**TUGAS CASE BASED 2  
MACHINE LEARNING**

Logo, company name

Description automatically generated

**Nama**:

Hilman Taris Muttaqin 1301204208

**Anda mengerjakan tugas ini dengan cara yang tidak melanggar aturan perkuliahan dan kode etik akademisi.**

**Kode Dosen**:

DDR

UNIVERSITAS TELKOM

TAHUN AKADEMIK 2022/2023

# **Bagian 1 Ikhtisar data**

Data yang digunakan adalah data yang terdapat pada Kaggle tentang perkembangan suatu negara  
<https://www.kaggle.com/datasets/rohan0301/unsupervised-learning-on-country-data>

Tujuan dari data ini adalah untuk mengkategorikan negara menggunakan faktor sosial-ekonomi dan kesehatan yang menentukan pembangunan negara secara keseluruhan. Terdapat 9 kolom pada data ini, yaitu:

1. country: nama dari negara
2. child\_mort: kematian anak umur dibawah 5 tahun per 1000 kelahiran
3. exports: jumlah eksport barang dan servis percapita. Diberikan sebagai persentase dari GDP perkapita
4. health: jumlah pengeluaran untuk biaya kesehatan perkapita. Diberikan sebagai persentase GDP perkapita
5. imports: jumlah import barang dan servis perkapita. Diberikan sebagai persentase GDP perkapita
6. Income: pendapatan bersih perorang
7. Inflation: pengukuran dari pertambahan jumlah GDP pertahun
8. life\_expec: rata-rata jumlah tahun dari anak yang belum lahir akan hidup jika pola mortalitas tetap sama
9. total\_fer: banyak anak yang akan lahir untuk setiap wanita jika age-fertility sekarang masih sama
10. gdpp: GDP perkapita dihitung dari total GDP dan dibagi dengan banyaknya populasi

Preview Data  
A black screen with white text

Description automatically generated with low confidence

Informasi Umum  
A picture containing text, plaque

Description automatically generated  
Jika dilihat dari data ini, terdapat 167 record dan memiliki 1 tipe data deskriptif yaitu yang terdapat pada kolom country. Sehingga untuk kolom ini, akan di drop saat proses perhitungan. Tidak terdapat data yang kosong atau null.

General Descriptive Statistics  
Text

Description automatically generated

# **Bagian 2 Ringkasan Pra-Pemrosesan Data**

Pra-Pemrosesan data merupakan proses menyiapkan data agar nantinya proses perhitungan atau learning dapat berjalan optimal dan menghasilkan hasil yang memuaskan. Terdapat beberapa proses yang diimplementasikan, yaitu:

1. Pengecekan duplikasi data  
   Text

   Description automatically generated

## Pengecekan data kosong Text Description automatically generated

## Scaling Text Description automatically generated

Scaling berfungsi untuk mempercepat proses perhitungan, karena data aslinya memiliki rentang angka yang sangat besar. Scaling yang digunakan adalah min max scaling dengan rentang dari 0 hingga 1.

Kesimpulannya data yang digunakan termasuk data yang lumayan rapih

# Bagian 3 Implementasi Model

Model yang digunakan untuk proses clustering pada tugas ini adalah metode kMeans. kMeans terdiri dari parameter k untuk menentukan jumlah cluster yang akan dibuat pada data yang ada. Berikut adalah rankaian prosesnya:

1. Generate random centroid  
   Centroid digenerate sebanyak fitur \* k. Setiap fitur atau kolom memiliki k centroid. Centroid diambil dari sample atau 1 record acak pada dataset.
2. Iterasi hingga centroid tidak berubah  
   Dalam Iterasi kMeans, terdapat langkah langkah yang dilakukan, yaitu:
   1. Labeling  
      Labeling dilakukan untuk melabeli apakah data termasuk ke label a, b, c, dst. Banyaknya label ditentukan dari nilai k. Label dihitung berdasarkan centroid pertama atau initial centroid.

Diagram, schematic

Description automatically generated  
Graphical user interface, text

Description automatically generated

* 1. Centroid  
     Pada tahap ini centroid baru dihitung. Perhitungan dilakukan dengan cara hitung masing-masing centroid dengan data yang ada dan ambil jarak terdekatnya (bisa menggunakan rumus Euclidean distance) yang pada akhirnya data dengan jarak terdekat itu akan menentukan kelasnya atau labeling. Centroid baru akan terbentuk dari perhitungan jumlah kelas dibagi dengan banyaknya jumlah kelas atau rata-ratanya.  
     Text

     Description automatically generated
  2. Visualisasi data  
     Pada tahap ini data divisualisasikan dalam bentuk 2 dimensi. Metode PCA (Principal Component Analysis) digunakan untuk mengkonversi atau mereduksi data dengan dimensi yang banyak menjadi sejumlah dimensi yang diingingkan. Pada kasus ini, data dikonversi dari 9 dimensi menjadi 2 dimensi. Data 2 dimensi itu merepresentasikan x dan y pada grafik.  
     A screenshot of a computer

     Description automatically generated with medium confidence  
     PCA diterapkan pada tahap ini untuk kebutuhan visualisasi data 2 dimensi.  
       
     Chart, scatter chart

     Description automatically generated  
     Contoh implementasi visualisasi data.
  3. Pemberian label kelas dan negara  
     Pada tahap ini, setelah mendapatkan label terakhir pada data, maka data perhitungan, label, dan negara disatukan  
     Graphical user interface, text

     Description automatically generated

# **Bagian 4 Evaluasi Hasil**

Pada metode kMeans ini, terdapat parameter yang diinputkan, yaitu jumlah k atau banyaknya kelas yang akan dibuat. Cara untuk mencari nilai k yang tepat salah satunya adalah menggunakan metode elbow.

Text

Description automatically generated  
Chart, line chart

Description automatically generated  
Hasil dari metode elbow adalah banyaknya cluster yang tepat adalah 3 cluster.

# **Bagian 5 Presentasi Video**

Berikut ini adalah kumpulan resource yang ada pada tugas ini

<https://drive.google.com/drive/folders/1j9OIYSrGEG79shI4udK7nxTMXRuBKGT6?usp=share_link>