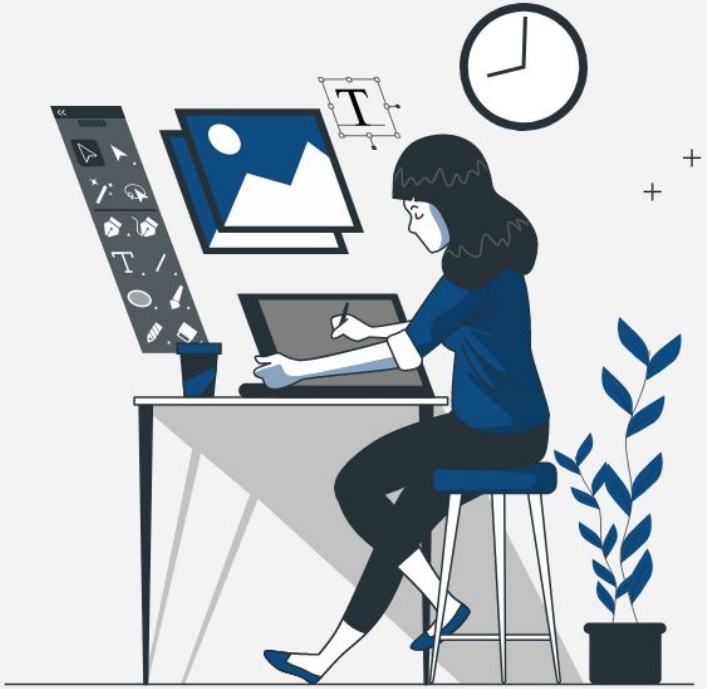
CASE BASED 02

MACHINE LEARNING



Berlian Muhammad G. A (1301204378)
Kelas: IF-44-10

CASE BASED 02

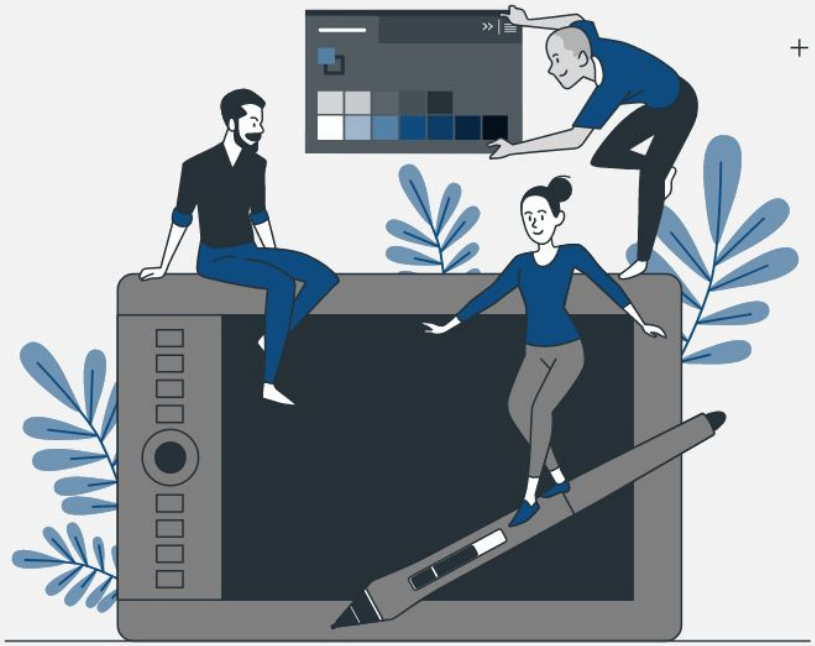
NIM GENAP = Dataset Country-data

Baris	167
Kolom	10
Tipe Data	float64 (7), int64 (2), object (1)

CASE BASED 02

country	Nama negara sebanyak 167 negara
child-mort	Kematian anak di bawah usia 5 tahun per 1000 kelahiran hidup
exports	Ekspor barang dan jasa per kapita. Diberikan sebagai % dari usia PDB per kapita
health	Total pengeluaran kesehatan per kapita. Diberikan sebagai % dari usia PDB per kapita
imports	Impor barang dan jasa per kapita. Diberikan sebagai % dari usia PDB per kapita
income	Pendapatan bersih per orang
inflation	Ukuran tingkat pertumbuhan tahunan Total PDB
life_expect	Jumlah rata-rata tahun hidup seorang anak yang baru lahir jika pola kematian saat ini tetap sama
total_fer	Jumlah anak yang akan dimiliki setiap wanita jika tingkat kesuburan usia saat ini tetap sama
gdpp	PDB per kapita. Dihitung sebagai Total PDB dibagi dengan total populasi

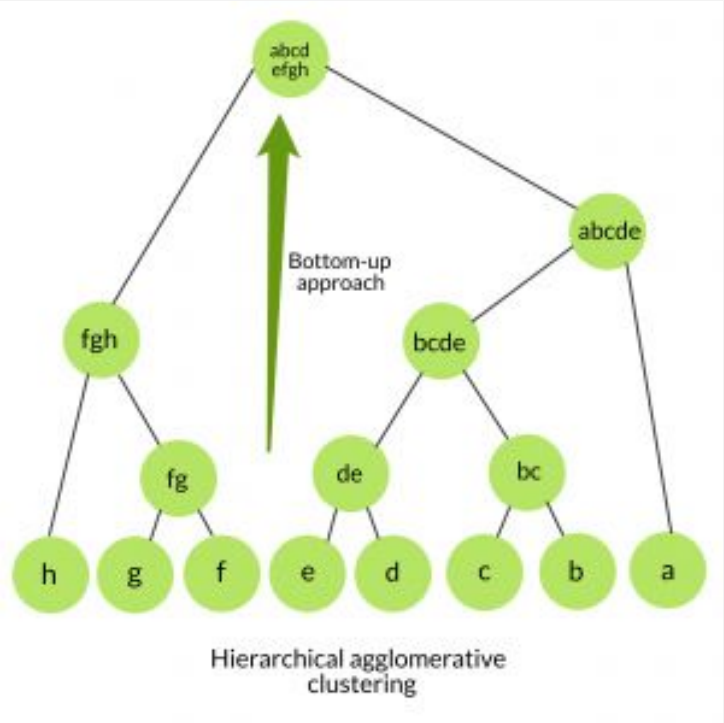
UNSUPERVISED LEARNING



Unsupervised learning adalah salah satu tipe algoritma machine learning yang digunakan untuk menarik kesimpulan dari dataset. Metode ini hanya akan mempelajari suatu data berdasarkan kedekatannya saja atau yang biasa disebut dengan clustering. Metode unsupervised learning yang paling umum adalah analisis cluster, yang digunakan pada analisa data untuk mencari pola-pola tersembunyi atau pengelompokan dalam data.

AGGLOMERATIVE HIERARCHICAL CLUSTERING

Biasa disebut juga sebagai agglomerative nesting dimana cara kerja dalam melakukan pengelompokan data menggunakan ***bottom-up***. Prosesnya dimulai dengan menganggap setiap data sebagai satu cluster kecil (*leaf*) yang hanya memiliki satu anggota saja, lalu pada tahap selanjutnya dua cluster yang memiliki kemiripan akan dikelompokkan menjadi satu cluster yang lebih besar (*nodes*). Proses ini akan dilakukan terus menerus hingga semua data menjadi satu cluster besar (*root*).



Hierarchical agglomerative clustering

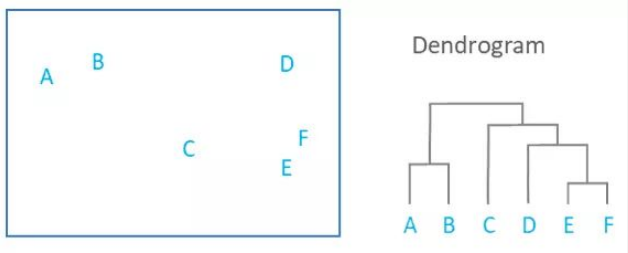
KELEBIHAN & KEKURANGAN

- Mampu menggambarkan kedekatan antar data dengan dendrogram.
- Cukup mudah untuk pembuatannya.
- Dapat menentukan banyak cluster yang terbentuk setelah dendrogram terbentuk.
- Tidak memiliki fungsi objektif alami yang sedang dioptimalkan (berbeda dengan K-Means)

- Tidak dapat menganalisis data kategorik secara langsung
- Tidak diperuntukkan untuk menghasilkan jumlah cluster optimal yang mutlak Sensitif terhadap data yang memiliki skala berbeda
- Sensitif terhadap *outlier.*
- Cukup berat komputasinya untuk data berukuran besar.
    a) Kompleksitas Ruang : O(N2)
    b) Kompleksitas Waktu : O(N3)

CARA KERJA

- Menyiapkan data dimana data yang digunakan adalah data bertipe numerik agar dapat digunakan untuk penghitungan jarak.
- Menghitung *(dis)similarity* atau jarak antar data yang berpasangan pada dataset. Nilai *(dis)similarity* tersebut kemudian akan disusun menjadi *distance matrix*.
- Membuat dendrogram dari *distance matrix* menggunakan *linkage method* tertentu. Kita juga dapat mencoba beberapa *linkage method* kemudian memilih dedrogram paling baik.
- Menentukan dimana akan melakukan pemotongan tree (dengan nilai *(dis)similarity* tertentu). Disinilah tahap dimana cluster akan terbentuk.
- Melakukan interpretasi dari dendrogram yang telah didapat.



PRE-PROCESSING DATA

EKSPLORASI

Converting - Display Info - Detecting Missing & Duplicate Value - Indexing - Standardization & Scaling Data

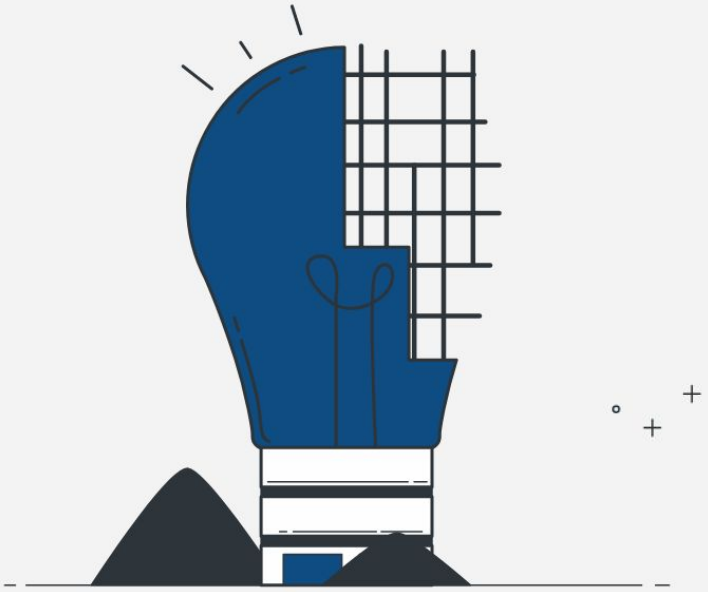
VISUALISASI

Display Info - Data Distribution - Boxplot Outliers - Correlation

PROCESS

Principal Component Analysis - Elbow Method - Cophenet & Silhouette Coefficient - Hierarchical Clustering (Agglomerative)

DEMO PROGRAM



KESIMPULAN

Hierarchical Clustering adalah algoritma yang mengelompokkan objek serupa ke dalam kelompok yang disebut cluster. Titik akhir adalah kumpulan cluster, di mana setiap cluster berbeda satu sama lain, dan objek dalam setiap cluster secara umum mirip satu sama lain.

Dengan penjelasan kasus diatas, dapat disimpulkan bahwa dengan penerapan algoritma *agglomerative hierarchical clustering*, dapat memberikan keluaran dengan baik, apabila dilakukan preproccesing terlebih dahulu sebagai tahap awal sebelum masuk ke dalam tahapan modeling. Dengan menggunakan metode tahapan linkage serta skor Silhouette yang terbaik akan dihasilkan output berupa pengelompokkan clustering yang terbaik.

THANK YOU