



PROJECT UAS KOMPUTASI STATISTIKA

Kelompok J





ANGGOTA

ADHA ABDULLAH - 2206053921

HAIKAL FIKRI RABANI - 2206823713

JOLIN FRANSIUS - 2206051374

MUHAMMAD ADLI RAHMAT SOLIHIN - 2006529184

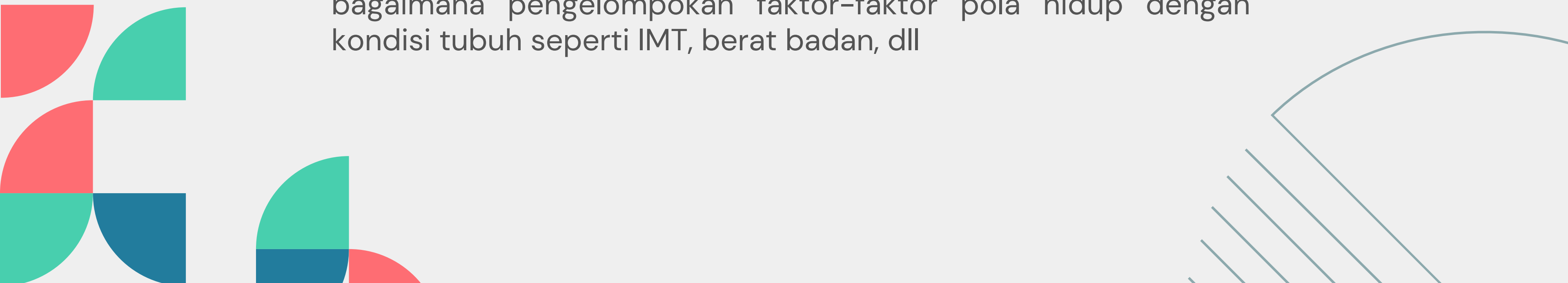
YIESHA REYHANI GHOZALI - 2206828115



The top-left corner features a series of thin, parallel diagonal lines in a light blue-grey color. The top-right corner contains several overlapping semi-circles in bright yellow, teal, and dark blue.

LATAR BELAKANG

Indeks Massa Tubuh (IMT) adalah suatu pengukuran yang umum digunakan untuk mengevaluasi proporsi berat badan seseorang terhadap tinggi badannya. IMT dapat memberikan gambaran tentang tingkat obesitas atau kekurangan berat badan seseorang, dan sering digunakan sebagai indikator kasar kesehatan tubuh. Di samping itu, pola hidup sehat juga merupakan faktor kunci dalam menjaga kesehatan tubuh secara keseluruhan. Oleh karena itu, kami tertarik untuk meneliti bagaimana pengelompokan faktor-faktor pola hidup dengan kondisi tubuh seperti IMT, berat badan, dll

The bottom-left corner features overlapping semi-circles in red, teal, and dark blue. The bottom-right corner contains thin, parallel diagonal lines in a light blue-grey color, mirroring the top-left design.




TUJUAN

1. Melakukan penelaahan dan analisis terkait karakteristik dan informasi dari data
2. Melakukan prapengolahan data sebagai input bagi K-Means Clustering yang dirancang dengan menggunakan rumus jarak *euclidean*.
3. Menganalisis hasil clustering, bagaimana pengelompokan fitur-fitur kategorik terkait pola hidup dan kondisi tubuh.

METODE

Metode yang kami gunakan adalah k-means clustering memanfaatkan jarak euclidean untuk menentukan jarak tiap data ke centroid.



01 - PREPROCESSING

02 - CLUSTERING

03 - COMPARISON

04 - INSIGHT

OUTLINE



TASK 1

Melakukan penelaahan Data (Eksplorasi & *preprocessing*) pada data yang diberikan (Data Wrangling, Statistika Deskriptif, & Aggregate).

PRE-PROCESSING

Sebelum melakukan clustering, perlu dilakukan prapengolahan data dengan mengenali karakteristik data hingga mengekstrak informasi dari data asli. Pada *project* ini, kami melakukan penelaahan pada data dengan proses sebagai berikut.

01 - STATISTIKA DESKRIPTIF

02 - VISUALISASI DATA

03 - PRAPENGOLAHAN

04 - DATA AGGREGATION

01 - STATISTIKA DESKRIPTIF

.describe()

Perintah ini digunakan untuk menampilkan statistika deskriptif dari data seperti count (jumlah nilai), min (nilai minimum), max (nilai maksimum), mean, quantile, dan standard deviasi

.info()

Perintah ini digunakan untuk dapat mengetahui informasi mengenai type data dan jumlah data yang tidak ***null*** atau ***missing value***

.head()

Perintah ini digunakan untuk menampilkan sebagian data pada baris-baris awal untuk mengetahui tampilan awal data

.nunique()

Perintah ini digunakan untuk menghitung jumlah nilai unik pada data

01 - STATISTIKA DESKRIPTIF

Berikut hasil dari fungsi '.describe()' yang kami gunakan

	Tinggi Badan	Berat Badan	Lingkar Pinggang	Usia	Nilai IMT	IMT
count	289794.000000	289794.000000	289794.000000	289794.000000	289794.000000	289794.000000
mean	5.476080	5.535350	1.012641	4.847283	5.527590	6.090298
std	1.969247	1.612795	0.016946	1.724122	1.672043	1.964583
min	1.000000	1.000000	1.000000	1.000000	1.000000	2.800000
25%	4.375000	4.375000	1.009983	3.446602	4.375000	6.400000
50%	5.725000	5.339286	1.013310	5.019417	5.324882	6.400000
75%	6.625000	6.625000	1.014299	6.067961	6.625000	8.200000
max	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000

01 - STATISTIKA DESKRIPTIF

Berikut hasil dari fungsi '.info()' yang kami gunakan

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 289801 entries, 1 to 289801
Data columns (total 18 columns):
#   Column                Non-Null Count  Dtype
---  -
0   Tinggi Badan          289801 non-null object
1   Berat Badan           289801 non-null float64
2   Lingkar Pinggang       289801 non-null float64
3   Usia                  289801 non-null float64
4   Nilai IMT             289801 non-null object
5   IMT                   289801 non-null object
6   Aktivitas Fisik        289801 non-null object
7   Air Mineral           289801 non-null object
8   Buah dan Sayur        289801 non-null object
9   Mencuci Badan         289801 non-null object
10  Mandi                 289801 non-null object
11  Merokok               289801 non-null object
12  Konsumsi Gula          289801 non-null object
13  Konsumsi Alkohol       289801 non-null object
14  Konsumsi Junk Food     289801 non-null object
15  Menggosok Gigi         289801 non-null object
16  Mengganti Pakaian Dalam 289801 non-null object
17  Mencuci Tangan         289801 non-null object
dtypes: float64(3), object(15)
memory usage: 42.0+ MB
```

01 - STATISTIKA DESKRIPTIF

Berikut hasil dari fungsi `'head()'` yang kami gunakan

	Tinggi Badan	Berat Badan	Lingkar Pinggang	Usia	Nilai IMT	IMT
ID						
1	168.9	85.0	100.0	34.0	29.80	4
2	169.5	70.0	80.0	43.0	24.36	3
3	169.5	60.0	78.0	21.0	20.88	3
4	166.8	80.0	34.0	40.0	28.75	4
5	164.5	62.0	32.0	28.0	22.91	3

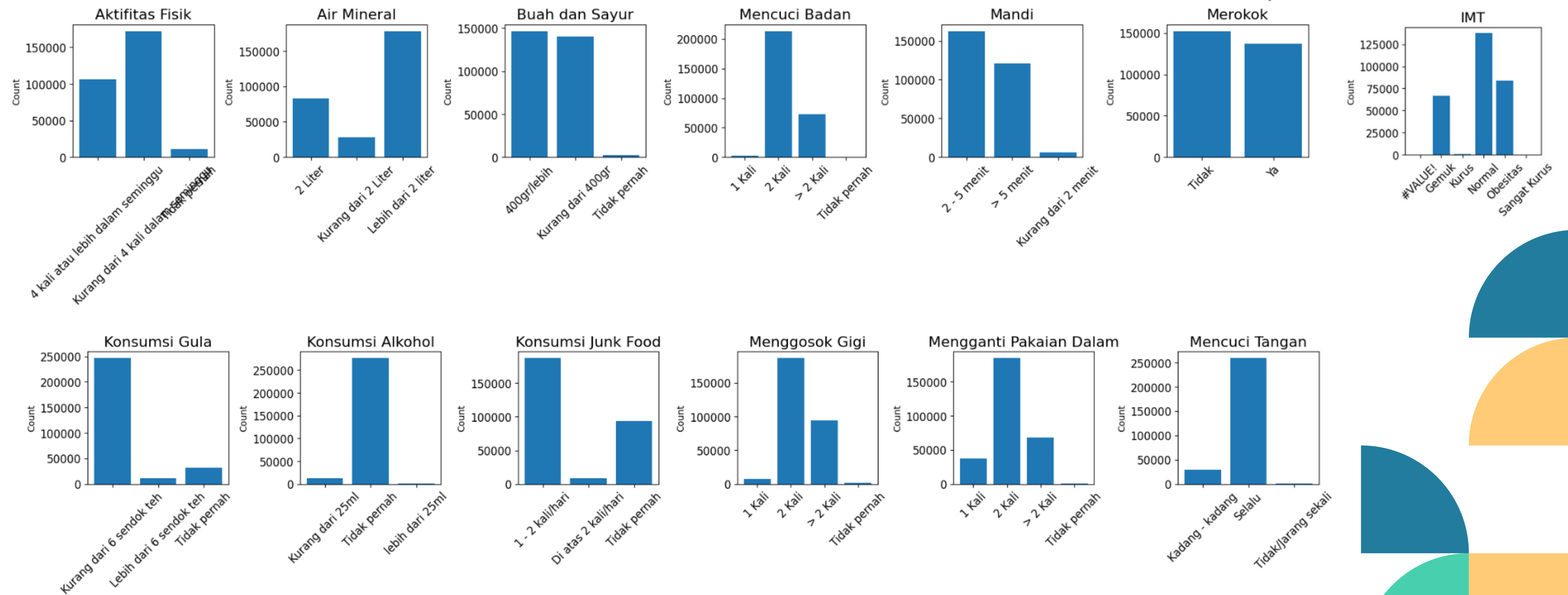
01 - STATISTIKA DESKRIPTIF

Berikut hasil dari fungsi `'nunique()'` yang kami gunakan

```
Kolom: Tinggi Badan, Jumlah Nilai Unik: 72
Kolom: Berat Badan, Jumlah Nilai Unik: 93
Kolom: Lingkar Pinggang, Jumlah Nilai Unik: 249
Kolom: Usia, Jumlah Nilai Unik: 54
Kolom: Nilai IMT, Jumlah Nilai Unik: 1230
Kolom: IMT, Jumlah Nilai Unik: 5
Kolom: Aktifitas Fisik, Jumlah Nilai Unik: 3
Kolom: Air Mineral, Jumlah Nilai Unik: 3
Kolom: Buah dan Sayur, Jumlah Nilai Unik: 3
Kolom: Mencuci Badan, Jumlah Nilai Unik: 4
Kolom: Mandi, Jumlah Nilai Unik: 3
Kolom: Merokok, Jumlah Nilai Unik: 2
Kolom: Konsumsi Gula, Jumlah Nilai Unik: 3
Kolom: Konsumsi Alkohol, Jumlah Nilai Unik: 3
Kolom: Konsumsi Junk Food, Jumlah Nilai Unik: 3
Kolom: Menggosok Gigi, Jumlah Nilai Unik: 4
Kolom: Mengganti Pakaian Dalam, Jumlah Nilai Unik: 4
Kolom: Mencuci Tangan, Jumlah Nilai Unik: 3
```

02 - VISUALISASI DATA

Pada bagian ini, visualisasi data yang dilakukan adalah pada bagian data dengan tipe kategorik yang kami gunakan untuk clustering agar mengetahui komposisi setiap fitur kategorik terlebih dahulu.



03 - PRAPENGOLAHAN

1

RENAME FEATURES

Pertama, dilakukan penamaan ulang pada kolom-kolom yang berupa pertanyaan untuk mempermudah dalam membaca dan memperseingkat *code*.

4

ENCODE FITUR KATEGORIK

Proses ini dilakukan untuk mengubah fitur-fitur kategorik yang digunakan agar menjadi data numerik. Proses ini dilakukan dengan menggunakan *module LabelEncoder* dari **scikitlearn**

2

FEATURE SELECTION

Setelah itu, kami memilih fitur-fitur yang cocok untuk tujuan dari project ini. yang terdiri dari data mengenai pola hidup serta yang berkaitan dengan individu seperti IMT, berat badan, tinggi badan, dan usia.

5

MENANGANI MISSING VALUE

Selanjutnya, dilakukan juga penanganan untuk missing value. Proses ini dilakukan dengan menghapus baris yang terdapat missing value menggunakan perintah **.dropna()**

3

MENANGANI OUTLIER

Kemdudian, dilakukan penanganan outlier pada data yang telah dideteksi dari analisis statistika deskriptif pada data numerik. Penanganan ini dilakukan dengan *Interquantile Range (IQR)*.

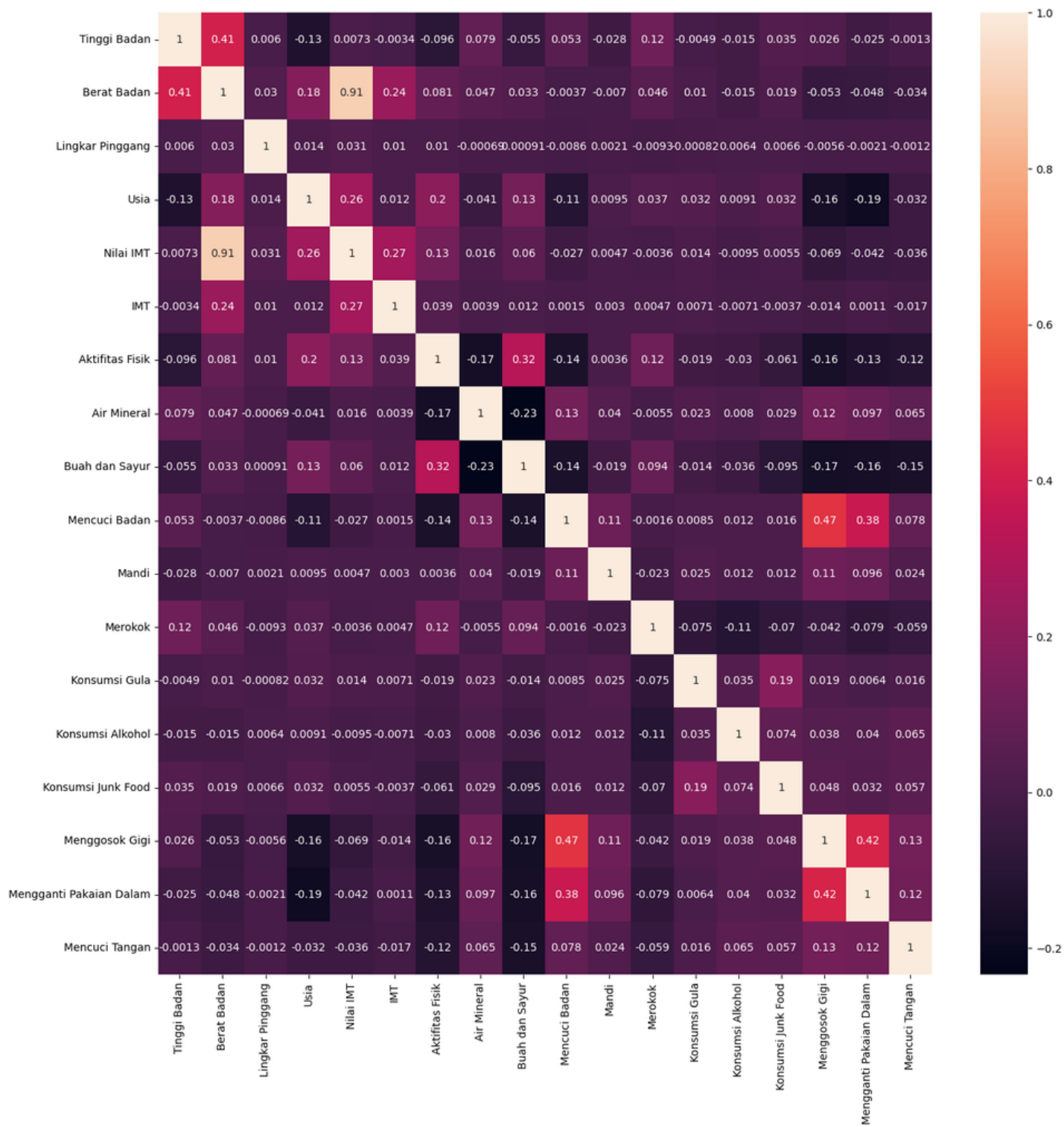
6

SCALING FITUR NUMERIK

Prapengolahan data terakhir yang kami lakukan adalah melakukan *scaling* pada fitur-fitur numerik sehingga memiliki range nilai yang sama agar dapat meningkatkan akurasi *clustering*. Namun, proses ini kami lakukan setelah **data aggregation** agar dapat menampilkan data numerik yang asli ketika ditampilkan

03 - PRAPENGOLAHAN

Setelah menyelesaikan data yang sudah siap untuk digunakan, dilakukan analisis korelasi pada fitur-fitur yang akan digunakan untuk *clustering*.



04 - DATA AGGREGATION

Dilaksanakan agregasi pada data agar dapat dengan mudah mendapatkan ringkasan statistik dari dataset dan dapat merangkum data menjadi informasi yang lebih ringkas

	TINGGI BADAN	BERAT BADAN	USIA	NILAI IMT
mean	169.421749	72.689561	36.874438	25.307553
min	159.500000	45.000000	15.000000	17.240000
max	179.500000	101.000000	66.000000	33.430000

Ringkasan statistik secara umum

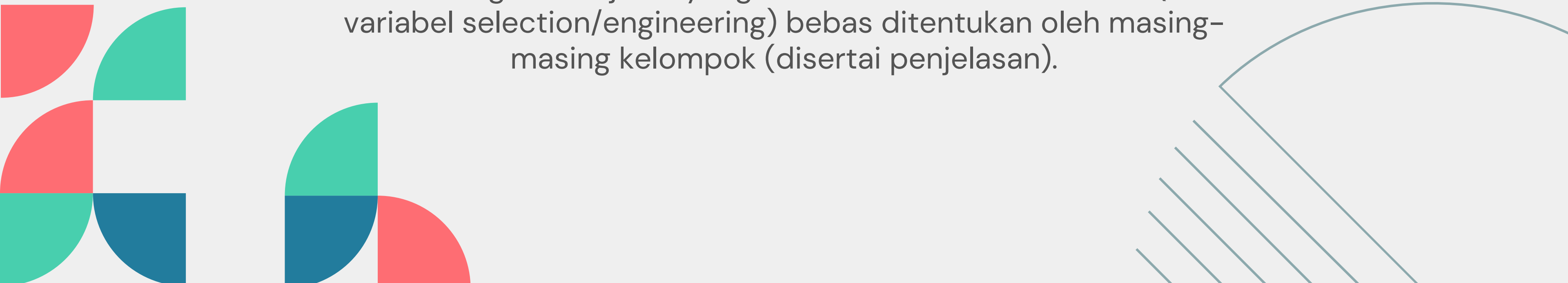
Ringkasan statistik berdasarkan jenis kelamin di mana 0 melambangkan laki-laki dan 1 melambangkan perempuan

JENIS KELAMIN	TINGGI BADAN			BERAT BADAN			USIA			NILAI IMT		
	mean	min	max	mean	min	max	mean	min	max	mean	min	max
0	169.775730	160.0	179.5	73.105289	45.0	101.0	37.027850	15.0	66.0	25.350814	17.24	33.43
1	163.633747	159.5	179.0	65.891923	45.0	100.0	34.365962	15.0	66.0	24.600186	17.26	33.43



TASK 2

Melakukan clustering ($k=4$) pada data yang diberikan dengan menggunakan algoritma k-Means (tanpa module) dan sembarang rumus jarak yang bersesuaian. Parameter lain (dan variabel selection/engineering) bebas ditentukan oleh masing-masing kelompok (disertai penjelasan).



K-MEANS CLUSTERING

FITUR

Pada bagian ini, dilaksanakan k-means clustering dengan jumlah cluster adalah 4. Variabel yang digunakan adalah

- 'JENIS KELAMIN',
- 'TINGGI BADAN',
- 'BERAT BADAN',
- 'USIA',
- 'NILAI IMT',
- 'olahraga',
- 'konsumsi air',
- 'konsumsi buah/sayur',
- 'mandi',
- 'durasi_mandi',
- 'merokok',
- 'konsumsi gula',
- 'konsumsi alkohol',
- 'konsumsi junk food',
- 'gosok gigi', dan
- 'cuci tangan'.

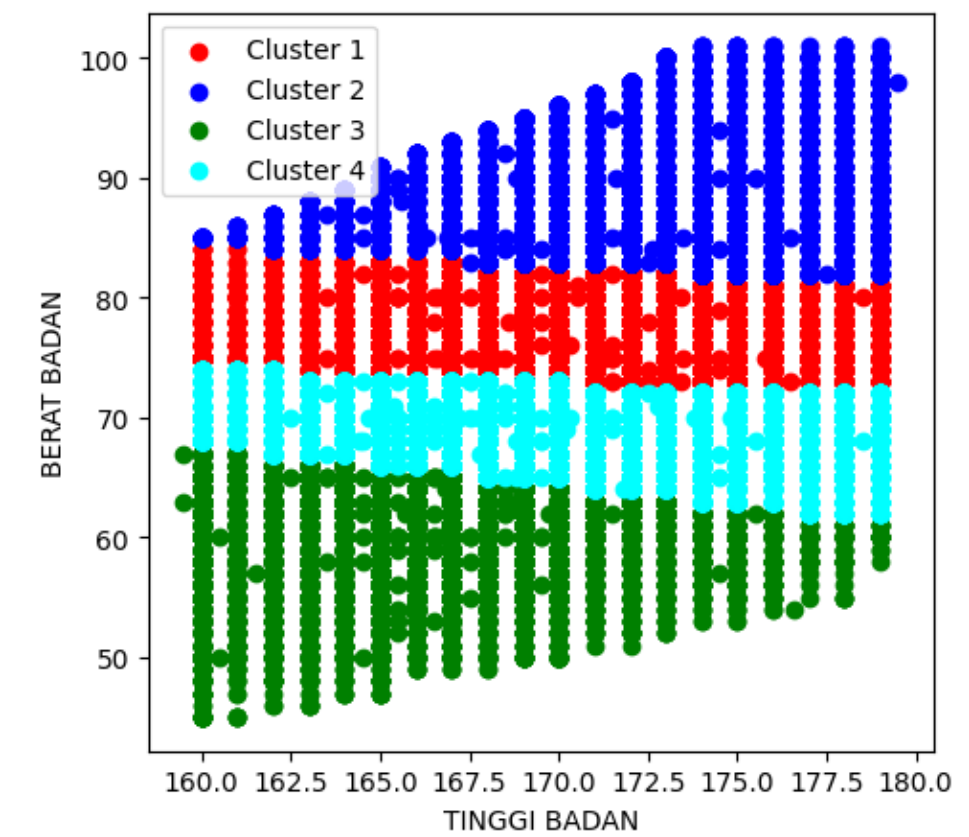
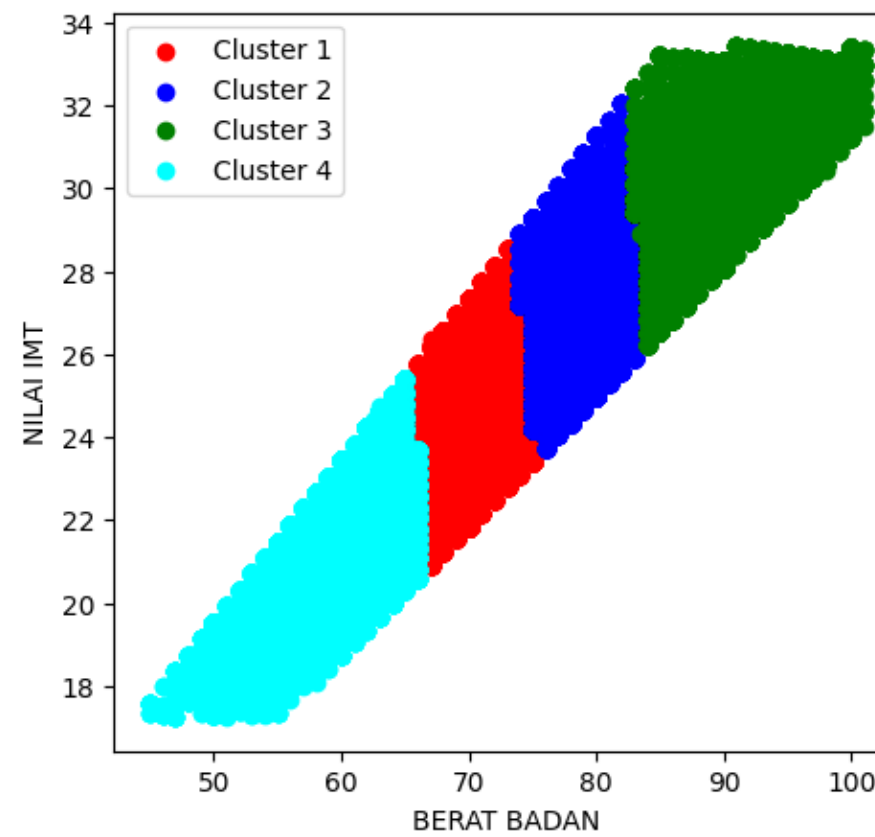
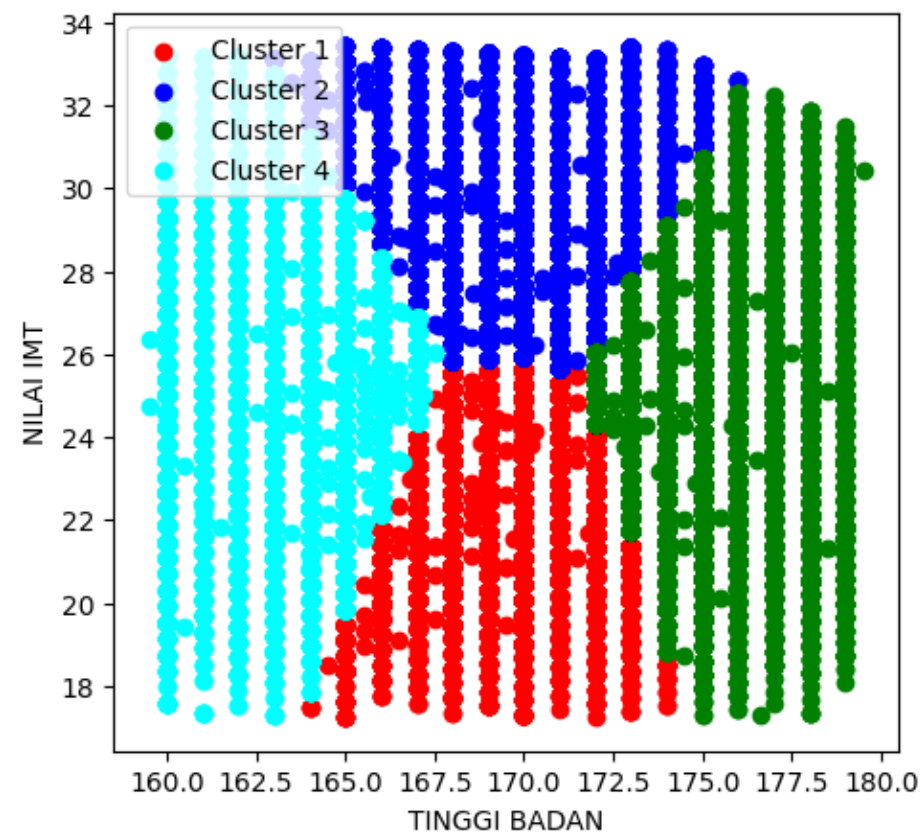
FUNGSI K MEANS

Untuk melakukan clustering tanpa module, pada project ini dibangun sebuah fungsi terlebih dahulu yang dapat melakukan clustering.

- Fungsi KMeansClustering, dimulai dengan menginisialisasi centroid awal secara acak.
- Kemudian, menggunakan rumus **jarak Euclidean** untuk menghitung jarak setiap data ke centroid terdekat.
- setelah itu, centroid akan diperbarui dengan berdasarkan titik rata-rata dalam setiap cluster.
- Pada akhirnya, fungsi akan memberikan keluaran berupa cluster setiap data.

K-MEANS CLUSTERING

Berikut adalah hasil dari k-means clustering yang didapatkan



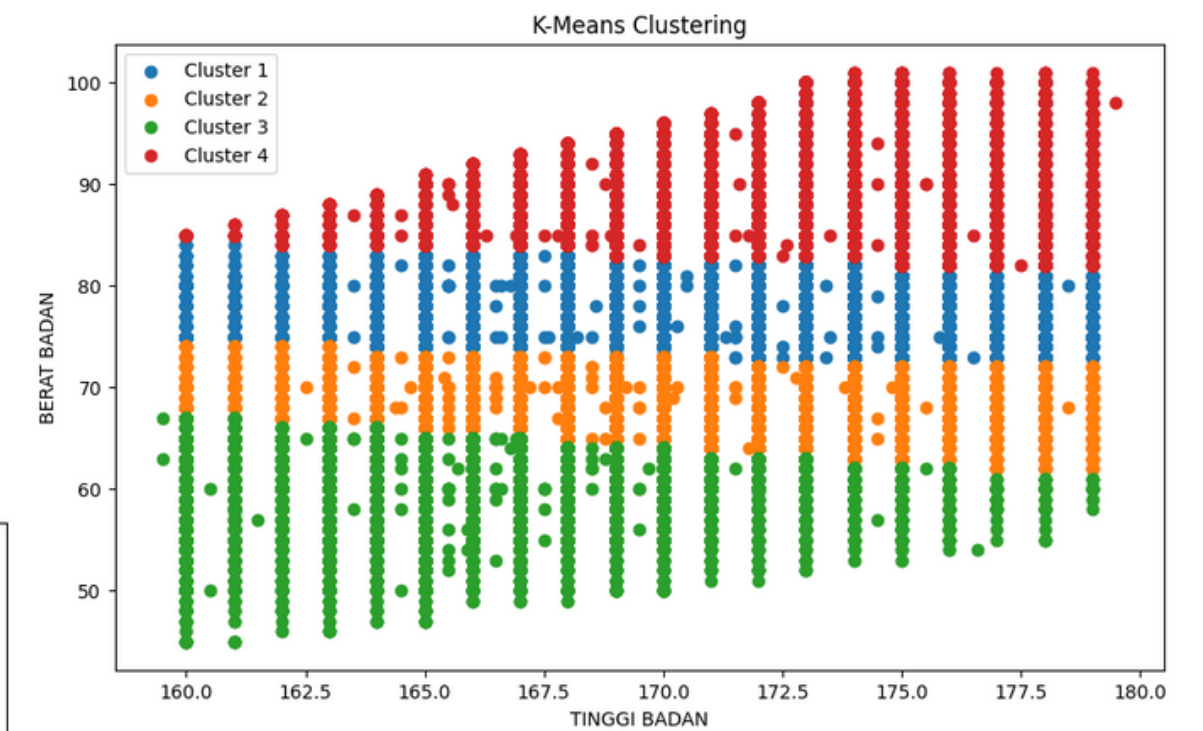
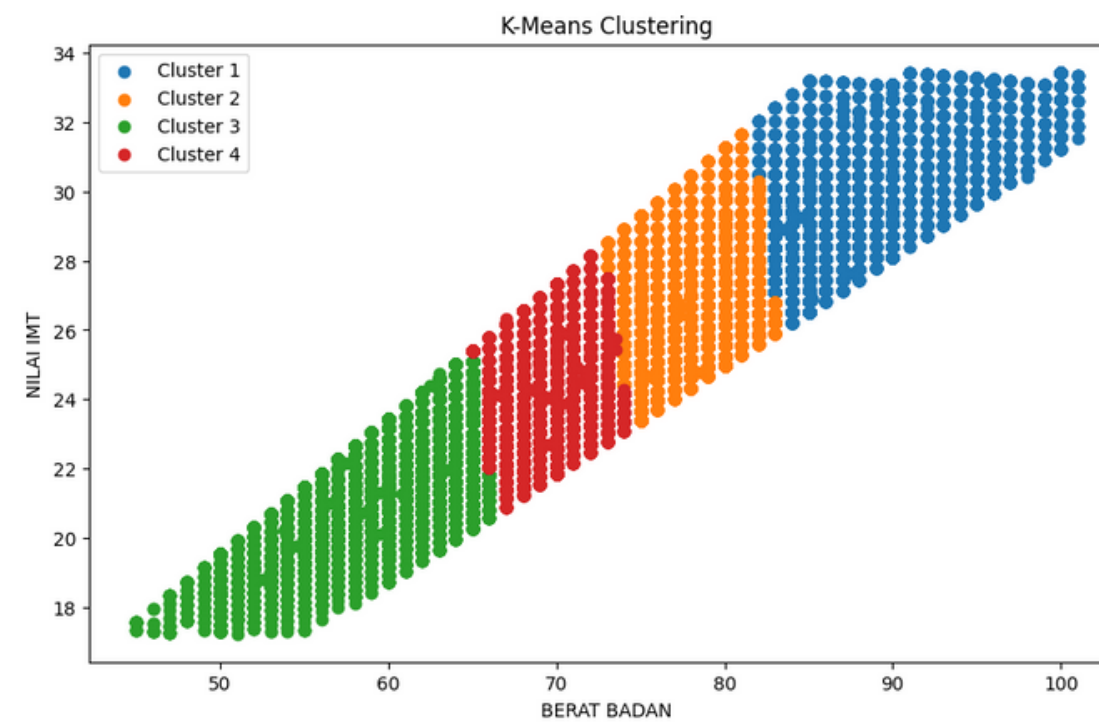
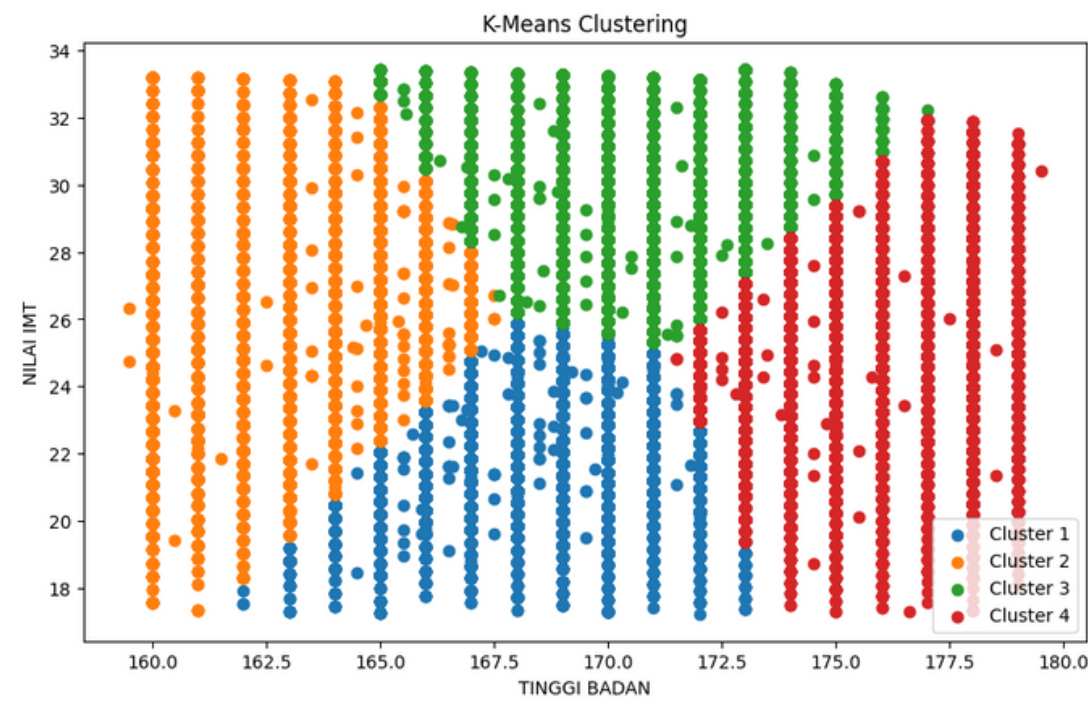


TASK 3

Membandingkan *code* pada TASK 2 (*Non-parallel programming*) dengan pemrograman parallel.

COMPARISON

HASIL CLUSTERING DENGAN PARALLEL PROGRAMMING



COMPARISON

TANPA PARALLEL

```
columns_to_cluster = [  
    ['TINGGI BADAN', 'NILAI IMT'],  
    ['BERAT BADAN', 'NILAI IMT'],  
    ['TINGGI BADAN', 'BERAT BADAN'],  
]  
  
import time  
  
start_time = time.time()  
results = KMeansClustering(df, columns_to_cluster=columns)  
time_non_parallel = time.time() - start_time  
print('durasi: ', time_non_parallel)
```

`durasi: 103.46909713745117`

DENGAN PARALLEL

```
columns_to_cluster = [  
    ['TINGGI BADAN', 'NILAI IMT'],  
    ['BERAT BADAN', 'NILAI IMT'],  
    ['TINGGI BADAN', 'BERAT BADAN'],  
]  
  
import time  
  
start_time = time.time()  
results = KMeansClusteringParallel(data, columns_to_cluster=columns)  
time_parallel = time.time() - start_time  
print('durasi: ', time_parallel)
```

`durasi: 0.5235445499420166`

Terlihat bahwa durasi waktu yang diperlukan dengan program parallel adalah 0.5235445499420166 sementara tanpa program parallel adalah 103.46909713745117

TASK 4

Insight atau informasi yang didapatkan dari TASK 1 dan 2 adalah sebagai berikut

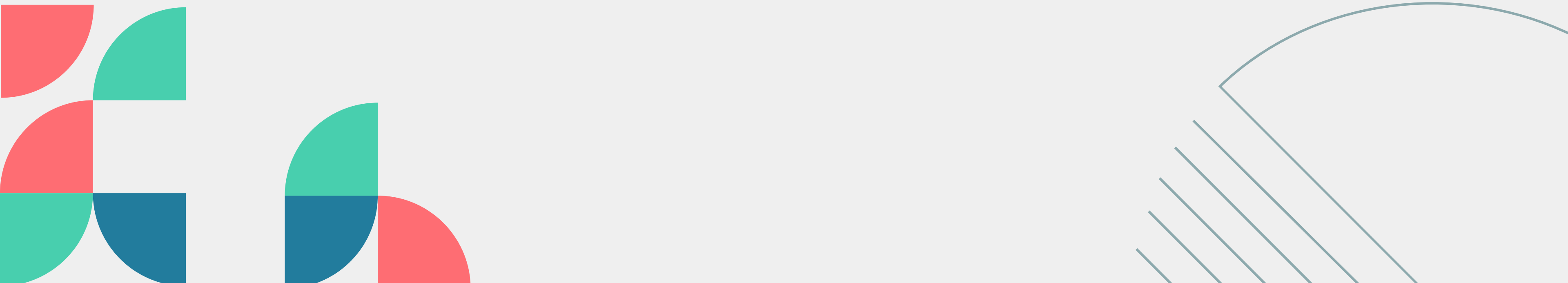
- Terlihat mayoritas data berbentuk kategorik sehingga harus dilakukan pengubahanan type variabel menjadi numerik
- Adanya korelasi yang cukup tinggi antara variabel IMT dan variabel berat badan, Nilai IMT serta aktifitas fisik sehingga variabel tersebut yang dapat berpengaruh terhadap variabel IMT
- Hasil yang didapatkan pada pengelompokan K-Means sudah sesuai ekspektasi, akan tetapi hasil akan lebih baik apabila jumlah iterasi ditingkatkan



KESIMPULAN

Clustering menggunakan algoritma k-Means menunjukkan hasil yang berbeda secara signifikan antara penggunaan tanpa dan dengan *parallel programming*.

Dapat dilihat bahwa waktu yang dibutuhkan k-Means tanpa parallel adalah 103.46909713745117, sementara dengan parallel adalah 0.5235445499420166. Dengan demikian, k-Means yang lebih efisien didapatkan dari dengan program parallel.



The background features four decorative geometric patterns in the corners. The top-left corner has a series of parallel diagonal lines in a light blue-grey color. The top-right corner contains a cluster of overlapping semi-circles in yellow, red, teal, and dark blue. The bottom-left corner also features a cluster of overlapping semi-circles in red, teal, and dark blue. The bottom-right corner has a series of parallel diagonal lines in a light blue-grey color, mirroring the top-left pattern.

**TERIMA
KASIH**