	2.47	13500
164	Vietnam	23.3	72.0	6.84	80.2	4490	12.10	73.1	1.95	1310
165	Yemen	56.3	30.0	5.18	34.4	4480	23.60	67.5	4.67	1310
166	Zambia	83.1	37.0	5.89	30.9	3280	14.00	52.0	5.40	1460

167 rows × 10 columns

Gambar 3. Dataset yang digunakan

Pada gambar 3, dapat dilihat bahwa dataset memiliki jumlah data sebanyak 167 serta memiliki 10 fitur. Berikut merupakan penjelasan fitur-fitur yang ada 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 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.

Selanjutnya yaitu mengecek type-type data pada datasetnya serta mengecek keunikan nilai kolom Negara pada dataset. Pengecekan keunikan pada kolom negara dilakukan karena kolom Negara merupakan Candidate Key dari data.

```
<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
```

```
[ ]  len(df['Negara'].unique())

167
```

Gambar 4. Info Dataset

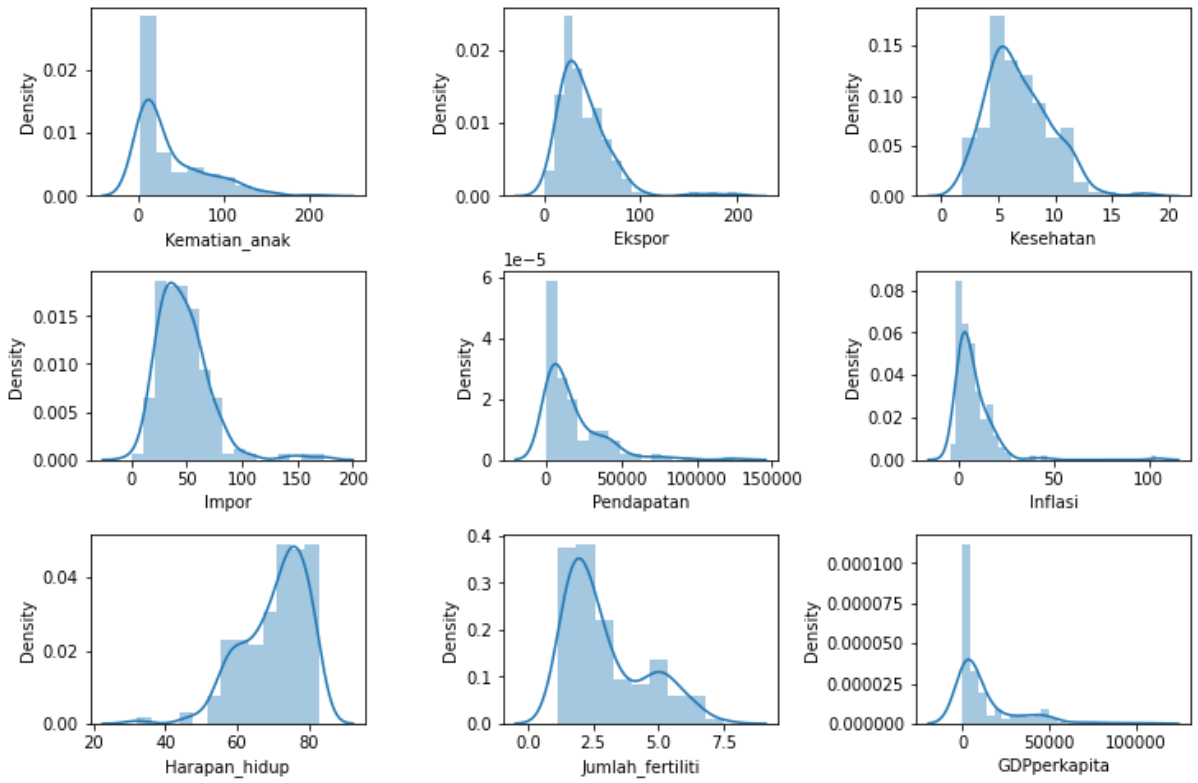
Pada Gambar 4, dapat dilihat bahwa fitur-fitur sudah memiliki type data float ataupun integer untuk yang bernilai numerik serta kolom negara tidak memiliki type numerik. Dari hal tersebut, maka tidak perlu dilakukan pengubahan type data. Dapat dilihat juga bahwa seluruh nilai pada fitur Negara merupakan nilai yang unik sehingga tidak perlu penanganan lebih lanjut terkait data yang redundan. Selain itu, dapat dilihat juga bahwa seluruh fitur tidak ada yang memiliki nilai kosong (*null*). Setelah itu akan dilakukan pengecekan nilai statistik dasar untuk mengetahui nilai statistik dasar serta memastikan bahwa pada seluruh kolom tidak terdapat data yang kosong.

```
[ ] df.describe()
```

	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

Gambar 5. Statistik Dasar Dataset

Setelah membaca dan memahami dataset yang digunakan, selanjutnya adalah melakukan Exploratory Data Analysis (EDA). Pertama adalah tahapan Data Cleaning. Pada dataset yang digunakan sudah dicek bahwa tidak terdapat missing value maupun double value sehingga dataset tidak memerlukan penanganan untuk hal-hal tersebut. Selanjutnya adalah melakukan analysis.

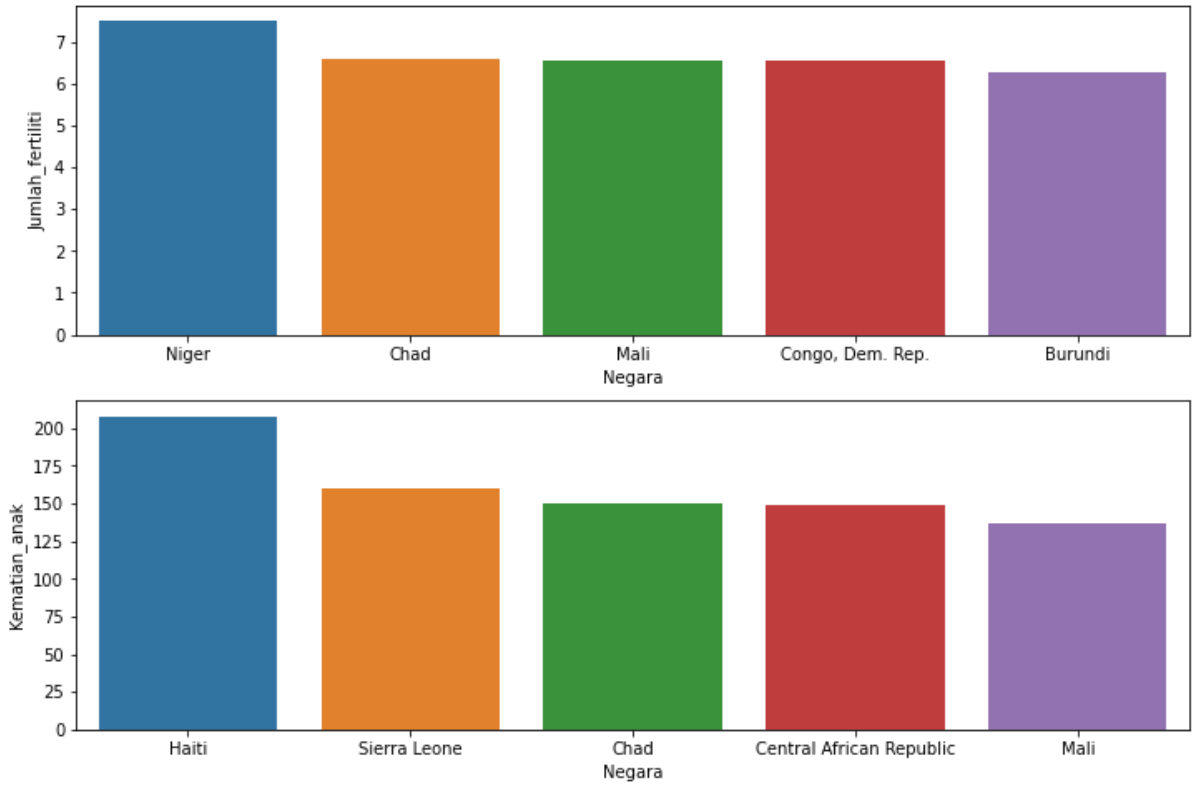


Gambar 6. Univariate Analysis

Univariate Analysis yang dilakukan adalah histogram agar dapat mengetahui gambaran persebaran data dari tiap fiturnya. Dapat dilihat bahwa hampir seluruh fitur

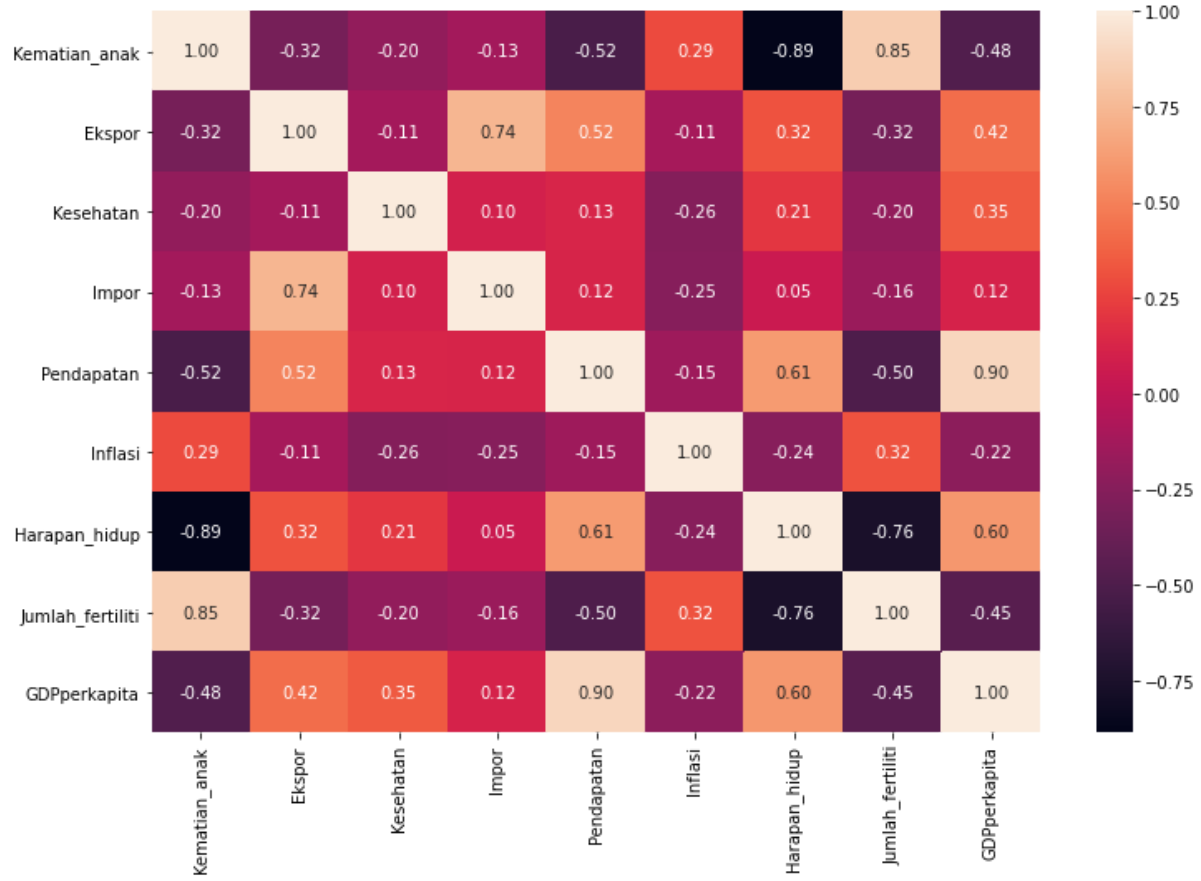
memiliki kemiringan data ke arah kanan kecuali pada fitur Harapan_hidup.Hal tersebut berarti modus data memiliki nilai yang lebih rendah dari rata-rata dari yang memiliki nilai lebih rendah daripada nilai tengahnya.

Selanjutnya adalah melakukan Bivariate Analysis. Bivariate Analysis yang digunakan adalah menggabungkan fitur yang bersifat kategorikal, yaitu negara, dengan fitur-fitur lain.



Gambar 7. Bivariate Analysis

Pada hal ini hanya dicek hubungannya dengan beberapa fitur yaitu Jumlah_fertiliti dan Kematian_anak. Dapat dilihat bahwa negara dengan Jumlah_fertiliti yang tertinggi yaitu Nigeria serta negara dengan Kematian_anak terendah yaitu Haiti. Setelah itu, hal yang dilakukan adalah melakukan multivariate analysis untuk setiap kolom-kolom numerik.

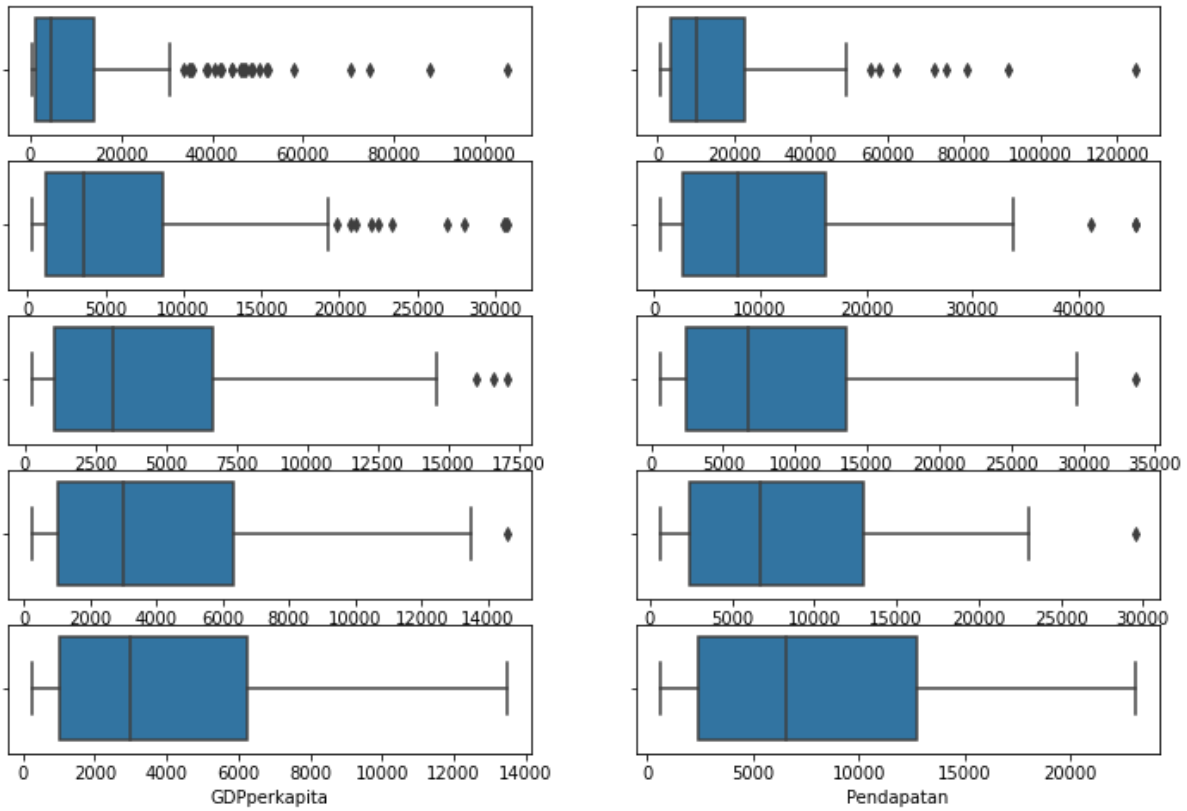


Gambar 8. Bivariate Analysis

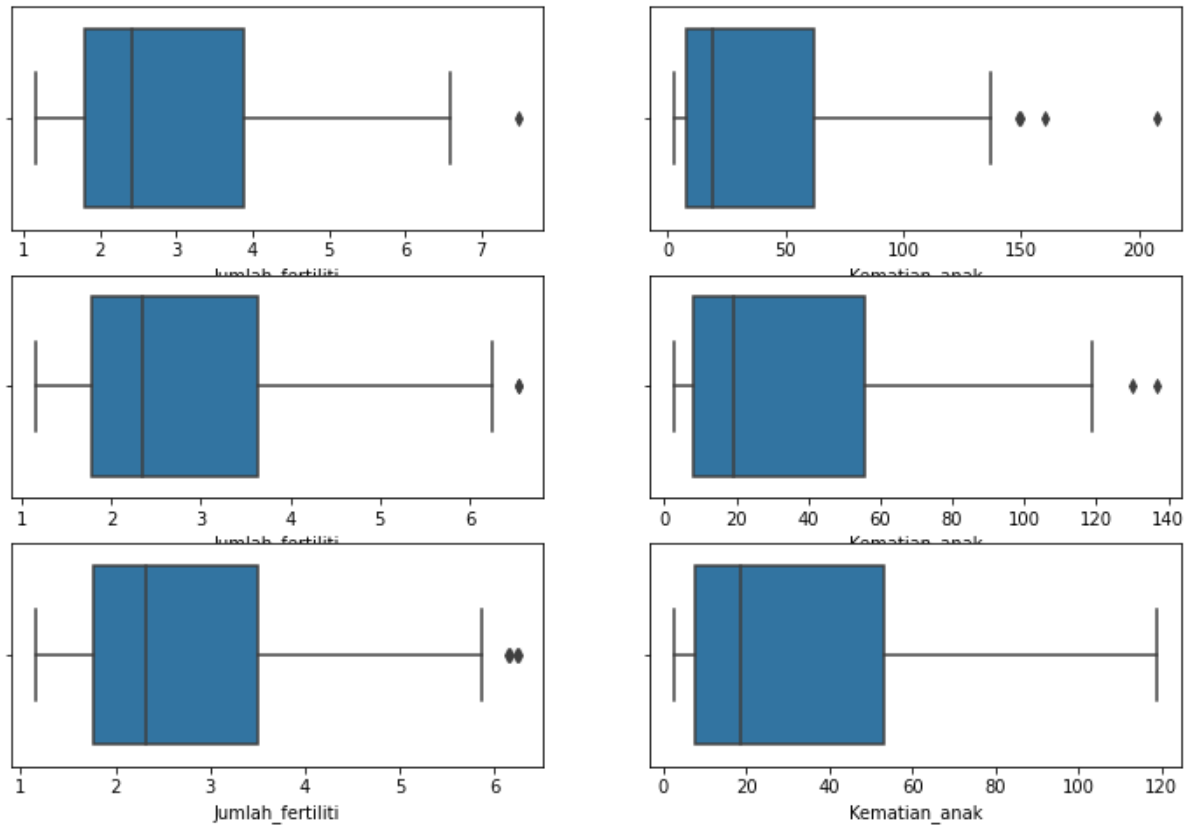
Korelasi antar fiturnya divisualisasikan dalam bentuk heatmap. Berdasarkan heatmap dapat dilihat bahwa kolom GDPperkapita dan Pendapatan memiliki korelasi yang paling tinggi disusul oleh Jumlah_fertiliti dengan Kematian_anak. Pemilihan kedua korelasi tersebut juga berdasarkan ciri-ciri negara berkembang yaitu negara yang memiliki pendapatan & GDPperkapita rendah serta Jumlah_fertiliti dan Kematian_anak yang tinggi [1]. Kita akan menggunakan kedua korelasi tersebut untuk menghasilkan clusteringnya. Selain itu, akan digunakan juga korelasi yang kecil, pada project ini yang digunakan yaitu Kesehatan dengan Inflasi.

Dari hubungan-hubungan fitur yang telah ditentukan maka hal selanjutnya adalah mulai memproses data menjadi suatu cluster. Untuk memproses data menjadi suatu cluster maka perlu terlebih dahulu untuk menghilangkan outliernya. Outlier disini dihilangkan agar saat scaling data tidak terbentuk jarak yang jauh. Untuk metode penanganan outliernya yaitu dengan menghilangkannya pada data melalui indikator IQR. Dalam pengerjaan project ini, penerapan outlier tidak hanya dilakukan sekali namun bisa dalam 3-5 kali iterasi. Hal tersebut dikarenakan setelah penerapan masih terdapat outlier yang tersisa dari fitur. Outlier tersebut

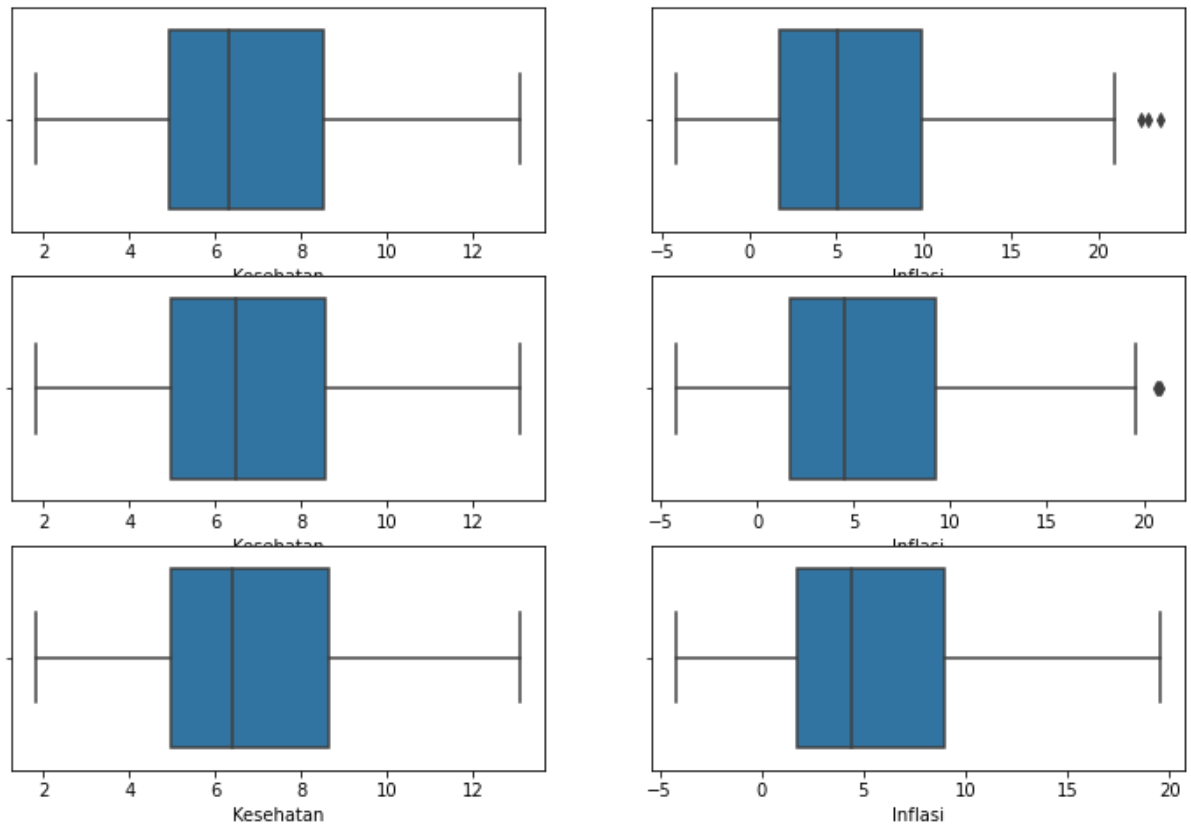
dihilangkan hingga mencapai hasil yang sedikit mungkin. Berikut merupakan visualisasi dari proses menghilangkan outlier untuk setiap hubungan cluster.



Gambar 9. Penanganan Outlier GDPperkapita & Pendapatan

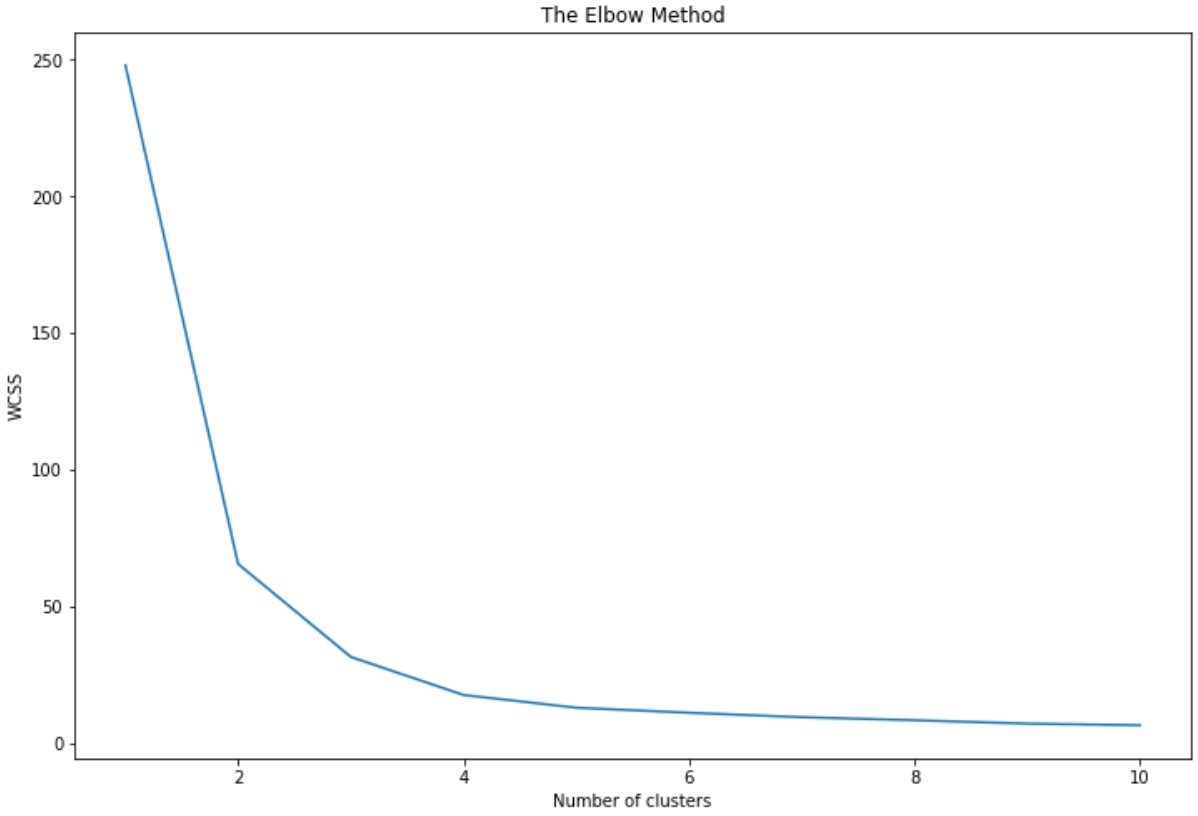


Gambar 10. Penangan Outlier Jumlah_fertiliti & Kematian_anak

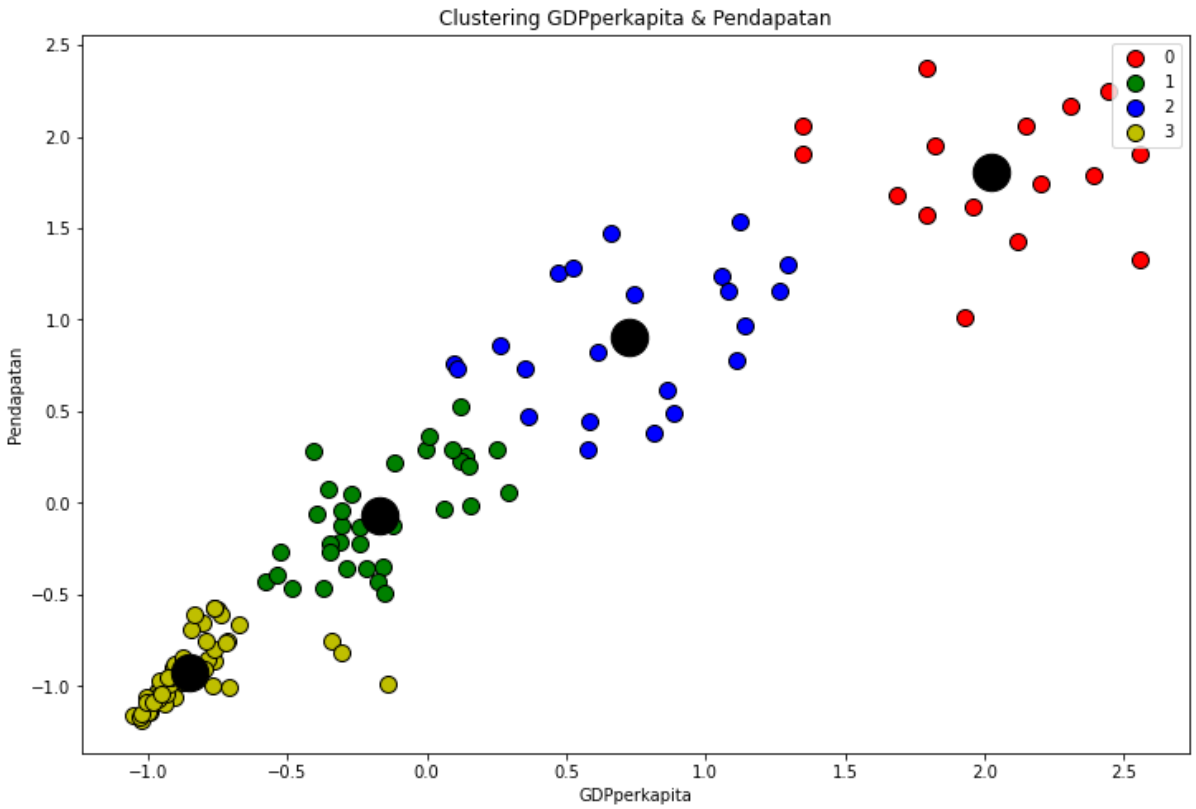


Gambar 11. Penanganan Outlier Kesehatan & Inflasi

Setelah outlier ditangani, maka selanjutnya adalah melakukan scaling data. Scaling yang digunakan yaitu Standard Scaler dari library *sklearn.* Setelah melakukan scaling selanjutnya adalah menggunakan elbow method untuk menentukan nilai k yang digunakan pada K-Mean clustering.

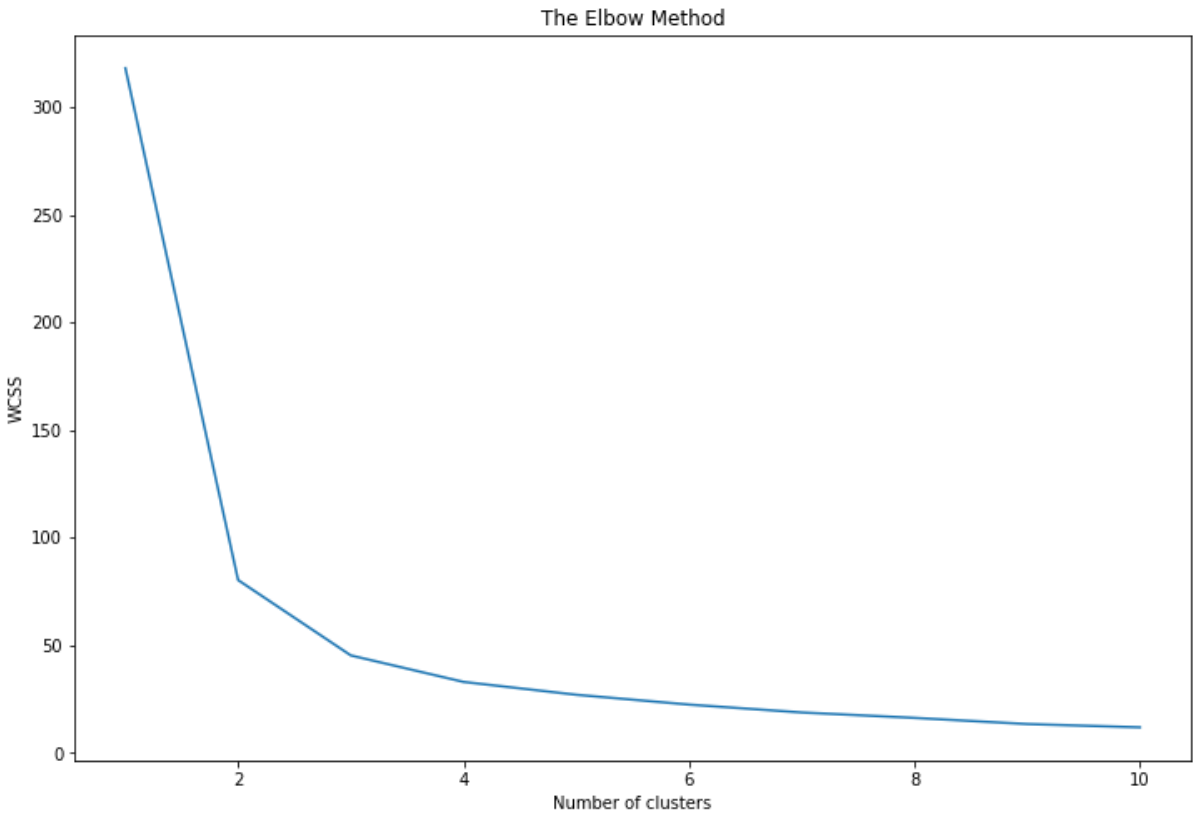


Gambar 12. Elbow Method GDPperkapita & Pendapatan

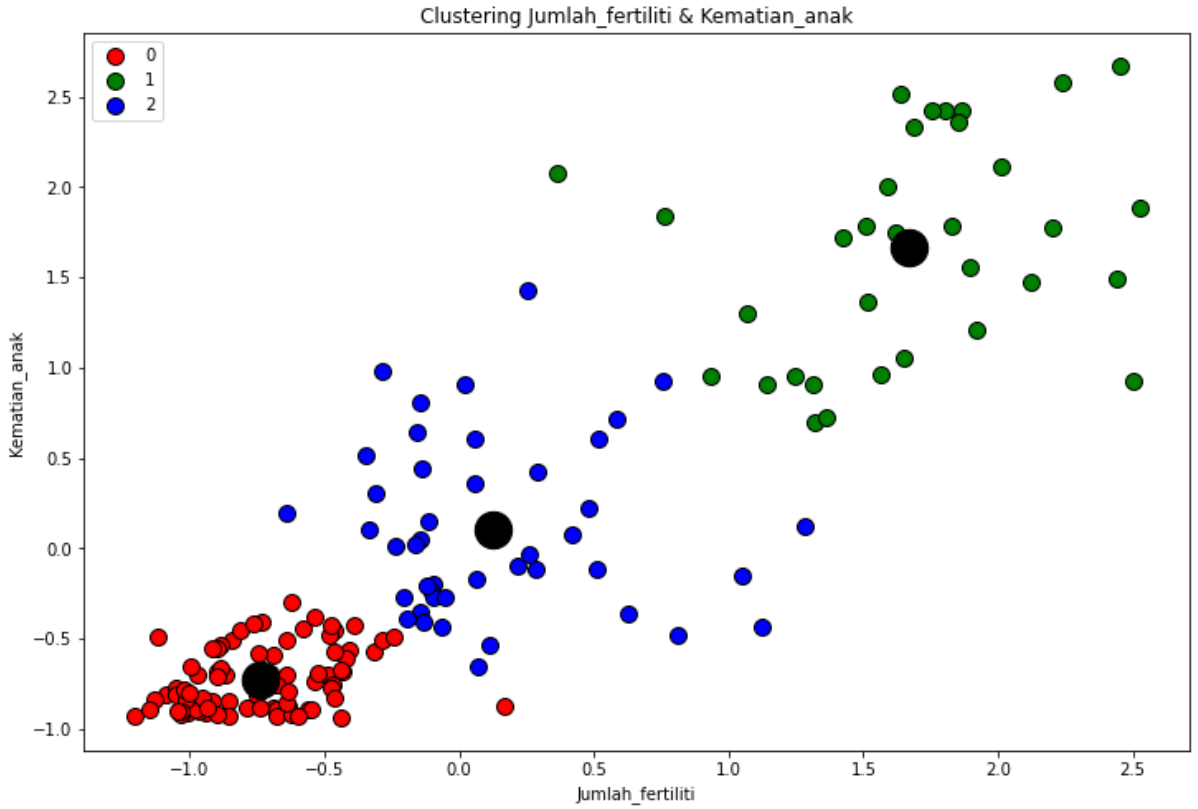


Gambar 13. Hasil Clustering GDPperkapita & Pendapatan

Pada fitur GDPperkapita & Pendapatan nilai k yang digunakan yaitu 4 sehingga dihasilkan clustering seperti Gambar 13.

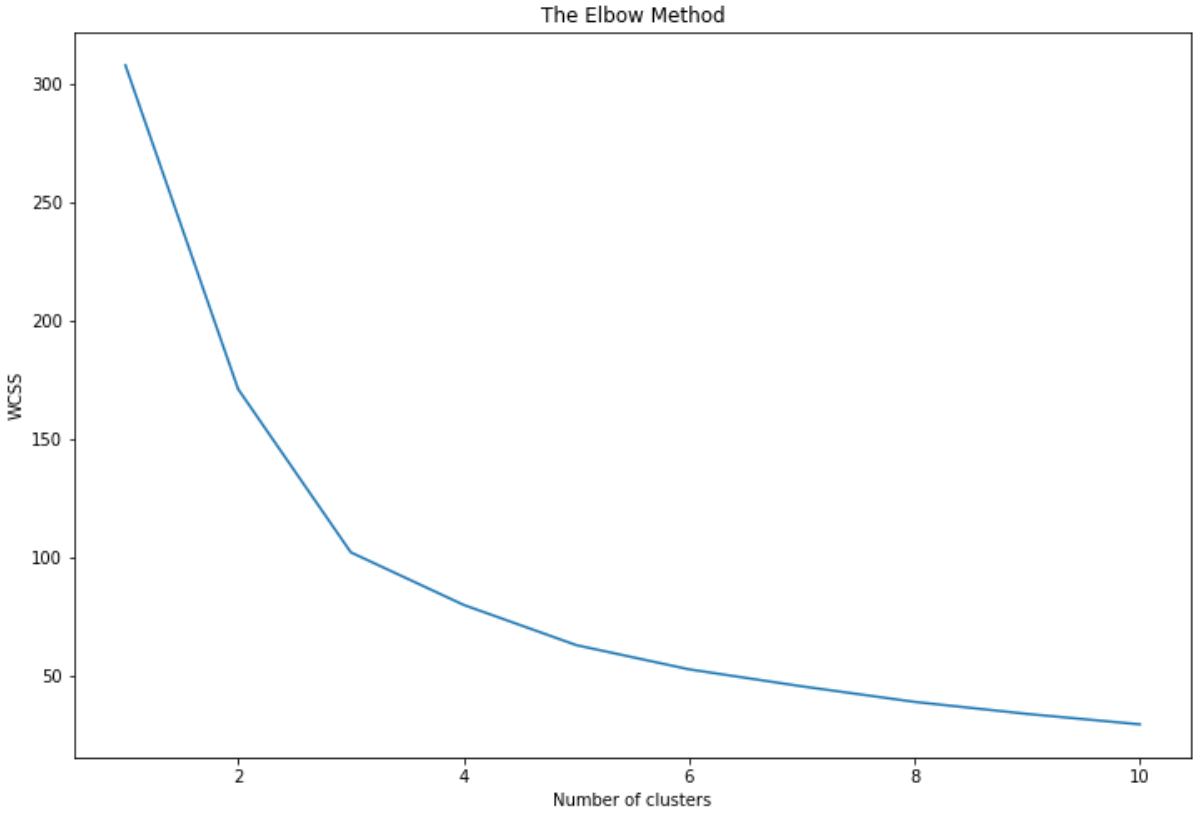


Gambar 14. Elbow Method Jumlah_fertiliti & Kematian_anak

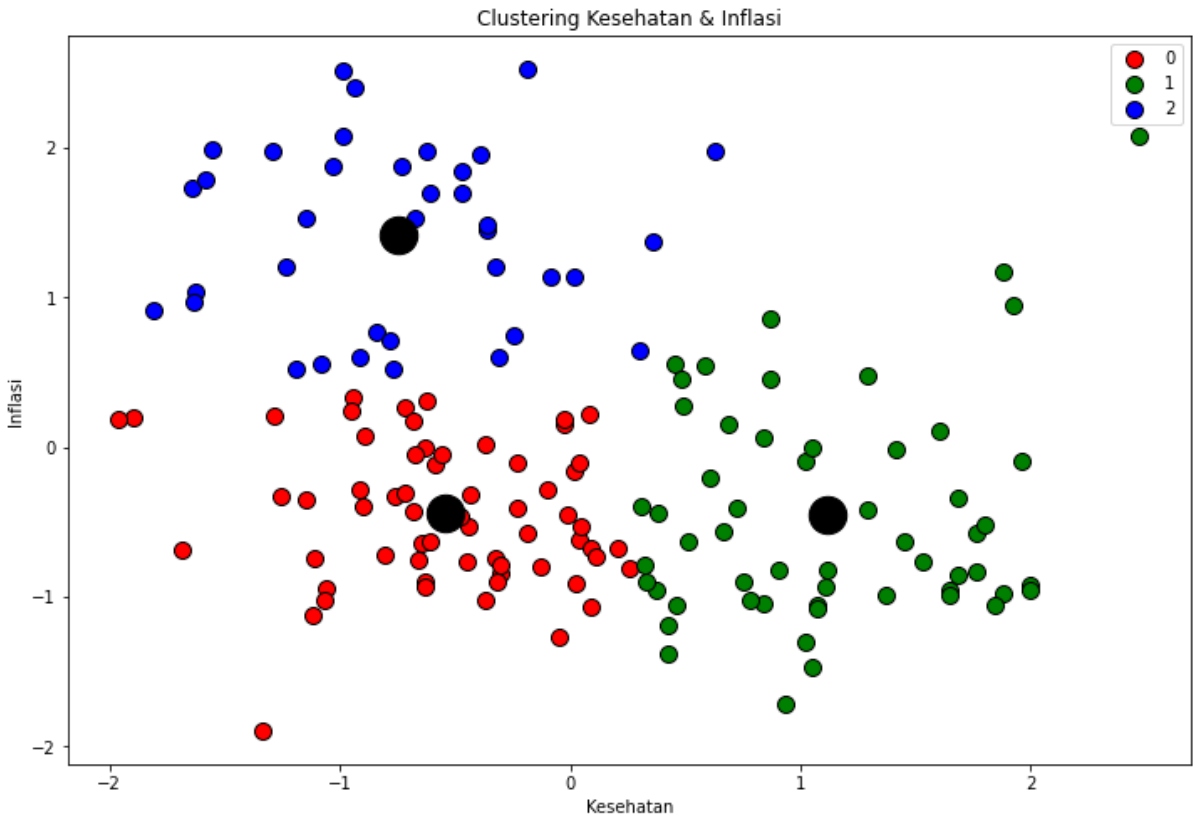


Gambar 15. Hasil Clustering Jumlah_fertiliti & Kematian_anak

Pada fitur Jumlah_fertiliti & Kematian_anak nilai k yang digunakan yaitu 3 sehingga dihasilkan clustering seperti Gambar 15.



Gambar 16. Elbow Method Kesehatan & Inflasi



Gambar 17. Hasil Clustering Kesehatan & Inflasi

Pada fitur Kesehatan & Inflasi nilai k yang digunakan yaitu 3 sehingga dihasilkan clustering seperti Gambar 17.

Berdasarkan hasil clustering yang didapatkan maka filtering yang dilakukan yaitu pada negara yang memiliki GDPperkapita dan Pendapatan rendah (label 3 pada Gambar 13), memiliki Jumlah_fertiliti dan Kematian_anak yang tinggi (label 1 pada Gambar 15), serta memiliki Kesehatan yang rendah dan Inflasi yang tinggi (label 2 pada Gambar 17). Kemudian pada project ini ditambahkan juga filter Impor lebih besar dari Ekspor karena salah satu ciri negara berkembang yaitu memiliki impor yang lebih besar dari ekspor [3]. Selain itu, dilakukan juga filter yaitu negara yang diberikan hanya negara yang digolongkan oleh Bank Dunia (*World Bank*) sebagai negara yang berpendapatan rendah.

Kesimpulan

Berdasarkan penyaringan-penyaringan yang dilakukan maka didapatkan hasil 6 buah negara yaitu, **Argentina**, **Eritrea**, **Guinea**, **Madagascar**, **Malawi**, dan **Tanzania**. Keenam negara inilah yang disarankan untuk mendapatkan bantuan dari HELP Internasional berdasarkan kriteria-kriteria yang ada.

	Negara	Kematian_anak	Ekspor	Kesehatan	Impor	Pendapatan	Inflasi	Harapan_hidup	Jumlah_fertiliti	GDPperkapita
0	Afghanistan	90.2	10.00	7.58	44.9	1610	9.44	56.2	5.82	553
1	Eritrea	55.2	4.79	2.66	23.3	1420	11.60	61.7	4.61	482
2	Guinea	109.0	30.30	4.93	43.2	1190	16.10	58.0	5.34	648
3	Madagascar	62.2	25.00	3.77	43.0	1390	8.79	60.8	4.60	413
4	Malawi	90.5	22.80	6.59	34.9	1030	12.10	53.1	5.31	459
5	Tanzania	71.9	18.70	6.01	29.1	2090	9.25	59.3	5.43	702

Gambar 18. Saran Negara

Referensi

[1] https://tirto.id/apa-saja-karakteristik-negara-maju-dan-negara-berkembang-f75N

[2] "How we Classify Countries"

[3] https://id.wikipedia.org/wiki/Negara_berkembang#cite_note-16