**Laporan Case-Based 2**

**Mata Kuliah Pembelajaran Mesin**

**K-MEANS CLUSTERING**

**Muhamad Azmi Rizkifar – 1301218586 – IFX-45-GAB – IKN**

“Saya mengerjakan tugas ini dengan cara yang tidak melanggar aturan perkuliahan dan kode etik akademisi”

****

**Program Studi Sarjana Informatika**

**Fakultas Informatika**

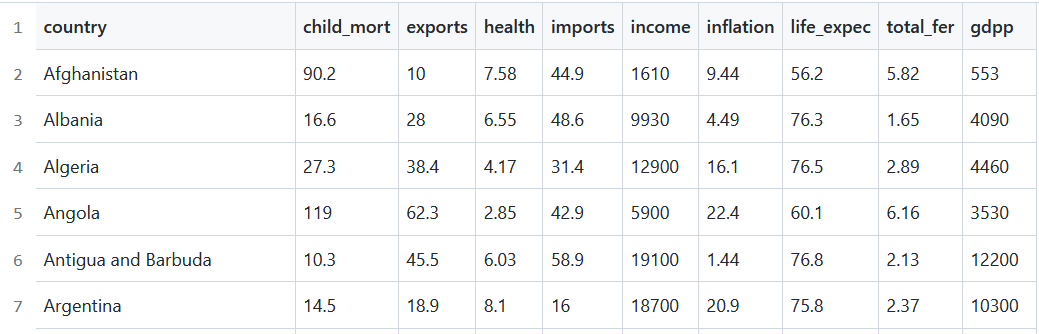
**Universitas Telkom**

**Bandung**

**2022**

# **BAGIAN I Ikhtisar Data Yang Dipilih**

Dataset yang akan dipilih dalam pengimplementasian *K-Means Clustering* adalah dataset *Country Data* yang didapatkan dari *Kaggle*. Tujuan penulis adalah untuk membangun sebuah model *clustering* untuk mengklasifikasikan data negara berdasarkan data label yang nanti akan dibuat berdasarkan data yang telah di analisis. Data yang didapat disajikan dalam bentuk format file csv seperti gambar dibawah:

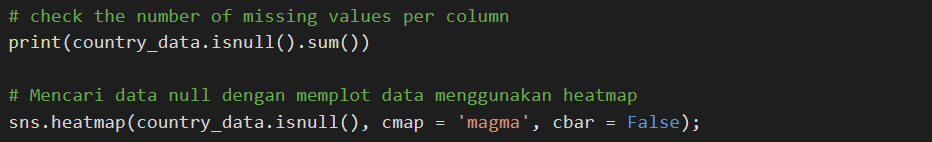


Gambar 1 Dataset Country Data

Dari keseluruhan data yang ada, terdapat total 167 data negara yang terdiri dari 10 kolom dimana terdapat 1 *categorical feature* dan 9 *numerical feature* dengan keterangan sebagai berikut :

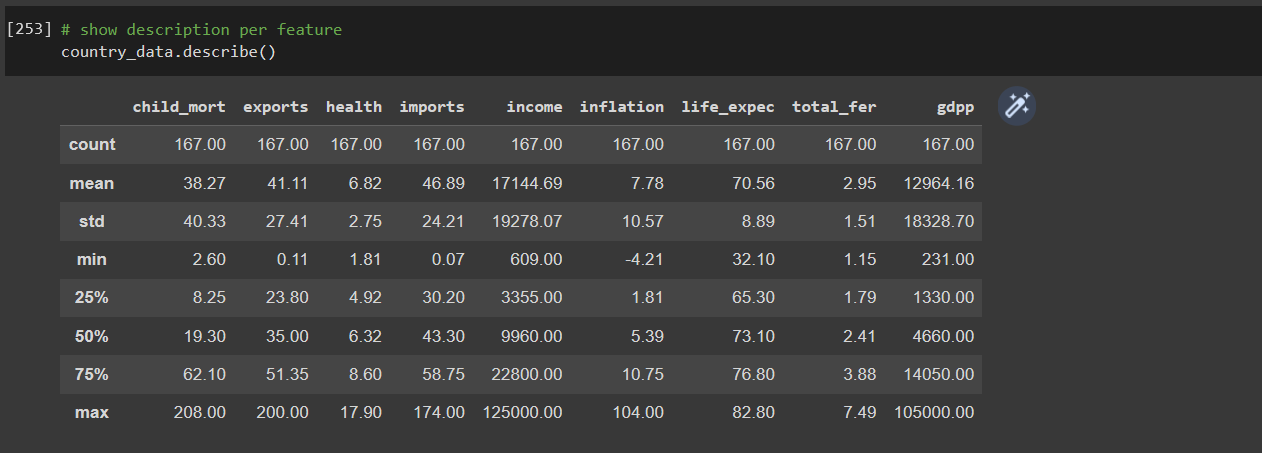
1. **country**: nama negara
2. **child\_mort**: anak di bawah usia 5 tahun per 1000 kelahiran hidup
3. **exports**: ekspor barang dan jasa per kapita
4. **health**: total belanja kesehatan per kapita
5. **imports**: impor barang dan jasa per kapita
6. **income**: pendapatan bersih per orang
7. **inflation**: pengukuran tingkat pertumbuhan tahunan Total PDB
8. **life\_expec**: rata-rata jumlah tahun seorang anak yang baru lahir akan hidup jika pola kematian saat ini tetap sama
9. **total\_fer**: jumlah anak yang akan dilahirkan oleh setiap wanita jika tingkat kesuburan usia saat ini tetap sama
10. **gdpp**: PDB per kapita. Dihitung sebagai Total PDB dibagi dengan total populasi

Sebelum masuk ke tahap *pre-*processing, dilakukan Exploratory Data Analysis untuk meninjau apakah terdapat kecatatan data atau tidak.

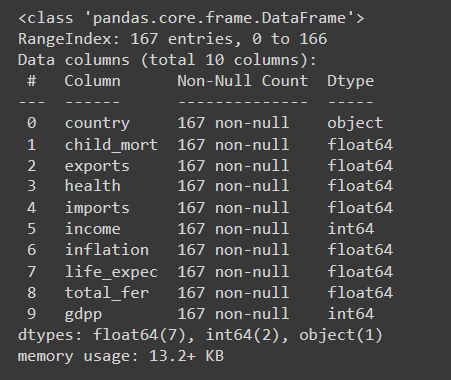




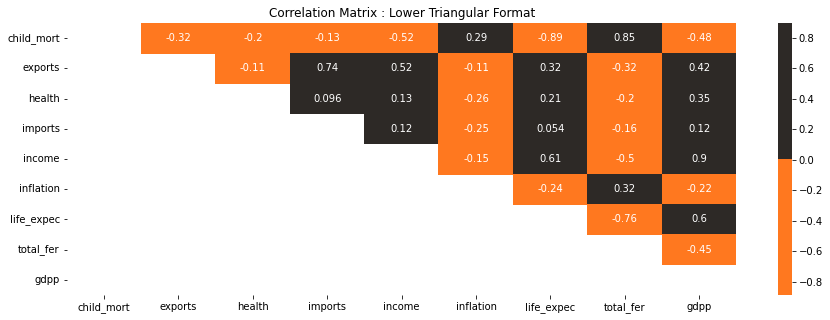
Dari hasil pengecekan data diatas, tidak ditemukan adanya *missing value* atau *null value* pada data tersebut.



Dari hasil pengecekan deskripsi data diatas, ditampilkan beberapa nilai jumlah, rata-rata, standar deviasi, dan lain sebagainya.



Terdapat 10 kolom atau fitur dengan 1 *categorical feature* dengan tipe data *object* dan 9 *numerical feature* dengan tipe data *integer* dan *float*. Pada dataset ini, dikarenakan memiliki fitur yang sedikit maka hanya dilakukan pengecekan secara manual seperti tahapan diatas. Berikut ditampilkan gambar *Correlation Matrix*:



Beberapa fitur pada dasarnya berasal dari kategori yang sama dan memiliki keterkaitan terhadap fitur lainnya dari kategori yang berbeda. Hal ini bisa dibuktikan dengan :

1. Kenaikan nilai *export* meningkatkan nilai *import*, *income*, dan *gdpp*.
2. Kenaikan nilai *child\_mort* terjadi ketika nilai inflasi dan *total\_fer* meningkat.
3. Nilai *life\_expec* yang tinggi menampilkan *total\_fer* yang rendah dan *gdpp* yang tinggi menyebabkan banyak pengeluaran untuk kesehatan (*health*).

Beberapa fitur berasal dari katageori yang sama karena memiliki keterkaitan dengan fitur lainnya. oleh karena itu, penulis membagi kategori tersebut menjadi 3 dengan keterangan :

1. kesehatan (*health*): *child\_mort*, *health*, *life\_expec*, *total\_fer*
2. perdagangan (*trade*): *import*, *export*
3. keuangan (*finance*): *income*, *inflation*, *gdpp*

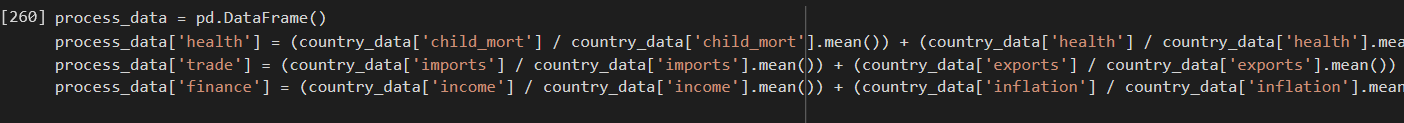
# **BAGIAN II Pra-Pemrosesan Data**

Sebelum dilakukannya pembuatan model dan klasifikasi data, perlu dilakukannya tahap pra-pemrosesan data dengan harapan untuk mendapatkan hasil klasifikasi yang sesuai. Berikut merupakan penjabaran tahapan pra-pemrosesan yang dilakukan oleh penulis :

## **Penggabungan fitur**

Dikarenakan beberapa fitur berasal dari katageori yang sama karena memiliki keterkaitan dengan fitur lainnya, dilakukan penggabungan fitur menjadi 3 kategori yang terdiri dari :

* kesehatan (*health*): *child\_mort*, *health*, *life\_expec*, *total\_fer*
* perdagangan (*trade*): *import*, *export*
* keuangan (*finance*): *income*, *inflation*, *gdpp*

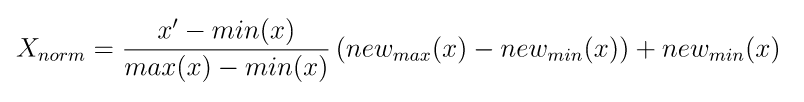


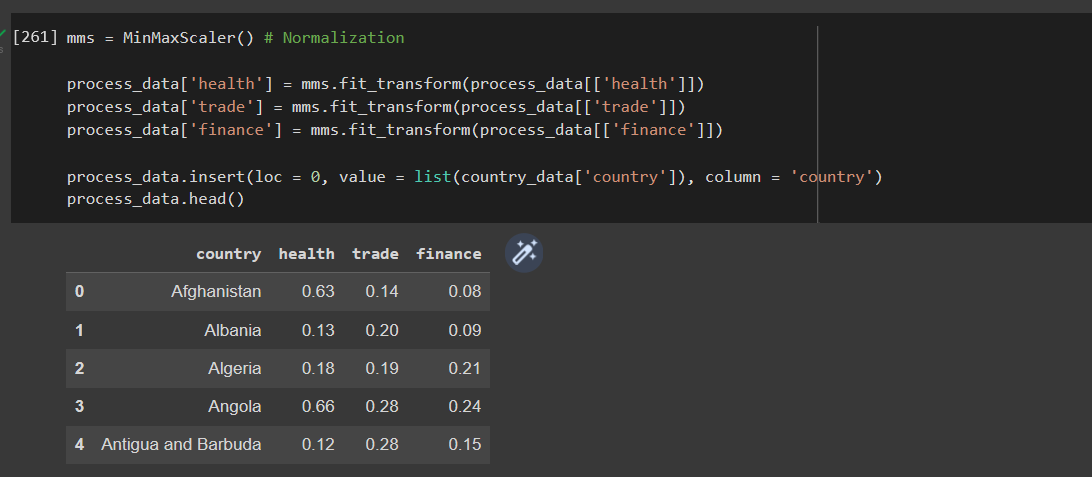
Selanjutnya, penulis melakukan penghapusan fitur *country* karena fitur tersebut merupakan *categorical data* yang dimana dalam algoritma *K-Means Clustering* tidak mendukung *categorical features*.



## **Scaling Feature**

Pada tahapan ini, dilakukan scaling fitur karena *machine learning* hanya memperlakukan nilai input sebagai angka sederhana dan tidak memahami satuan nilai fitur. Maka dari itu, perlu dilakukan normalisasi data untuk fitur yang datanya tidak menampilkan distribusi secara normal. Penulis melakukan normalisasi *Min-Max* dengan mengubah sekumpulan data menjadi skalai mulai dari 0 (Min) hingga 1 (Max) dengan bantuan function *MinMaxScaler()*. Berikut terlampir formula perhitungan normalisasi Min-Max dan penerapannya :





Dari hasil normalisasi data tersebut, data yang saat ini disajikan sudah berada di rentang 0 sampai dengan 1 dan distribusi datanya sudah terlihat normal (tidak ada data yang melebihi range 0 – 1.

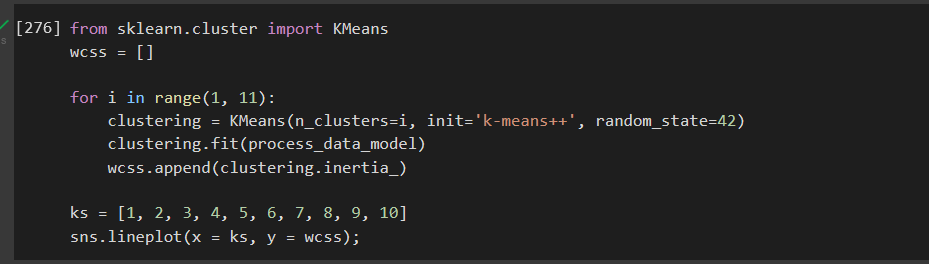
# **BAGIAN III Penerapan Algoritma**

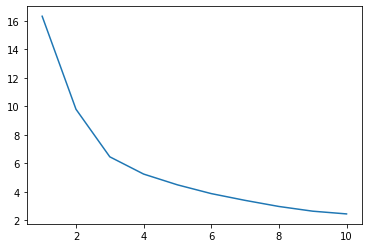
Pada proses penerapan algoritmanya, penulis menggunakan algoritma *K-Means Clustering* untuk melakukan clustering dan mengklasifikasikan data negara berdasarkan data label yang nanti akan dibuat berdasarkan data yang telah di analisis.

Berikut penjabaran implementasi *K-Means Clustering* dalam penyelesaian studi kasus *Country Data* dengan bantuan *Python* :

## **Menentukan jumlah nilai K**

Dalam menentukan jumlah nilai K digunakan *Elbow Method* sebagai salah satu metode yang memplot sum of squared error untuk rentang nilai K. Biasanya digunakan analogi lengan karena untuk menentukan nilai K, kita dapat memilih nilai yang menyerupai siku. Dari nilai ini, jumlah nilai kuadrat (inersia) mulai menurun secara linier dan karenanya dianggap sebagai nilai optimal.



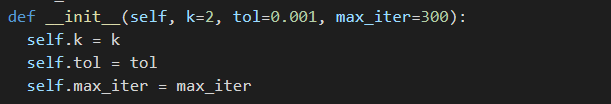


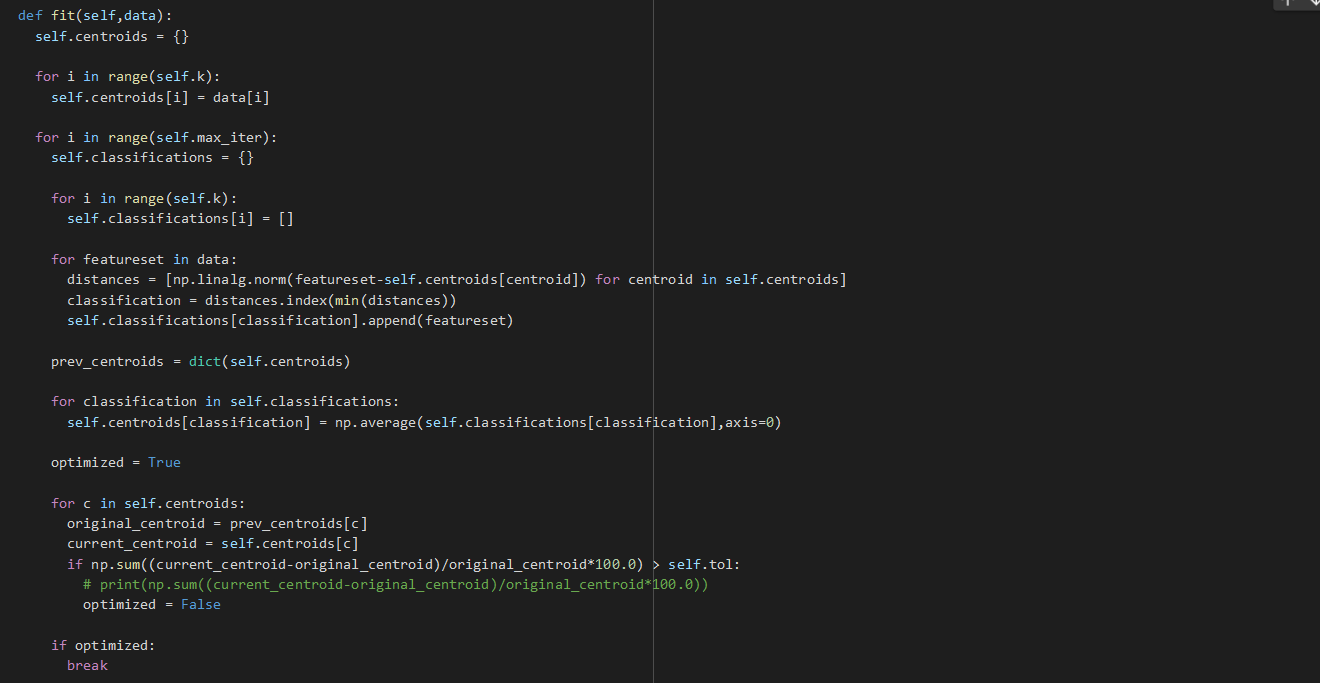


Dari hasil pembuatan plot diatas, diambil nilai 3 karena membentuk siku bawah yang menunjukkan nilai optimalnya.

## **Pembangunan model**

Pada pembangunan algoritmanya, penulis melakukan pembuatan model class K\_Means yang memiliki 3 function yang berisi *\_\_init\_\_()* dan *fit()*. Pada function *\_\_init\_\_()* berfungsi sebagai inisialisasi class yang didalamnya menerima input parameter dan melakukan set atribut di dalam class *K\_Means*. Kemudian pada function *fit()* berfungsi sebagai proses pelatihan data yang menentukan lokasi *centroid* dan melakukan pengelompokkan data pada setiap titik dengan menghitung jarak terdekat dari *centroid*.

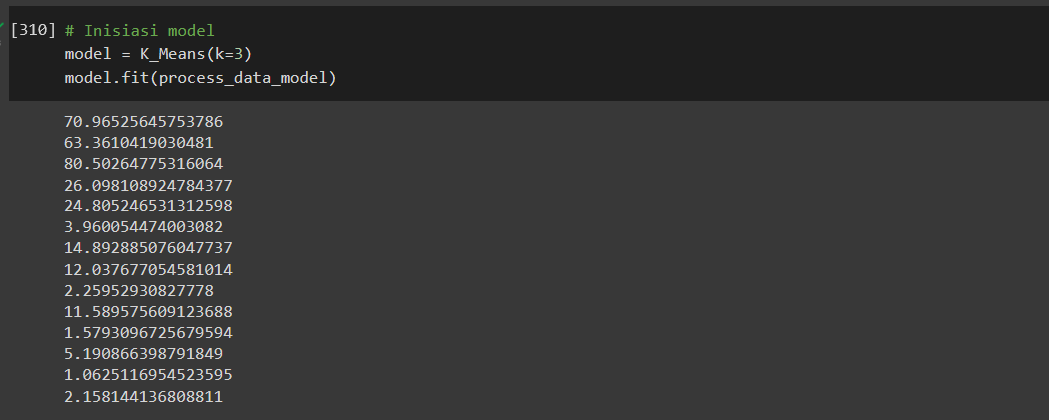




## **Training Model**

Pada proses training modelnya, penulis mendefinisikan paramter x\_train sebagai fitur yang akan di latih, y\_train sebagai target fitur, epoch sebanyak 50 iterasi, data validation yang digunakan untuk memvalidasi model dan mencegah *overfitting*, dan callbacks untuk menghentikan iterasi apabila nilai akurasi dan akurasi validasinya lebih besar dari 90%.

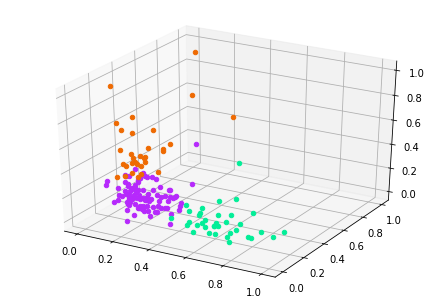
Pada proses training modelnya, penulis mendefinisikan parameter k dengan nilai 3 sesuai dengan hasil yang didapat dari *Elbow Method*. Kemudian dilakukan fitting model dengan menginputkan data yang berisi 3 fitur yang terdiri dari *health*, *trade*, dan *finance*.



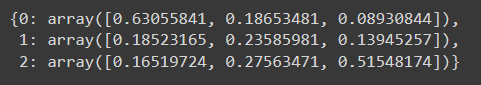
**Link Google Collab :** <https://colab.research.google.com/drive/1kYxTvZVYc-M80lAp0dEH-gkiDno5cqsV?usp=sharing>

# **BAGIAN IV Evaluasi Hasil**

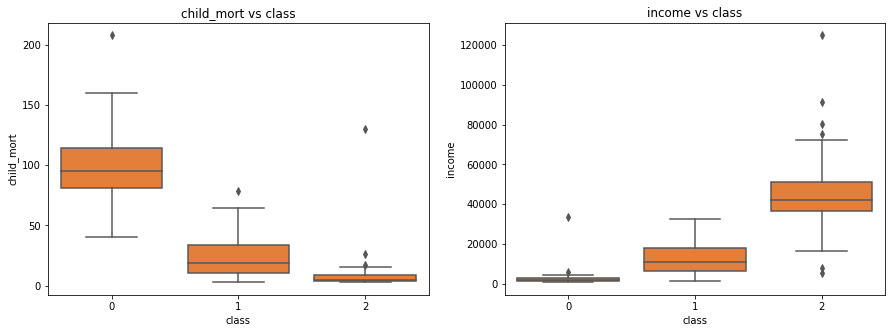
Berikut terlampir hasil clustering country data dengan nilai k = 3 :



Hasil *centroids* pada proses *clustering* diatas :



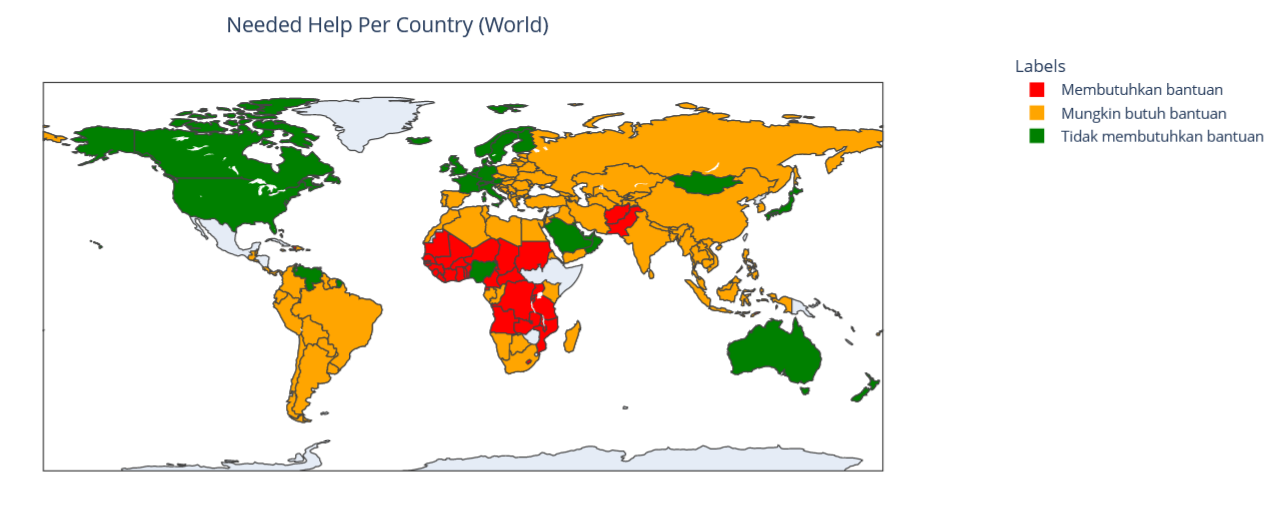
Seperti yang kita ketahui apabila nilai *income* rendah dan *child\_mort* (kematian anak) tinggi adalah tanda bahwa negara tersebut terbelakang secara ekonomi, maka digunakan fitur *income* dan *child\_mort* untuk menentukan label dari setiap class atau kelompok.

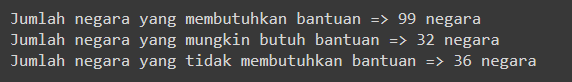


Dari hasil plot diatas, dapat disimpulkan bahwa :

* 0 : Membutuhkan bantuan
* 1 : Mungkin butuh bantuan
* 2 : Tidak membutuhkan bantuan

Selanjutnya dilakukan pembuatan plot peta dunia dengan pelabelan sesuai dengan class dan dataset country data.





Dari hasil plot pemetaan data dan class tersebut, terdapat 99 negara yang membutuhkan bantuan, 32 negara yang mungkin butuh bantuan, dan 36 negara yang tidak membutuhkan bantuan.

# **BAGIAN V Presentasi Video**

Link slide + source code :

<https://github.com/azmirizkifar20/Machine-Learning-K-Means-Clustering>

Link Video :

<https://youtu.be/qxiVU1e1l2o>