I

UB

# Big Data Assignment

# K-Means Clustering: Clustering the Countries

2 April 2021

DIANNA AGUSTINA S. | 1202912003

# K-Means Clustering

K-Means Clustering merupakan suatu metode penganalisaan data atau metode Data Mining yang melakukan proses pemodelan tanpa supervisi (unsupervised) dan merupakan salah satu metode yang melakukan pengelompokan data dengan sistem partisi. Metode K-Means Clustering berusaha mengelompokkan data yang ada ke dalam beberapa kelompok, dimana data dalam satu kelompok mempunyai karakteristik yang sama satu sama lainnya dan mempunyai karakteristik yang berbeda dengan data yang ada di dalam kelompok yang lain.

III

# Country Data for HELP International Dataset

## Intro

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.

## Info

Kategori negara menggunakan faktor sosial ekonomi dan kesehatan yang menentukan pembangunan negara secara keseluruhan.

## Sumber Dataset :

https://www.kaggle.com/rohan0301/unsupervised-learning-on-country-data


# Business Understanding

Ruang Lingkup Bisnis : Sosial / Kemanusiaan

HELP International memiliki dana sebesar $10 juta. Kemudian, pimpinan organisasi ini perlu menentukan bagaimana agar dana tersebut dapat digunakan untuk negara yang membutuhkan bantuan secara strategis dan efektif. Oleh karena itu, perlu melakukan kategori negara menggunakan beberapa faktor sosial ekonomi dan kesehatan yang menentukan perkembangan negara secara keseluruhan. Hasil dari penelitian ini akan membantu pimpinan organisasi dalam mengambil keputusan terkait penyaluran dana tersebut.

# Data Understanding

| Field Name | Description | Data Type |
|---|---|---|
| country | *Negara* | object |
| child_mort | *Kematian anak dibawah usia 5 tahun (ribuan)* | float64 |
| exports | *Ekspor barang jasa per kapita* | float64 |
| health | *Pengeluaran untuk kesehatan per kapita* | float64 |
| imports | *Impor barang jasa per kapita* | float64 |
| income | *Penghasilan bersih per-orang* | int64 |
| inflation | *Inflasi* | float64 |
| life_expec | *Rata-rata harapan hidup masyarakat* | float64 |
| total_fer | *Rata-rata kehamilan wanita* | float64 |
| gdpp | *GDP per kapita* | int64 |


# Data Preparation

## ◆ Menghapus kolom tertentu

Kolom yang tidak digunakan dalam pemodelan dihapus dari DataFrame. Kolom dimaksud merupakan kolom dengan tipe data Object (String).

## ◆ Mengecek ringkasan statistik dan hubungan antar variabel

Pengecekan ringkasan data statistik dengan menggunakan fungsi describe() pada DataFrame. Ringkasan data statistik yang ditampilkan adalah jumlah data, nilai min, max, standar deviasi, median, mean, kuartil bawah, kuartil tengah, dan kuartil atas.

## ◆ Melihat korelasi antar variabel

Untuk melihat korelasi antar variabel dapat menggunakan fungsi corr() pada DataFrame. Pada bagian ini dapat dilihat bahwa variabel gdpp memiliki korelasi positif dengan income, imports, exports, child_mort, dan total_fert. Sementara variabel life_expec dan child_mort memiliki korelasi negatif.

X

# Modeling (Training)

### Melakukan Penskalaan

Penskalaan nilai pada dataset dilakukan dengan menggunakan fungsi StandarScaler(). Fungsi StandarScaler() mengubah data mean menjadi 0 dan varian 1.

### Menentukan Jumlah Klaster

Jumlah klasterditentukan dengan menggunakan metode elbow. Posisi 'elbow' berada di angka 3, maka nilai cluster untuk data ini adalah 3.

### Menghitung KMeans Clustering

Setelah mendapatkan jumlah klaster dengan metode elbow, kemudian dilakukan perhitungan KMeans terhadap data hasil pengskalaan. Dari proses ini dihasilkan jumlah data pada setiap klaster.

# Evaluation (Testing and Accuracy)

### Pengetesan (Testing)

Testing dilakukan dengan menjalankan hasil pemodelan terhadap data hasil penskalaan dengan fungsi fit_predict().

### Pengecekan Akurasi

Pengecekan akurasi dengan menggunakan fungsi silhouette_score dan menghasilkan nilai 0.28329575683463126. Nilai ini menunjukan bahwa akurasi pemodelan yang telah dibuat sebesar 28,32%.

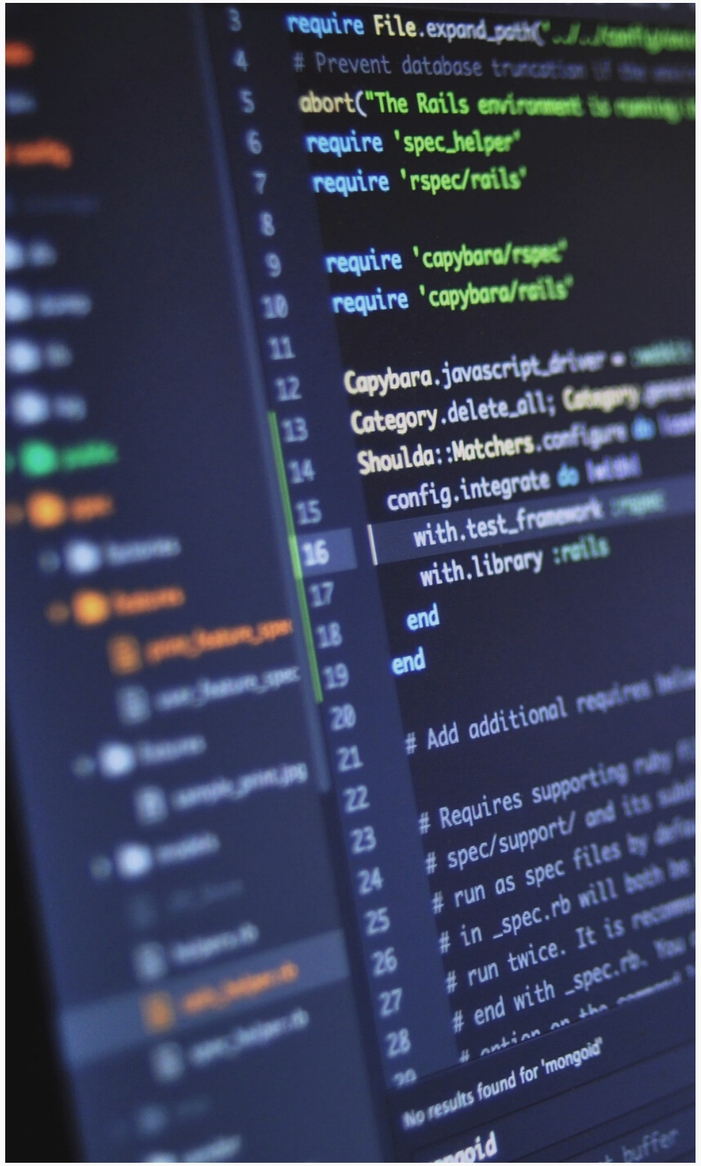
### Menambahkan Kolom Hasil Klaster

Melakukan penambahan kolom pada dataset, dimana kolom ini akan memuat nilai klaster pada setiap data.

# Conclusion

Perhitungan K-Means Cluster menghasilkan 3 klaster. Apabila dilihat dari hubungan antara nilai Child Mortality terhadap nilai GDPP dan nilai Inflation terhadap nilai GDPP pada plot yang ditampilkan, maka dapat disimpulkan bahwa 3 klaster tersebut terdiri dari klaster negara miskin, negara berkembang, dan negara maju.

Pada bagian akhir deployment disertakan fitur pengujian untuk mengetahui klaster dari suatu negara.




# Deployment

Tahap ini dilakukan menggunakan bahasa pemrograman Python dengan IDE Visual Studio Code.

Code secara keseluruhan terdapat pada link Github :

https://github.com/dyanaagustina/Learn-BigData/tree/Big-Data/KMeans

# Thank You