Clustering Countries By Using K-Means

HELP Internasional

Final Project

Dibuat Oleh:

Jonathan Adithya

TANGERANG

2021

DAFTAR ISI

BAB I
PENDAHULUAN

1.1 Latar Belakang

Di dunia ini selalu ada kesenjangan sosial maupun kesenjangan ekonomi antara negara maju dengan negara berkembang. Masih ada banyak negara-negara yang sebagian besar masyarakatnya berada dibawah garis kemiskinan, kurangnya tingkat pendidikan, kecenderungkan kenaikan harga yang terus menerus meningkat, dan selalu bertambah pengangguran. Faktor-faktor tersebut merupakan faktor-faktor yang menyebabkan kemiskinan.

Masalah kemiskinan belum dapat teratasi secara optimal juga diakibatkan karena penyebab kemiskinan yang beragam dan kompleks. Faktor yang mempengaruhi kemiskinan bukan hanya faktor ekonomi saja. Ada banyak faktor eksternal lainnya seperti tingkat pendidikan, lowongan pekerjaan, tenaga kerja, dan masih banyak lagi.

Terdapat sebuah lembaga internasional yang bernama HELP International yang memiliki visi misi untuk memerangi kemiskinan, menyediakan fasilitas, dan bantuan dasar bagi masyarakat di negara-negara terbelakang saat terjadi bencana alam. Untuk saat ini, organisasi HELP International telah berhasil mengumpulkan dana sebanyak $10 juta. CEO LSM perlu mengambil keputusan bagaimana caranya agar dana yang telah berhasil dikumpulkan tersebut bisa tersalurkan secara efektif ke negara-negara yang paling membutuhkan bantuan.

1.2 Tujuan

- Mengetahui negara apa saja yang sangat membutuhkan bantuan agar dapat membantu CEO LSM dalam mengambil keputusan
- Mengklasifikasi negara-negara yang ada ke dalam kelompok negara maju, negara menengah, atau negara berkembang dengan menggunakan *K-Means Clustering*.

BAB II
LANDASAN TEORI

2.1 Klasifikasi Negara

2.1.1 Negara Maju

Negara maju adalah istilah yang diberikan untuk negara-negara yang memiliki taraf hidup yang relatif tinggi baik dari segi teknologi, ekonomi, maupun sosialnya. Pembangunan mereka itu bisa dilakukan secara merata, sehingga tingkat kesejahteraan juga tinggi.

Pendapatan penduduk per kapita di negara maju biasanya di atas $2,000. Negara maju juga biasanya memiliki infrastruktur yang tertata lebih rapi dan canggih. Hal ini tak dipungkiri karena perilaku masyarakatnya juga lebih mudah diatur dan memiliki tingkat Pendidikan yang lebih tinggi. Contohnya kondisi jalanan di negara Jepang yang bersih di mana tidak ada sampah-sampah yang berserakan. Hal itu juga didukung dari kebiasaan masyarakatnya yang biasanya lebih profesional dan memiliki kesadaran diri.

Ciri-ciri negara maju:

- Pendapatan rata-rata tinggi
- Harapan hidup tinggi
- Pertumbuhan penduduk rendah (Angka kelahiran rendah, Angka kematian rendah)
- Tingkat pendidikan yang tinggi
- Memiliki fasilitas yang memadai
- Berbasis industri dan jasa ekonomi
- Sebagian besar penduduk tinggal di daerah perkotaan
- Angka kematian bayi kecil

2.1.2 Negara Berkembang

Negara berkembang adalah negara yang kesejahteraan masyarakatnya masih rendah. Mereka memiliki tenaga ahli yang masih kurang karena tingkat pendidikan masih rendah. Hal ini tentu akan menyulitkan untuk bersaing secara internasional dengan negara maju. Pengelolaan sumber daya, baik sumber daya alam maupun sumber daya manusianya juga belum maksimal.

Hal ini disebabkan karena modal dalam perihal pertumbuhan ekonominya juga relatif lebih kecil dan kebanyakan justru didapat dari pinjaman berbunga besar. Tentu saja hal ini sangat berisiko yang dapat membuat negara memiliki kerugian yang semakin besar dan bisa menghambat pertumbuhan negaranya sendiri. Pendapatan penduduk per kapita untuk negara berkembang biasanya di bawah $2,000.

Karena negara bergerak cukup lambat, akhirnya perkembangan infrastruktur juga melambat. Jika hal ini terus dibiarkan, bisa menyebabkan terus berkurangnya daya saing dan produktivitas. Jadi, produk bisa kalah saing di luar negeri dan hanya mampu diperjualbelikan di dalam negeri.

Ciri-ciri negara berkembang:

- Berbasis pertanian (ekspor bahan dasar pangan).
- Memiliki banyak masalah, antara lain:
    - Tingginya tingkat pengangguran.
    - Jumlah ledakan populasi tinggi.
    - Produktivitas Rendah/Survival rate yang rendah.
    - Ketergantungan pada negara-negara lain yang tinggi dan rentan.
- Sumber daya alam mentah (belum di proses).
- Kurangnya modal.
- Ketergantungan pada produk industri impor, dan ekspor produk bahan baku (industri dan pertanian).
- Warga berpenghasilan rendah, dengan GNP <US $ 1.000.
- Penguasaan rendah ilmu pengetahuan dan teknologi.
- Tingkat pendidikan yang rendah.

2.2 Machine Learning

Machine learning (ML) adalah mesin yang dikembangkan untuk bisa belajar dengan sendirinya tanpa arahan dari penggunanya. Pembelajaran mesin dikembangkan berdasarkan disiplin ilmu lainnya seperti statistika, matematika dan data mining sehingga mesin dapat belajar dengan menganalisa data tanpa perlu di program ulang atau diperintah. *Machine Learning* juga dapat mempelajari data yang ada dan data yang ia peroleh sehingga bisa melakukan tugas tertentu atau dapat menginferensikan sesuatu.

Metode belajar mesin dibagi menjadi 2, yaitu *Supervised Learning* dan *Unsupervised Learning*. *Supervised learning* adalah metode belajar mesin yang membutuhkan training data sets untuk belajar, melainkan *Unsupervised Learning* adalah metode untuk mengelompokkan sebuah no-labeled data. Diharapkan dengan metode Unsupervised Learning ini, dapat membantu untuk menemukan suatu pola atau struktur tertentu yang ada di dataset.

Salah satu metode *Unsupervised Learning* adalah *K-Means Clustering*. *K-Means Clustering* ini menggunakan metode centroid dan menghitung jarak dari setiap data yang ada ke centroidnya dan *assign* data tersebut ke centroid masing-masing (jarak yang paling dekat). Namun salah satu kerugian dari K-Means Clustering adalah metode tersebut tidak cukup baik untuk data yang memiliki pencilan data.

K-Means merupakan salah satu metode pengelompokan data non-hirarki yang berusaha mempartisi data yang ada ke dalam bentuk dua atau lebih cluster. Metode ini mempartisi data ke dalam cluster sehingga data berkarakteristik sama dimasukkan ke dalam satu cluster yang sama dan data yang berkarakteristik

berbeda dikelompokkan ke dalam cluster yang lain. Adapun tujuan pengelompokan data ini adalah untuk meminimalkan fungsi objektif yang diatur dalam proses pengelompokan, yang pada umumnya berusaha meminimalkan variasi di dalam suatu cluster dan memaksimalkan variasi antar cluster. (Supranto, 2004)

2.3 *Feature Scaling*

Feature Scaling adalah teknik untuk menstandarisasi fitur/variabel yang ada dalam data dalam rentang yang sama. *Feature Scaling* dilakukan saat data *pre-processing* untuk menangani besaran atau nilai atau unit yang sangat bervariasi. Jika *Feature Scaling* tidak dilakukan, algoritme *Machine Learning* cenderung membuat prediksi yang kurang akurat.

Terdapat 2 tipe teknik *Feature Scaling* yang cukup populer, yaitu *Standard Scaling* dan *Normalization*. Standardisasi adalah *Feature Scaling* di mana nilai dipusatkan di sekitar mean dengan *standard deviation* dan *variance* yang bernilai satu. Artinya mean dari fitur/variabel tersebut menjadi nol dan distribusi memiliki *standard deviation* dan *variance* satuan. Kunci terbesar dari Standarisasi adalah sifatnya yang tidak memiliki batas, tidak seperti Normalisasi, sehingga data tersebut tidak akan dipengaruhi oleh *outliers*.

Tipe teknik *Feature Scaling* yang lain adalah Normalisasi. Normalisasi adalah teknik penskalaan di mana nilai digeser dan diskalakan kembali sehingga akhirnya berada dalam rentang 0 dan 1. Teknik ini juga dikenal sebagai penskalaan Min-Max.

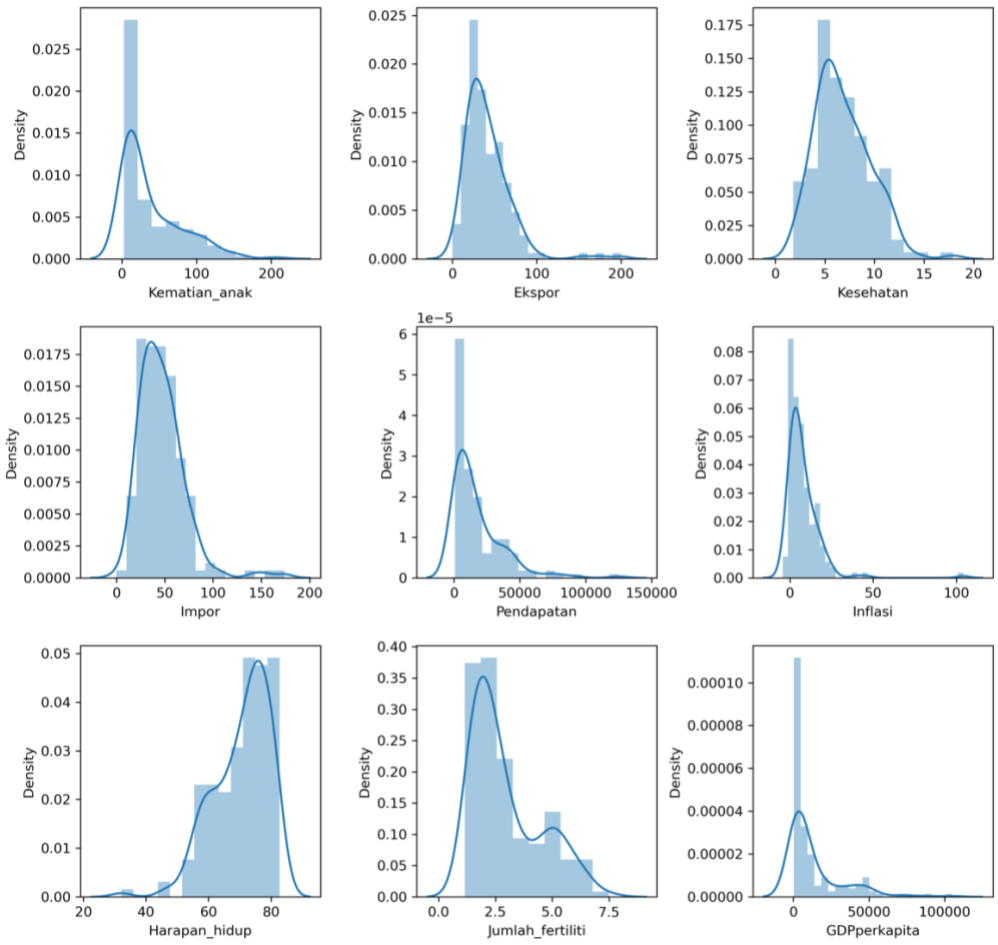
BAB III
ANALISIS DATA

3.1 Penjelasan Dataset

Dataset yang telah disediakan adalah data statistik negara seperti *Gross Domestic Product* dari masing-masing negara, harapan hidup, jumlah fertilitas, angka kematian anak, dan masih banyak lagi. Dataset tersebut memiliki 9 indikator atau variabel utama dan 167 negara yang berbeda. Semua variabel tersebut memiliki tipe data numerik. Berikut adalah penjelasan dari masing-masing variabel/indikator :

Indikator	Penjelasan
Kematian Anak	Kematian anak balita (bawah lima tahun) per 1000 kelahiran
Ekspor	Ekspor barang jasa per kapita
Kesehatan	Total pengeluaran dalam bidang Kesehatan per kapita
Impor	Impor barang dan jasa per kapita
Pendapatan	Penghasilan bersih per orang
Inflasi	Pengukuran tingkat pertumbuhan tahunan dari Total GDP
Harapan Hidup	Jumlah umur rata-rata seorang anak yang baru lahir akan hidup
Jumlah Fertilitas	Jumlah anak yang akan lahir dari setiap wanita
GDP per kapita	Besarnya pendapatan rata-rata semua penduduk di suatu negara.

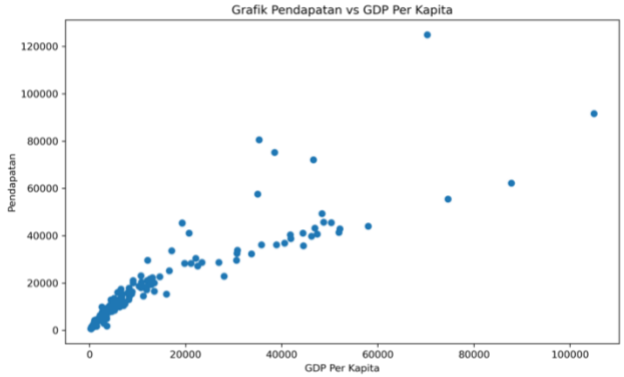
3.2 Univariate Analisis



Figur di atas merupakan distribution plot dari dataset dari HELP International. Figur tersebut menggambarkan sebaran data dari masing-masing indikator. Sebagian besar dari data tersebut memiliki sebaran data yang *positive skewed*, hanya indicator harapan hidup yang sebaran datanya *negative skewed*. Dari figure di atas, dapat dilihat bahwa Sebagian besar negara memiliki angka kematian anak sekitar 0 – 5, dan setelah itu jumlah negaranya menurun pesat. Dapat juga dilihat bahwa ada pula negara yang memiliki angka kematian anak lebih dari 200. Untuk kesehatan, sebaran datanya cukup normal dan dapat dilihat titik maksimumnya berada di 5 – 6. Pada bagian ekspor dan impor, sebaran datanya mirip, sehingga dapat dilihat kedua indikator tersebut memiliki korelasi yang cukup tinggi dan dapat dilihat juga bahwa terdapat pencilan data yang cukup jauh. Untuk indikator pendapatan dan GDP per kapita, jika dilihat dari figurnya, mereka memiliki sebaran data yang cukup jauh. Oleh sebab itu kedua indikator tersebut memiliki banyak pencilan data (*outliers*). Indikator jumlah fertilitas menunjukkan jumlah anak yang akan lahir dari setiap wanita. Data tersebut memiliki rentang dari 1 – 8 anak yang akan lahir dari setiap wanita.

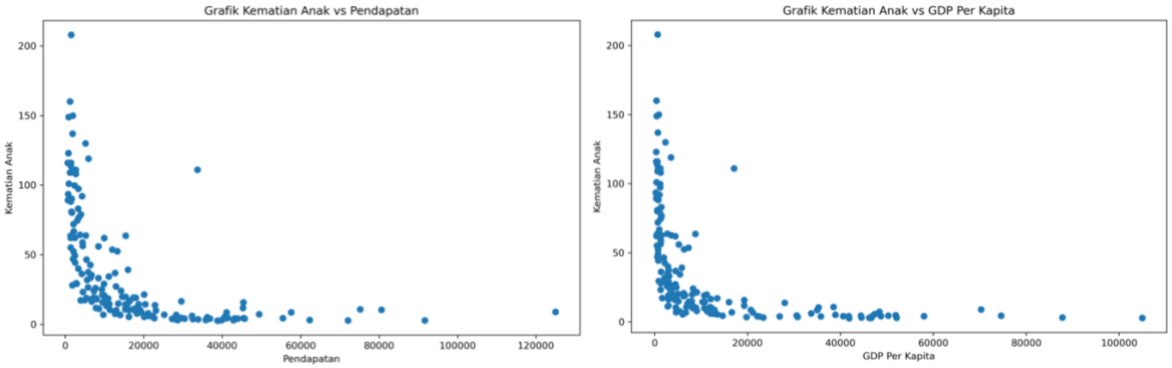
3.3 Bivariate Analisis

3.3.1 Pendapatan dan GDP per Kapita



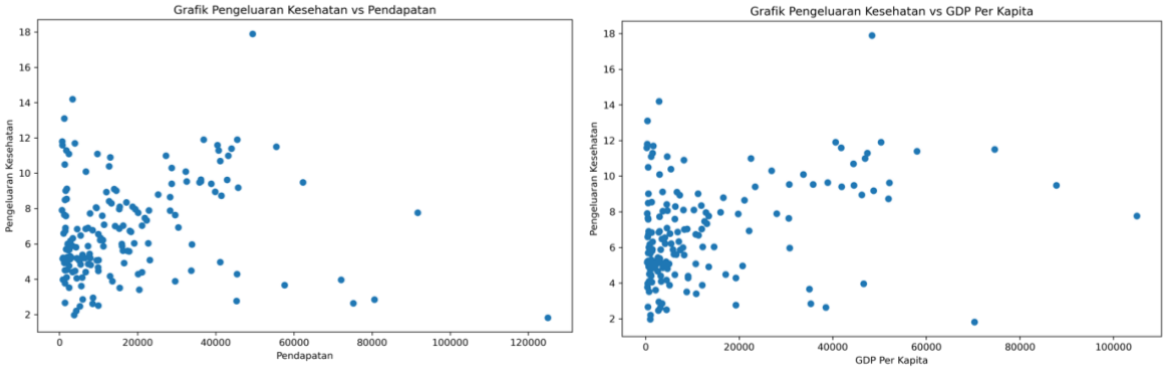
Grafik di atas merupakan grafik antara pedapatan bersih per orang dengan GDP per kapita. Dapat dilihat dari grafik di atas, kedua indikator tersebut memiliki korelasi yang erat dan berbanding lurus. Jika GDP perkapita dari suatu negara, maka pendapatan bersih per orang juga naik. Data tersebut juga terlihat memusat pada GDP per kapita dan pendapatan bersih per orang kurang dari 20.000. Seperti yang ada pada landasan teorinya, negara berkembang memiiki GDP per kapita dan juga pendapatan bersih per orang yang rendah. Sedangkan negara maju memiliki GDP dan pendapatan yang tinggi.

3.3.2 Angka Kematian Anak dan Pendapatan



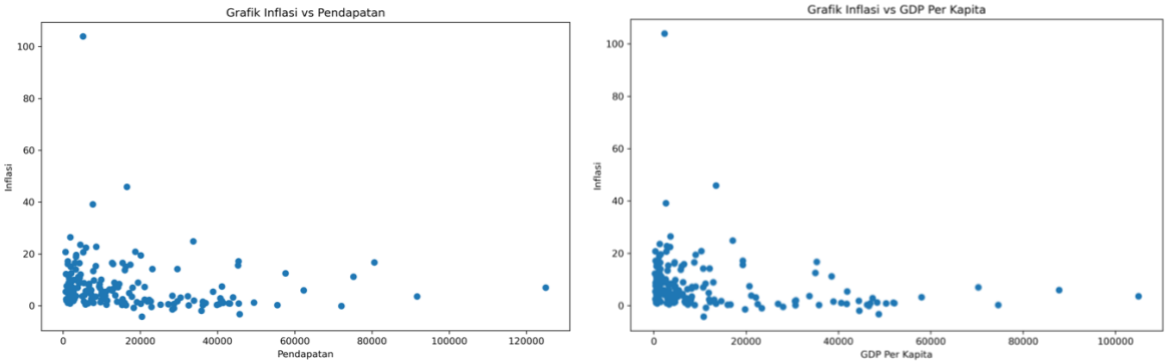
Grafik di atas merupakan grafik antara kematian anak dengan pendapatan bersih per orang dan juga GDP per kapita. Dari grafik di atas, dapat disimpulkan bahwa jika pendapatan per orangnya tinggi, maka kemungkinan besar angka kematian anaknya akan rendah. Di sisi lain, jika pendapatannya rendah maka kemungkinan untuk kematian anak dibawah lima tahun semakin besar. Dari landasar teori yang sudah ada, negara berkembang memiliki pendapatan yang cukup rendah dan angka kematian yang tinggi. Sedangkan negara maju memiliki angka kematian yang rendah dan pendapatan yang tinggi

3.3.3 Pengeluaran Kesehatan dan Pendapatan



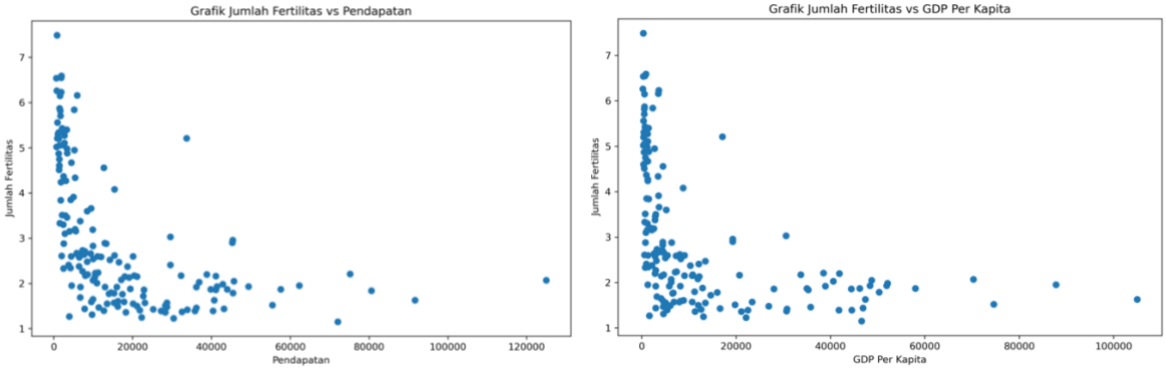
Grafik di atas merupakan grafik antara pengeluaran dalam bidang kesehatan per kapita dengan pendapatan bersih per orang dan GDP per kapita. Dilihat dari grafik di atas, kedua indikator tersebut tidak memiliki korelasi yang cukup kuat. Sebagian besar datanya mengumpul pada Pendapatan/GDP per kapita kurang dari 10.000 dan pengeluaran kesehatan sekitar 4 – 6. Jika data tersebut dimasukkan ke dalam klasifikasi *K-Means Clustering*, maka data tersebut belum tentu memberikan data yang akurat. Hal ini dikarenakan pengeluaran kesehatan merupakan suatu indikator yang cukup ambigu, yang artinya tidak diketahui secara jelas apakah pengeluaran yang besar itu berasal dari perlengkapan medisnya yang sudah canggih atau gaya hidupnya yang tidak baik dan penanganannya yang kurang baik.

3.3.4 Inflasi dan Pendapatan



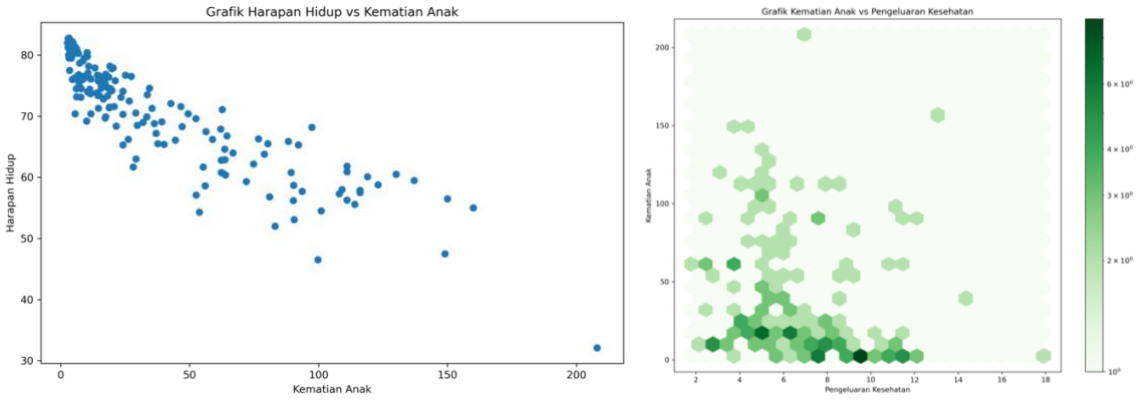
Grafik di atas merupakan grafik antara laju inflasi terhadap pendapatan bersih per orang dan GDP per kapita. Grafik di atas juga menunjukkan korelasi yang lemah antar kedua indikator tersebut. Sama seperti pengeluaran dalam bidang kesehatan, indikator inflasi ini juga tidak dapat dimasukkan ke dalam klasifikasi *K-Means Clustering*, karena interpretasinya yang ambigu. Tidak semua negara maju memiliki laju inflasi yang selalu rendah, dan tidak semua negara berkembang memiliki laju inflasi yang tinggi. Laju inflasi itu bersifat tidak menentu. Jika terdapat suatu peristiwa yang sedang terjadi di suatu negara, laju inflasinya bisa berubah dengan cepat. Oleh karena itu, laju inflasi tidak dapat dimasukkan ke dalam salah satu kategori negara maju dan berkembang.

3.3.5 Jumlah Fertilitas dan Pendapatan



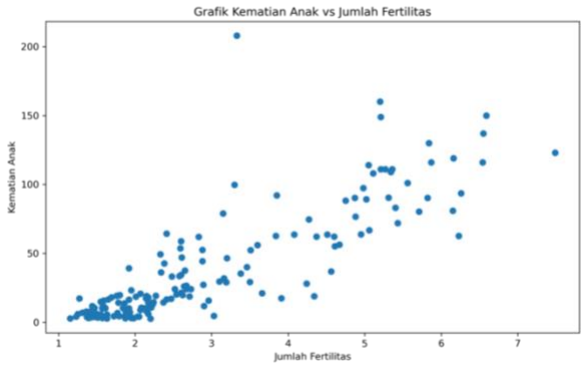
Grafik di atas merupakan grafik antara jumlah fertilitas suatu negara terhadap pendapatan bersih per orang dan GDP per kapita. Grafik tersebut juga menunjukkan korelasi yang lemah. Walaupun begitu, grafik di atas cocok untuk di klasifikasi menggunakan *K-Means Clustering*. Hal ini dikarenakan grafik tersebut membentuk kelompok-kelompok. Ada yang jumlah fertilitas tinggi dan pendapatan rendah, ada yang jumlah fertilitas rendah dan pendapatan rendah, dan ada juga yang jumlah fertilitas rendah dan pendapatan tinggi. Indikator tersebut juga cocok untuk mengkategorikan sebuah negara, karena negara berkembang memiliki pendapatan yang rendah dan angka kelahiran yang tinggi, sedangkan negara maju memiliki pendapatan yang tinggi dan angka kelahiran yang rendah. Hal ini menunjukkan tingkat edukasi dari negara tersebut yang baik.

3.3.6 Angka Kematian Anak, Harapan Hidup, dan Pengeluaran Kesehatan



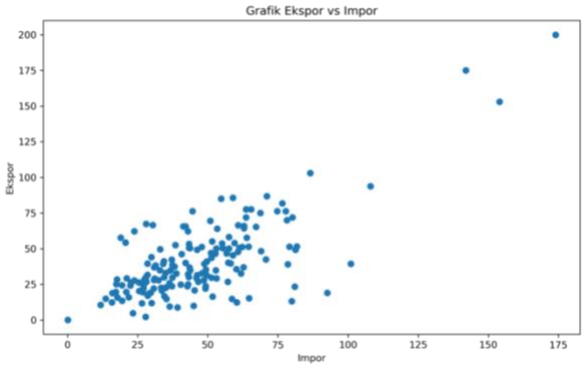
Grafik di atas merupakan grafik antara harapan hidup dengan angka kematian anak dan angka kematian anak dengan pengeluaran kesehatan. Seperti yang sudah dijelaskan sebelumnya, pengeluaran kesehatan tidak cocok untuk dijadikan sebuah kategori negara maju dan berkembang, namun harapan hidup memiliki kaitan erat dengan negara maju dan berkembang. Negara maju tentu saja memiliki harapan hidup yang tinggi dan kematian anak yang rendah, karena pendidikan yang tinggi dan teknologi yang canggih. Di sisi lain, negara berkembang memiliki harapan hidup yang rendah dan angka kematian anak yang tinggi, karena pendidikannya yang belum maju mengakibatkan gaya hidupnya yang tidak sesuai, penanganan medis yang kurang, dan fasilitas yang kurang memadai.

3.3.7 Angka Kematian Anak dan Jumlah Fertilitas



Grafik di atas merupakan grafik antara angka kematian anak dengan jumlah anak yang lahir dari setiap wanita. Grafik tersebut juga menunjukkan korelasi antar kedua indikator tersebut yang cukup tinggi. Itu berarti jika jumlah fertilitasnya meningkat, maka kemungkinan besar angka kematian anaknya juga meningkat. Grafik tersebut juga cocok untuk dijadikan sebagai kategori negara maju dan berkembang. Negara maju memiliki jumlah fertilitas yang rendah menunjukkan tingkat pendidikannya yang tinggi, dan angka kematian anak yang rendah menunjukkan penanganan medis yang cukup baik. Di sisi lain negara berkembang memiliki jumlah fertilitas yang tinggi dan kematian anak yang tinggi dikarenakan oleh gaya hidup yang atau penanganan medis yang kurang baik.

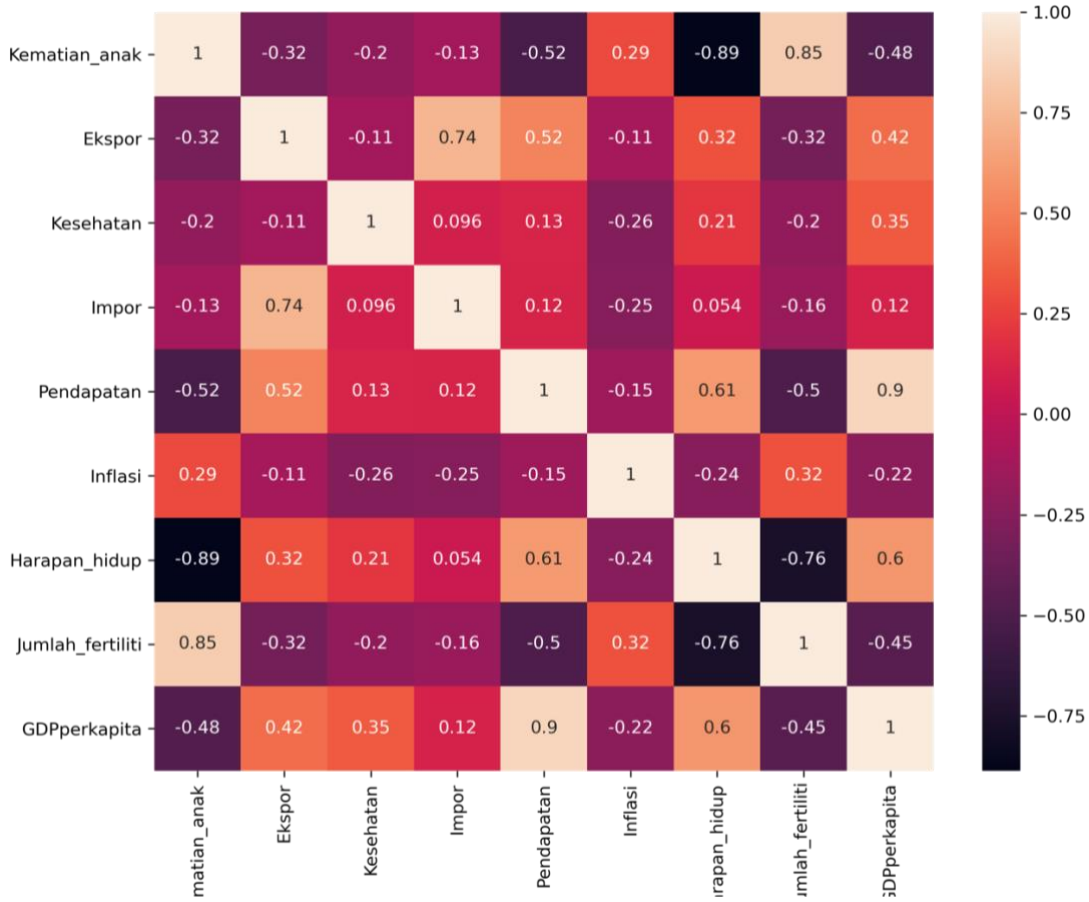
3.3.8 Ekspor dan Impor



Grafik di atas merupakan grafik antara kegiatan ekspor jasa atau barang per kapita dengan impor jasa atau barang per kapita. Grafik tersebut menunjukkan korelasi yang cukup baik antara ekspor dan impor. Walaupun begitu, indikator tersebut tidak cocok untuk dijadikan kategori negara maju dan berkembang karena interpretasinya ambigu. Artinya data tersebut tidak menunjukkan negara maju dan berkembang karena tidak dijelaskan secara spesifik barang atau jasa seperti apa yang di ekspor atau impor. Negara berkembang juga bisa memiliki kegiatan ekspor yang tinggi, namun barang atau jasa yang diekspor itu berbeda dengan negara maju. Negara berkembang biasanya mengekspor bahan-bahan pokok seperti hasil pertanian. Sedangkan negara-negara maju biasanya mengekspor barang-barang perindustrian.

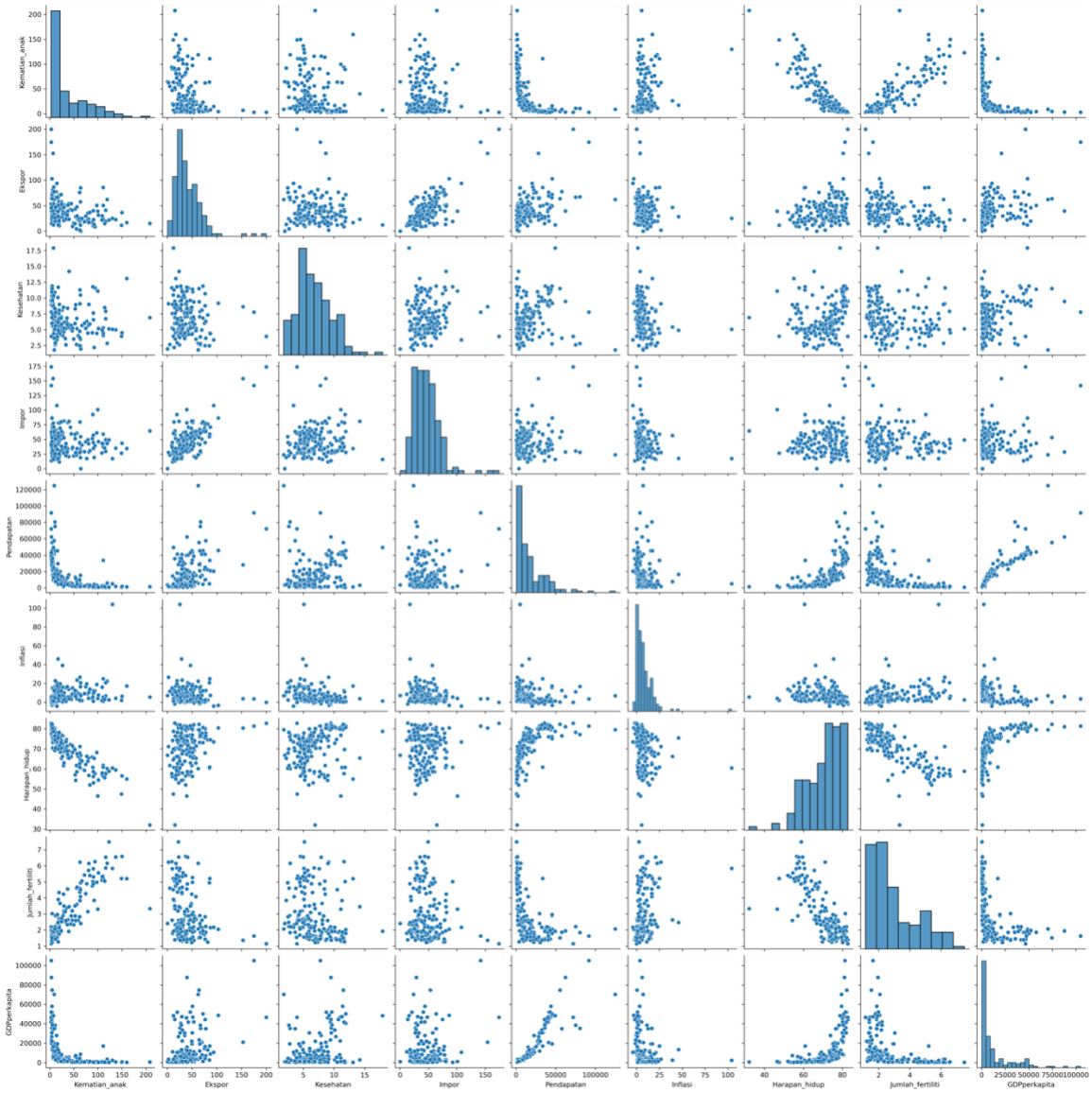
3.4 Multivariate

3.4.1 Heatmap



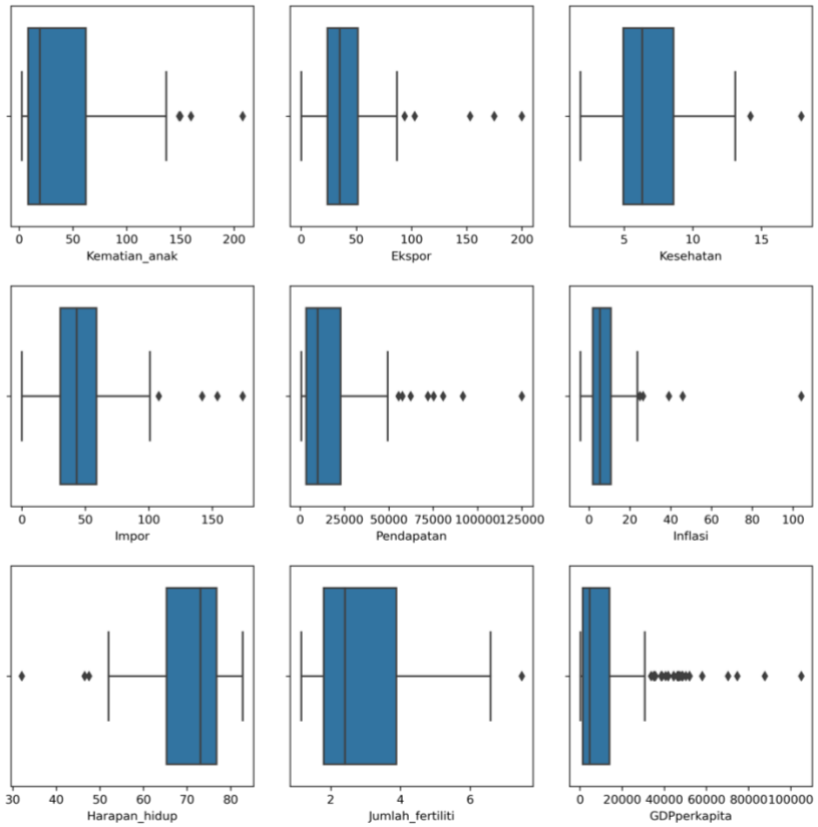
Figur di atas merupakan *heatmap* dari semua indikator dari dataset yang ada. *Heatmap* tersebut menjelaskan nilai *pearson correlation* antar indikator-indikatornya. Dapat dilihat dari *heatmap* di atas, ada beberapa pasangan indikator yang memiliki nilai korelasi yang tinggi, antara lain, harapan hidup dengan angka kematian anak, jumlah fertilitas dengan angka kematian anak, impor dan ekspor, GDP per kapita dengan pendapatan bersih per orang, dan jumlah fertilitas dengan harapan hidup. Namun, seperti yang sudah dijelaskan sebelumnya, nilai korelasi ini tidak menentukan indikator mana yang cocok menjadi kategorinya. Ada indikator yang memiliki nilai korelasi yang tinggi tetapi tidak cocok untuk menjadi kategorinya, dan ada juga indikator yang memiliki nilai korelasi yang rendah namun cocok untuk menjadi kategorinya.

3.4.2 Pairplot



Figur di atas merupakan hasil *pairplot* dari semua indikator yang ada di dataset. *Pairplot* dan *heatmap* ini memiliki kemiripan, yaitu sama-sama merupakan *multivariate*. Namun dalam *pairplot*, kita dapat melihat secara langsung ketika indikator tersebut diplot secara *scatter*. Sehingga kita tidak hanya melihat dari *correlationnya* saja tetapi juga grafiknya secara langsung, karena tidak semua kategori dapat dilihat dari *pearson correlationnya*.

3.5 *Outliers Treatment*

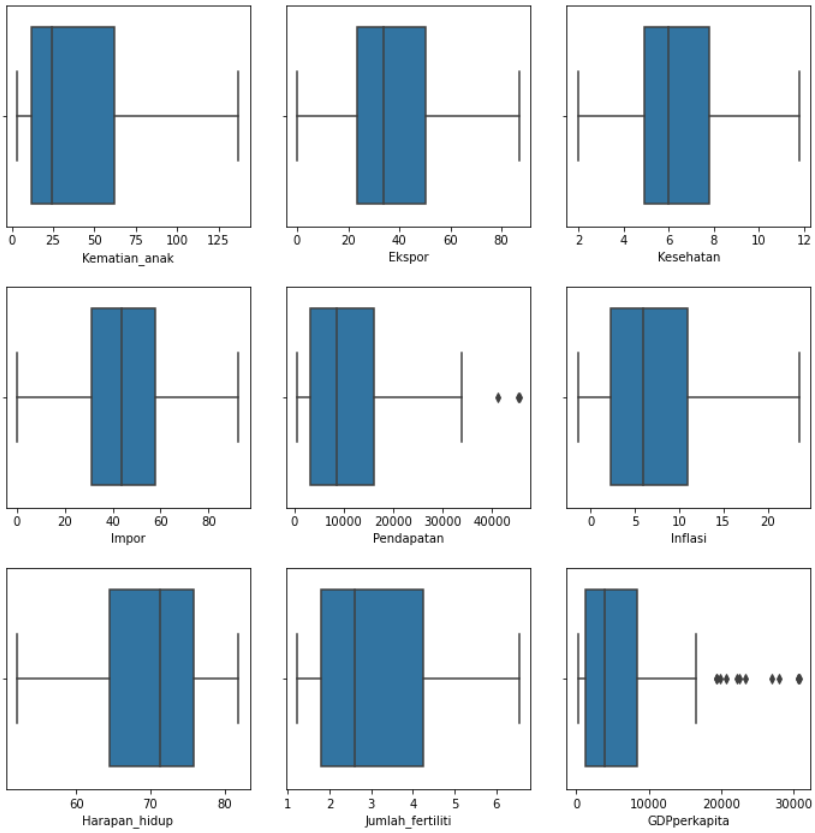


Figur di atas merupakan hasil *boxplot* dari semua indikator yang ada di dataset. Boxplot tersebut menunjukkan pencilan data dari dataset tersebut. Pencilan datanya merupakan data-data yang berada diluar range *boxplotnya* yang ditandai dengan titik atau *diamond*. Pencilan data tersebut dapat membuat prediksi atau klasifikasi dari *K-Means Clustering* menjadi tidak akurat. Oleh karena itu pencilan data yang ada di setiap indikatornya harus di *handle* terlebih dahulu.

Metode *outliers handling* yang akan digunakan adalah filtering data menggunakan interquartile range dari data tersebut. Berikut adalah formulanya:

$$q1 = 25^{th}\ percentile\ of\ the\ data$$
$$q3 = 75^{th} percentile\ of\ the\ data$$
$$IQR\ (Interquartile\ Range) = q3 - q1$$
$$Lower\ Boundary = q1 - 1.5 * IQR$$
$$Upper\ Boundary = q3 + 1.5 * IQR$$

Setelah sudah mendapatkan *lower dan upper boundarynya*, baru melakukan filtering data. Hanya mengambil data yang berada di antara *lower dan upper boundarynya* saja. Data yang di luar tersebut akan dihapuskan. Dengan datanya dihapus, maka pencilan data yang ada di datasetnya sudah berkurang.



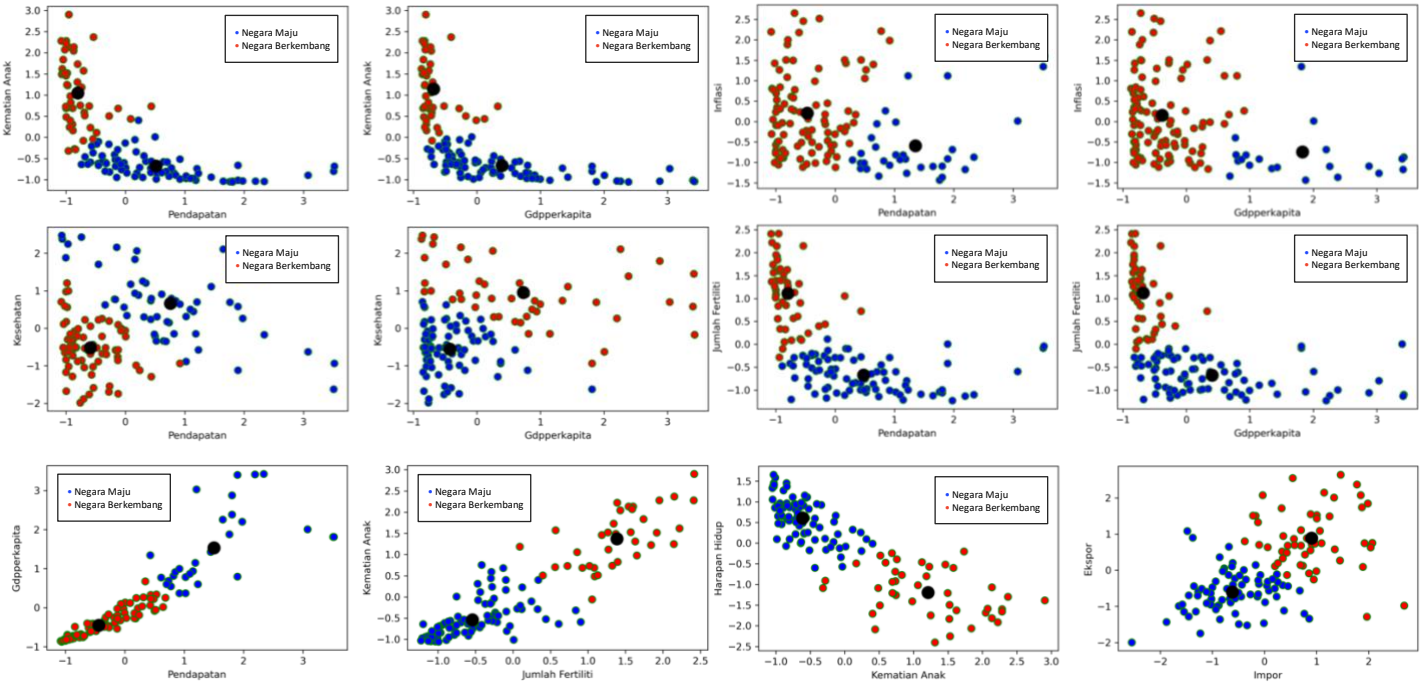
Figur di atas merupakan hasil boxplot setelah dilakukan *outliers handling*. Hampir semua indikator, pencilan datanya sudah hilang. Namun dapat dilihat pada indikator pendapatan bersih per orang dan GDP per kapita masih memiliki pencilan data setelah *outliers handling*. Hal ini dikarenakan kedua indikator tersebut memiliki rentang sebaran data yang cukup jauh, dalam kata lain, memiliki varians yang besar. Walaupu masih ada pencilan data, kita tidak bisa melakukan *outliers handling* dengan metode *remove outliers* lagi, karena itu akan menghilangkan cukup banyak data yang menyebabkan *loss of info*. Oleh karena itu, dalam *project* ini, pencilan data tersebut hanya dibiarkan.

3.6 *Scaling* Data

Scaling data sebelum melakukan *clustering* itu sangat penting. Apabila penyebaran datanya cukup jauh, maka *scaling* data itu wajib dilakukan agar *K-Means* dapat mengklasifikasikan datanya dengan akurat. Selain itu, jika satuan dari y-axis dan x-axisnya tidak sama, maka sangat dianjurkan untuk melakukan *scaling* data sebelum *clustering*.

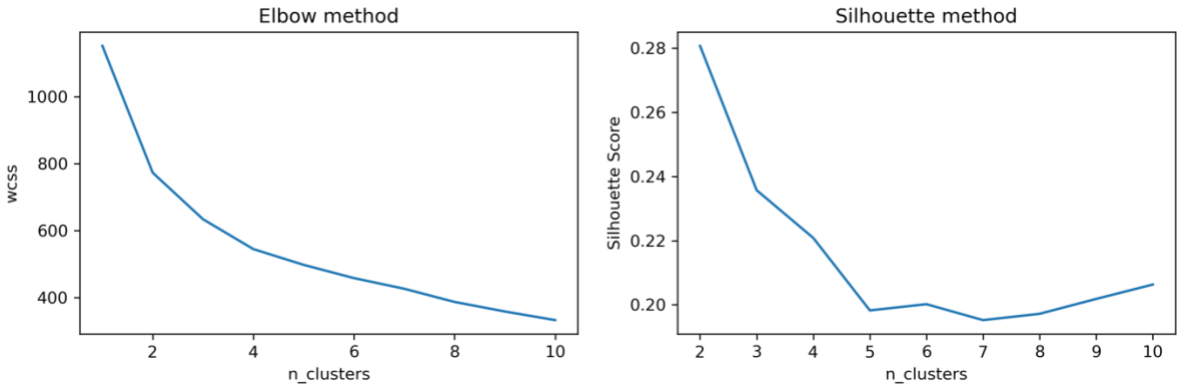
Ada 2 tipe scaling data yang populer di dunia *data science*, yaitu *Standard Scaling* dan *Normalization*. *Standard Scaling* memiliki rentang data dari -1 sampai 1 (tergantung dari *standard deviation* juga), sedangkan *Normalization* memiliki rentang data dari 0 sampai 1. Dengan riset lebih lanjut, saya menemukan bahwa *Standard Scaling* lebih cocok digunakan untuk *clustering*, sedangkan *Normalization* lebih cocok digunakan untuk *linear regression*. Oleh karena itu, dalam *project* ini, *Standard Scaling* akan digunakan untuk *scaling* datasetnya.

3.7 *K-Means Clustering* Dengan Jumlah *Cluster* 2



Grafik di atas merupakan hasil dari *K-Means Clustering* dengan jumlah *clusternya* 2. Negara maju ditandai oleh titik biru, sedangkan negara berkembang ditandai oleh titik merah (Kecuali Kesehatan, Impor, dan Ekspor). Dapat dilihat dari 2 bari pertama, negara-negara yang tergolong negara berkembang adalah negara yang memiliki ciri-ciri, pendapatan bersih yang rendah, GDP per kapita yang rendah, angka kematian anak yang tinggi, jumlah fertilitas yang tinggi, dan harapan hidup yang rendah. Di sisi lain, negara maju memiliki ciri-ciri, pendapatan bersih yang tinggi, GDP per kapita yang tinggi, angka kematian anak yang rendah, jumlah fertilitas yang rendah, dan harapan hidup yang tinggi.

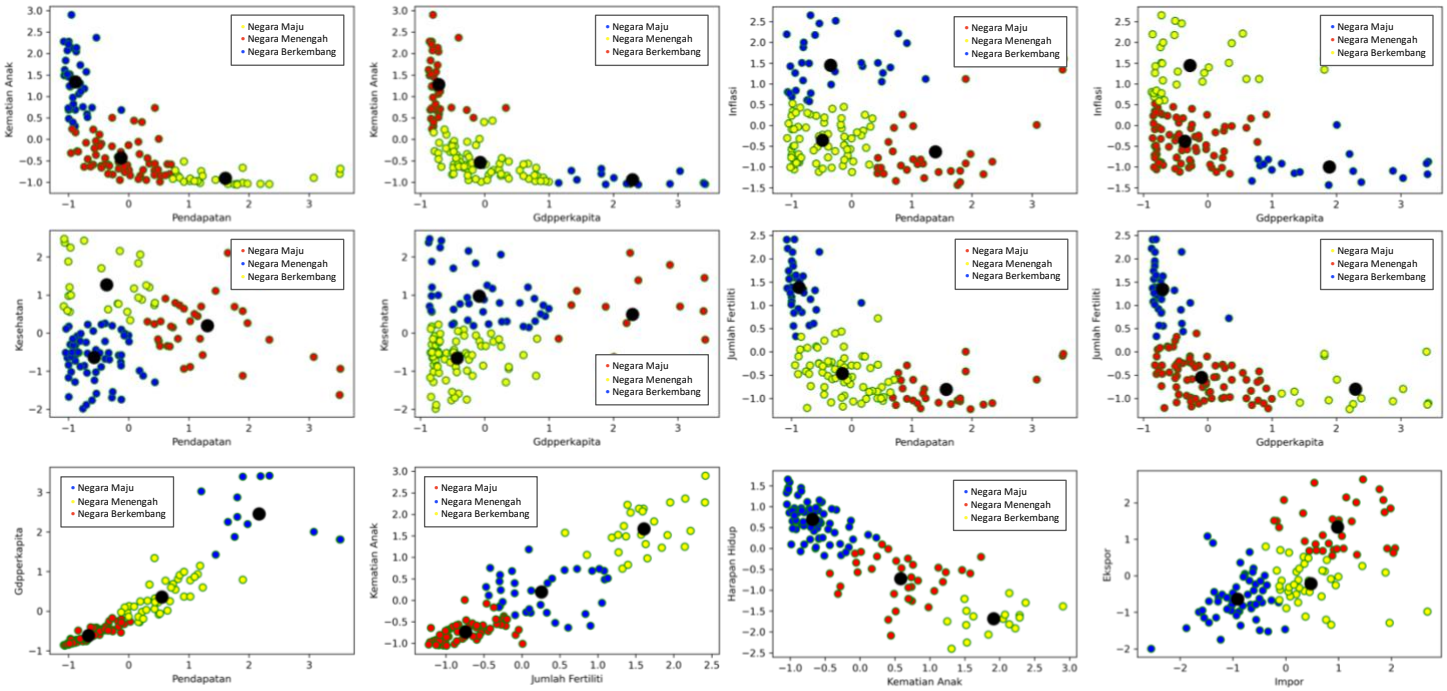
3.8 Mencari Jumlah *Cluster* Optimal



Untuk mendapatkan jumlah *cluster* optimal, dapat dilakukan dengan 2 cara atau lebih. Tetapi untuk *project* ini hanya akan digunakan cara *elbow method* dan *silhouette method*. Untuk elbow method, cari titik dimana penurunan wcssnya tidak drastic. Dalam hal ini, pada elbow method didapatkan *n_clusters* optimalnya

berada pada n = 4. Sedangkan untuk *silhouette method*, cari titik tertinggi (silhouette score yang paling tinggi). Dalam hal ini, pada silhouette method, didapatkan n_clusters optimalnya berada pada n = 2. Karena hasilnya berbeda, maka akan diambil rata-rata dari kedua method tersebut, yaitu n = 3. Oleh karena itu label negara yang akan digunakan adalah Negara Maju, Negara Menengah, dan Negara Berkembang.

3.9 *K-Means Clustering* Dengan Jumlah *Cluster* 3



Grafik di atas merupakan hasil dari *K-Means Clustering* dengan jumlah *clusternya* 3. Warna dan labelnya tidak semuanya sama karena *K-Means Clustering* menghasilkan label dengan random. Seperti yang sudah dijelaskan sebelumnya, laju inflasi, pengeluaran kesehatan, impor, dan ekspor bukan merupakan indikator yang cocok. Sehingga pada saat filtering data nanti, keempat variabel tersebut tidak akan dimasukkan.

Dari grafik di atas, dapat dilihat bahwa negara berkembang memiliki ciri-ciri, pendapatan bersih rendah, GDP per kapita rendah, angka kematian anak tinggi, jumlah fertilitas tinggi, dan harapan hidup rendah. Negara menengah memiliki ciri-ciri, pendapatan bersih rendah ke menengah, GDP per kapita rendah ke menengah, anga kematian anak rendah, jumlah fertilitas rendah ke menengah, dan harapan hidup menengah. Negara maju memiliki ciri-ciri, pendapatan bersih menengah ke tinggi, GDP per kapita menengah ke tinggi, angka kematian anak rendah, jumlah fertilitas rendah, dan harapan hidup tinggi

BAB IV

SUMMARY

4.1 *Data Summary*

Negara	Kematian_anak	Ekspor	Kesehatan	Impor	Pendapatan	Inflasi	Harapan_hidup	Jumlah_fertiliti	GDPperkapita
Burundi	93.6	8.92	11.60	39.2	764.0	12.30	57.7	6.26	231.0
Liberia	89.3	19.10	11.80	92.6	700.0	5.47	60.8	5.02	327.0
Congo, Dem. Rep.	116.0	41.10	7.91	49.6	609.0	20.80	57.5	6.54	334.0
Mozambique	101.0	31.50	5.21	46.2	918.0	7.64	54.5	5.56	419.0
Malawi	90.5	22.80	6.59	34.9	1030.0	12.10	53.1	5.31	459.0
Togo	90.3	40.20	7.65	57.3	1210.0	1.18	58.7	4.87	488.0
Guinea-Bissau	114.0	14.90	8.50	35.2	1390.0	2.97	55.6	5.05	547.0
Afghanistan	90.2	10.00	7.58	44.9	1610.0	9.44	56.2	5.82	553.0
Burkina Faso	116.0	19.20	6.74	29.6	1430.0	6.81	57.9	5.87	575.0
Uganda	81.0	17.10	9.01	28.6	1540.0	10.60	56.8	6.15	595.0

Tabel di atas merupakan 10 negara yang paling membutuhkan bantuan dana, fasilitas, dan bantuan dasar lainnya. Tabel ini didapatkan melalui proses filtering dan diambil semua perpotongan negara berkembang dari hasil *K-Means Clustering* dengan jumlah *cluster* 3 di atas. Indikator-indikator yang dipakai, antara lain, pendapatan bersih per orang, GDP per kapita, angka kematian anak, jumlah fertilitas, dan harapan hidup. Setelah datanya difilter, datanya itu diurutkan menurut GDP per kapitanya dari yang paling rendah ke tinggi. Terakhir, ambil 10 data teratas yang menghasilkan table berikut. Tabel tersebut adalah negara-negara yang memiliki GDP per kapita rendah, pendapatan bersih per orang rendah, jumlah fertilitas tinggi, angka kematian anak tinggi, dan harapan hidup rendah.

4.2 Kesimpulan

Dari table di atas, dapat diambil kesimpulan bahwa 10 negara yang paling membutuhkan bantuan dari HELP International adalah Burundi, Liberia, Congo, Dem. Rep., Malawi, Togo, Guinea-Bissau, Afghanistan, Burkina Faso, dan Uganda.

BAB V
REFERENSI

Supranto. (2004). *Analisis Multivariat: Arti dan Interpretasi*. Jakarta: PT. Rineka Cipta.

Bhandari, A. (2020, July 31). Feature Scaling: Standardization Vs Normalization. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2020/04/feature-scaling-machine-learning-normalization-standardization/

Dicoding Intern. (2020, August 25). Apa itu Machine Learning? Beserta Pengertian dan Cara Kerjanya. Dicoding Blog. https://www.dicoding.com/blog/machine-learning-adalah/.

Gupta, M. (2019, July 24). ML: Feature Scaling – Part 2. GeeksforGeeks. https://www.geeksforgeeks.org/ml-feature-scaling-part-2/.