

# IMPLEMENTASI EXPLORATORY DATA ANALYSIS UNTUK ANALISIS DATA LEMAK TUBUH

Anas Syaifudin<sup>1\*)</sup>, Risqiati<sup>2)</sup>, Hermanus Wim Hapsoro<sup>3)</sup>

Institut Widya Pratama<sup>123)</sup>

anzt.07@gmail.com<sup>1)</sup>, risqiati24@gmail.com<sup>2)</sup>, wimhapsoro06@gmail.com<sup>3)</sup>

## Abstrak

*Penelitian ini berfokus pada penerapan metode Exploratory Data Analysis (EDA) untuk menganalisis dataset publik tentang lemak tubuh. Masalah yang diidentifikasi adalah pentingnya memahami pola distribusi dan hubungan antar variabel dalam dataset lemak tubuh. Penelitian ini memanfaatkan dataset lemak tubuh yang mencakup berbagai pengukuran lingkaran tubuh pria, seperti berat badan, tinggi badan, dan persentase lemak tubuh. Solusi yang ditawarkan melibatkan serangkaian langkah EDA, termasuk pembersihan data, analisis univariat dan multivariat, serta visualisasi data menggunakan grafik seperti histogram, scatter plot, dan heatmap. Penanganan outlier dan rekayasa fitur seperti Body Mass Index (BMI) dilakukan untuk meningkatkan kualitas data dan menghasilkan wawasan baru tentang distribusi lemak tubuh. Hasil penelitian menunjukkan hubungan kuat antara persentase lemak tubuh dengan variabel seperti berat badan, lingkaran perut, dan kepadatan tubuh. Sebagian besar individu dalam dataset berada dalam kategori BMI normal, dengan perbedaan distribusi ukuran tubuh berdasarkan kategori BMI. Hasil dari penelitian ini adalah dengan menggunakan EDA yang menyeluruh, yang mencakup analisis univariat dan multivariat, pola dan hubungan antara persentase lemak tubuh dan variabel lain seperti berat badan, tinggi badan, usia, dan lingkaran tubuh telah ditemukan. Penelitian ini merekomendasikan pengembangan model prediktif berbasis machine learning dengan menggunakan dataset lemak tubuh.*

**Kata kunci:** Exploratory Data Analysis, EDA, Data Lemak Tubuh.

## 1. Pendahuluan

Dalam era Big Data, kemampuan untuk memahami dan menafsirkan data menjadi semakin penting, terutama dalam domain kesehatan dan kebugaran. Salah satu aspek penting dari kesehatan adalah berat badan, yang sering kali diukur melalui Persentase Lemak Tubuh (*Body Fat Percentage*) (Widiastuti, Priyambodo, and Buanayuda 2020). Lemak tubuh merupakan parameter yang krusial karena berkaitan langsung dengan risiko berbagai penyakit, seperti kardiovaskular, stroke iskemik, dan diabetes melitus tipe 2 serta obesitas (Salim, Wihandani, and Dewi 2021). Pemahaman yang mendalam mengenai distribusi dan faktor-faktor yang mempengaruhi lemak tubuh dapat membantu dalam pengambilan keputusan medis

serta perancangan program kesehatan yang lebih efektif (Powell-Wiley et al. 2021).

Untuk menganalisis data lemak tubuh secara efektif, implementasi Exploratory Data Analysis (EDA) menjadi sangat penting (Sagala and Aryatama 2022). EDA merupakan pendekatan analisis statistik yang digunakan untuk memahami struktur data, menemukan pola, mengidentifikasi anomali, dan menguji asumsi-asumsi statistik sebelum dilakukan analisis lebih lanjut (Rahmany, Zin, and Sundararajan 2020). Dalam konteks dataset lemak tubuh. Cara ini memungkinkan peneliti untuk mendapatkan wawasan awal tentang distribusi data, keterkaitan antar variabel, serta potensi *outlier* yang dapat mempengaruhi hasil analisis (Ramadhani, Ramadhanu, and Hidayat 2024). Oleh karena itu, EDA tidak hanya memberikan fondasi yang kuat

untuk analisis lebih lanjut, tetapi juga memastikan bahwa data yang digunakan dalam penelitian lebih bersih dan siap untuk diproses lebih lanjut.

Penelitian sebelumnya telah menggunakan EDA untuk menganalisis data tentang lemak tubuh. Seperti yang dilakukan oleh Hedgepeth (2016) melaporkan analisis hubungan antara usia dan persentase lemak tubuh serta perbedaan berdasarkan jenis kelamin. Metode ini, tidak dapat mengeksplorasi pola multidimensi yang lebih kompleks karena terbatas pada hubungan bivariat antar variabel. Meskipun demikian, penelitian oleh Venables and Ripley (2002) menunjukkan bahwa analisis multivariat sangat penting untuk memahami data secara menyeluruh, tetapi masih terbatas untuk digunakan pada dataset publik seperti dataset lemak tubuh.

Selain itu, seringkali eksplorasi data multidimensi memerlukan teknik yang lebih canggih untuk menemukan pola tersembunyi yang mungkin tidak terlihat melalui analisis konvensional (Valis, Zak, and Vint 2019). Sebagian besar penelitian saat ini belum menggunakan metode eksplorasi ini secara menyeluruh dalam konteks dataset lemak tubuh.

Dalam penelitian ini, dataset publik tentang lemak tubuh digunakan untuk mengeksplorasi hubungan antar variabel secara mendalam, baik dalam dimensi tunggal maupun multidimensi. Dengan memanfaatkan metode EDA, penelitian ini bertujuan untuk mengidentifikasi pola penting, mengungkap anomali, dan menghasilkan visualisasi yang dapat memberikan wawasan baru yang sebelumnya belum banyak dibahas. Penelitian ini tidak hanya bertujuan untuk memberikan kontribusi pada literatur terkait analisis dataset lemak tubuh, tetapi juga mendemonstrasikan pentingnya pendekatan eksplorasi data multidimensi untuk aplikasi di dunia nyata.

## 2. Metode Penelitian

Dalam melakukan penelitian ini disusun rancangan atau kerangka pemikiran untuk sebagai acuan dalam menyelesaikan penelitian, agar sesuai dengan pokok bahasan yang diteliti. Berikut ini menunjukkan kerangka pemikiran yang digunakan dalam penelitian ini.



Gambar 1. Alur Penelitian

### 2.1 Identifikasi Masalah

Penelitian dimulai dengan mengidentifikasi pentingnya memahami distribusi lemak tubuh dan faktor-faktor yang mempengaruhinya dalam konteks kesehatan. Ini menjadi dasar untuk melakukan *Exploratory Data Analysis* (EDA).

### 2.2 Pengumpulan Data

Data yang digunakan dalam penelitian ini adalah dataset lemak tubuh yang mencakup berbagai variabel seperti density, body fat, umur, berat badan, tinggi badan, lingkar leher, lingkar dada, lingkar perut, lingkar pinggul, lingkar paha, lingkar lutut, pergelangan kaki, lingkar bisep, lingkar lengan bawah, lingkar pergelangan tangan. Dataset merupakan dataset publik yang diambil dari kaggle.com

### 2.3 Penerapan EDA

Tahap awal EDA melibatkan pembersihan data, termasuk penanganan *missing values*, *outlier*, dan inkonsistensi data. Selanjutnya dilakukan analisis deskriptif untuk memahami distribusi data dan statistik dasar seperti mean, median, dan *standard deviation* dari masing-masing variabel. Kemudian data divisualisasikan

menggunakan grafik seperti histogram, scatter plot, dan box plot untuk mendeteksi pola, tren, dan anomali dalam data. Dilakukan analisis korelasi untuk memahami hubungan antara variabel seperti BMI dan persentase lemak tubuh.

## 2.4 Kesimpulan dan Rekomendasi

Kesimpulan diambil berdasarkan temuan penelitian, dan rekomendasi dibuat untuk penelitian lanjutan serta aplikasi praktis dalam pengelolaan kesehatan. Rekomendasi yang dihasilkan berupa pengolahan lanjutan ataupun penerapan algoritma berdasarkan temuan.

## 3. Hasil dan Pembahasan

Dari penelitian yang sudah dilakukan didapatkan hasil sesuai dengan tahapan penelitian sebagai berikut:

### 3.1 Identifikasi Masalah

Dalam melakukan sebuah analisis pada bidang *data science* penting untuk mendefinisikan latar belakang masalah (*problem statement*) yang akan dipecahkan. Dengan adanya *problem statement* yang jelas dapat mencegah analisis yang dihasilkan menjadi lebih terarah dan tujuan yang jelas. Pada penelitian ini permasalahan yang akan dipecahkan yaitu Bagaimana penerapan metode EDA dapat digunakan untuk menggali pola dan hubungan antar variabel dalam dataset lemak tubuh.

Untuk dapat menyelesaikan permasalahan yang akan dipecahkan, maka dirumuskan tujuan (*goals*) sebagai berikut:

- Menentukan karakteristik distribusi dari setiap variabel numerik dalam dataset lemak tubuh menggunakan metode EDA.
- Mengidentifikasi dan memvisualisasikan hubungan antar variabel numerik, khususnya korelasi antara persentase lemak tubuh (BodyFat) dengan variabel lainnya.
- Menyelidiki bagaimana kategori Body Mass Index (BMI) berpengaruh terhadap persentase lemak tubuh.
- Menentukan keberadaan outlier dalam setiap variabel numerik menggunakan metode visualisasi.
- Membuat ringkasan hasil analisis: Membuat deskripsi yang dapat dibaca dan dipahami dengan mudah, dengan tujuan agar orang awam dapat memahami hasil analisis.

### 3.2 Pengumpulan Data

Data yang akan digunakan pada penelitian ini menggunakan data pengukuran lingkaran tubuh pria sebanyak 252 orang. Data yang digunakan mencakup berbagai variabel seperti density, body fat, umur, berat badan, tinggi badan, lingkaran leher, lingkaran dada, lingkaran perut, lingkaran pinggul, lingkaran paha, lingkaran lutut, pergelangan kaki, lingkaran bisep, lingkaran lengan bawah, lingkaran pergelangan tangan. Data yang akan digunakan dapat dilihat pada Tabel 1 yang berisi data pengukurannya.

Tabel 1. Dataset Lemak Tubuh

Density	Body Fat	Age	Weight	Height	Neck	Chest	Abdomen	Hip	Thigh	Knee	Ankle	Biceps	Forearm	Wrist
1.0708	12.3	23	154.25	67.75	36.2	93.1	85.2	94.5	59	37.3	21.9	32	27.4	17.1
1.0853	6.1	22	173.25	72.25	38.5	93.6	83	98.7	58.7	37.3	23.4	30.5	28.9	18.2
1.0414	25.3	22	154	66.25	34	95.8	87.9	99.2	59.6	38.9	24	28.8	25.2	16.6
1.0751	10.4	26	184.75	72.25	37.4	101.8	86.4	101.2	60.1	37.3	22.8	32.4	29.4	18.2
1.034	28.7	24	184.25	71.25	34.4	97.3	100	101.9	63.2	42.2	24	32.2	27.7	17.7
....	....	....	....	....	....	....	....	....	....	....	....	....	....	....
1.0328	29.3	72	186.75	66	38.9	111.1	111.5	101.7	60.3	37.3	21.5	31.3	27.2	18
1.0399	26	72	190.75	70.5	38.9	108.3	101.3	97.8	56	41.6	22.7	30.5	29.4	19.8
1.0271	31.9	74	207.5	70	40.8	112.4	108.5	107.1	59.3	42.2	24.6	33.7	30	20.9

### 3.3 Penerapan EDA

#### a. Data Understanding

Pada tahapan pemahaman data (Data Understanding) kita akan memahami atribut atau variabel dalam dataset lemak tubuh, seperti nama, tipe data (numerik, kategori, teks), dan arti dari setiap kolom/ fitur. Penjelasan dari setiap fitur

yang ada pada dataset lemak tubuh dapat dilihat pada Tabel 2 berikut ini:

Tabel 2. Deskripsi Fitur Dataset Lemak Tubuh

Fitur	Keterangan
Density	Massa jenis ditentukan dari penimbangan bawah air
BodyFat	Persentase lemak tubuh dari persamaan Siri (1956)

Age	Merepresentasikan umur dalam tahun
Weight	Merepresentasikan berat badan dalam pound
Height	Merepresentasikan tinggi badan dalam inci
Neck	Merepresentasikan lingkaran leher dalam centimeter
Chest	Merepresentasikan lingkaran dada dalam centimeter
Abdomen	Merepresentasikan lingkaran perut dalam centimeter
Hip	Merepresentasikan lingkaran pinggul dalam centimeter
Thigh	Merepresentasikan lingkaran paha dalam centimeter
Knee	Merepresentasikan lingkaran lutut dalam centimeter
Angkle	Merepresentasikan pergelangan kaki dalam centimeter

Biceps	Merepresentasikan lingkaran bicep dalam centimeter
Forearm	Merepresentasikan lingkaran lengan bawah dalam centimeter
Wrist	Merepresentasikan lingkaran pergelangan tangan dalam centimeter

Dengan mendapatkan nilai-nilai statistiknya dapat dilihat pada Tabel 3 seperti *count*, *mean*, standar deviasi, nilai kuartil, *min* dan *max*. Dengan mengetahui nilai-nilai ini kita dapat mengetahui sebaran data. Dalam python, dapat menggunakan fungsi *describe()* untuk menampilkan informasi tersebut.

Tabel 3. Sebaran Data Dataset Lemak Tubuh

	Density	Body Fat	Age	Weight	Height	Neck	Chest	Abdomen	Hip	Thigh	Knee	Ankle	Biceps	Forearm	Wrist
Count	250	250	250	250	250	250	250	250	250	250	250	250	250	250	250
Mean	1.05	19.15	44.88	178.92	70.14	37.99	100.82	92.55	99.90	59.40	38.59	23.10	32.27	28.66	18.22
Std	0.019	8.368	12.60	29.38	3.662	2.430	8.430	10.78	7.164	5.249	2.411	1.694	3.021	2.020	0.933
Min	0.995	0.0	22	118.5	29.5	31.1	79.3	69.4	85	47.2	33	19.1	24.8	21	15.8
25%	1.0414	12.47	35.75	159	68.25	36.4	94.35	84.575	95.5	56	36.97	22	30.2	27.3	17.6
50%	1.0549	19.2	43	176.5	70	38	99.65	90.95	99.3	59	38.5	22.8	32.05	28.7	18.3
75%	1.0704	25.3	54	197	72.25	39.42	105.37	99.32	103.5	62	39.92	24	34.32	30	18.8
Max	1.1089	47.5	81	363.15	77.75	51.2	136.2	148.1	147.7	87	49.1	33.9	45	34.9	21.4

### b. Data Preprocessing

Pada proses ini memiliki tujuan membersihkan, mengubah dan mengorganisir data agar dapat digunakan secara optimal. Tahap pertama akan dilakukan pengecekan nilai yang kosong (missing value) dengan menggunakan fungsi *isnull().sum()* untuk menjumlahkan setiap fitur memiliki data kosong.

Hasil pengecekan nilai *null* menghasilkan nilai 0 hal ini menunjukkan bahwa pada setiap fitur tidak menunjukkan adanya data yang kosong. Selanjutnya akan membuat sebuah fitur baru (*feature engineering*), hal ini dilakukan untuk mendapatkan informasi yang lebih informatif dan representatif yaitu membuat fitur *Body Mass Indeks* (BMI) dengan Persamaan 1 di bawah ini.

$$BMI = \frac{\text{Berat Badan}}{(\text{Tinggi Badan})^2} \quad (1)$$

Pada perhitungan BMI satuan berat yang semula pound akan dikonversi ke dalam kilogram

dan satuan tinggi badan yang semula inch akan dikonversi ke dalam meter. Hasil dari *feature engineering* dapat dilihat pada Tabel 4 seperti di bawah ini.

Tabel 4. Hasil *Feature Engineering*

Weight_kg	Height_kg	BMI
69.954649	1.720853	23.622661
78.571429	1.835154	23.330271
69.841270	1.682753	24.664436
83.786848	1.835154	24.878889
83.560091	1.809754	25.512910
...	...	...
84.693878	1.676403	30.136650
86.507937	1.790704	26.977925
94.104308	1.778004	29.767630

Setelah nilai BMI didapatkan, dapat kita buat fitur baru yaitu BMI Kategori dapat dilihat pada Tabel 5. Fitur ini akan kita gunakan untuk mengelompokkan data berdasarkan 4 kategori berdasarkan nilai BMI (Ramírez et al. 2020), yaitu:

- 1) Jika nilai BMI kurang dari 18.5 masuk ke dalam kategori *Underweight*.
- 2) Jika nilai BMI lebih dari atau sama dengan 18.5 dan kurang dari 25 masuk ke dalam kategori Normal.
- 3) Jika nilai BMI lebih dari atau sama dengan 25 dan kurang dari 30 masuk ke dalam kategori *Overweight*.
- 4) Jika nilai BMI lebih dari atau sama dengan 30 masuk ke dalam kategori *Obese*.

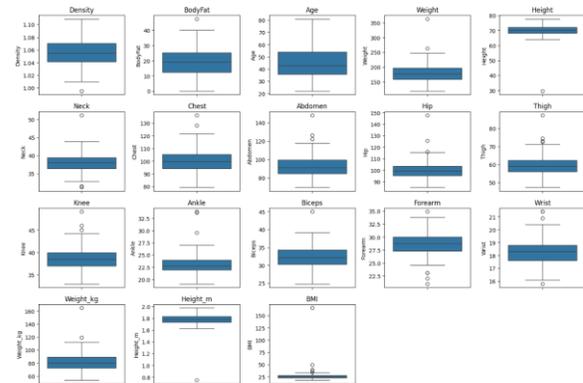
Tabel 5. Pembuatan Fitur BMI Kategori

BMI	BMI_Kat
23.622661	Normal
23.330271	Normal
24.664436	Normal
24.878889	Normal
25.512910	Overweight
...	...
30.136650	Obese
26.977925	Overweight
29.767630	Overweight

### c. Visualisasi Data

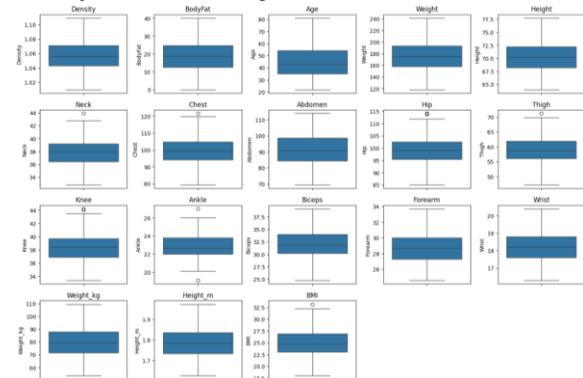
Setelah semua fitur siap, selanjutnya kita akan masuk ke dalam tahap visualisasi data untuk mempermudah dalam mengetahui distribusi data dari dataset Lemak Tubuh.

Langkah pertama adalah melakukan pengecekan apakah ada data pencilan (*outlier*) pada dataset. Pengecekan *outlier* pada setiap fitur menggunakan box plot (Gambar 1). Jika dalam sebuah dataset terdapat sebuah *outlier* maka dapat mempengaruhi analisis statistik yang dihasilkan. Pada penelitian ini penanganan *outlier* akan dilakukan dengan metode IQR (*Interquartile Range*). Kemudian data *outlier* akan dihapus dari dataset sehingga total data yang semula 252 data menjadi 233 data.



Gambar 1. Pengecekan Outlier Dataset Lemak Tubuh menggunakan Box Plot

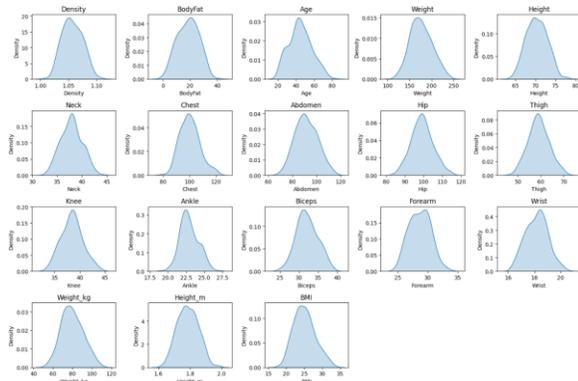
Box Plot menunjukkan adanya outlier pada beberapa fitur seperti BodyFat, Density, Age, Weight, Height, Neck, Abdomen, Hip, Thigh, Knee, Ankle, Biceps, Forearm, dan Wrist. Outlier ini ditunjukkan oleh titik-titik yang berada di luar batas atas dan bawah box plot. Hasil dari penerapan metode IQR dapat dilihat pada Gambar 2, terdapat data outlier tetapi rentang nilainya tidak terlalu jauh.



Gambar 2. Box Plot setelah Outlier dihapus

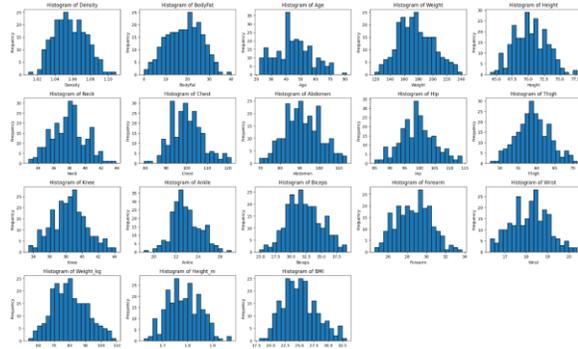
### 1) Analisis Univariat

Analisis Univariat adalah analisis yang berfokus pada satu variabel saja. Tujuannya adalah memahami karakteristik distribusi, pola, atau tren variabel tersebut.



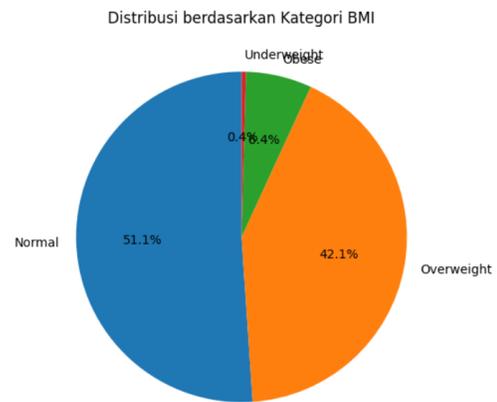
Gambar 3. Density Plot Fitur Dataset Lemak Tubuh

Density plot pada Gambar 3 menunjukkan distribusi dari setiap fitur numerik. Beberapa fitur seperti *Age*, *Weight*, *Height*, *Abdomen*, *Hip*, dan *Thigh* menunjukkan distribusi yang mendekati normal. Fitur-fitur lainnya seperti *BodyFat*, *Density*, *Neck*, *Chest*, *Knee*, *Ankle*, *Biceps*, *Forearm*, dan *Wrist* menunjukkan distribusi yang sedikit miring.



Gambar 4. Histogram Plot Fitur Dataset Lemak Tubuh

Berdasarkan histogram yang dihasil pada Gambar 4 menunjukkan beberapa pola seperti Variabel seperti *Weight* dan *Height* menunjukkan rentang yang lebih luas dibandingkan variabel seperti *Wrist* dan *Ankle*.

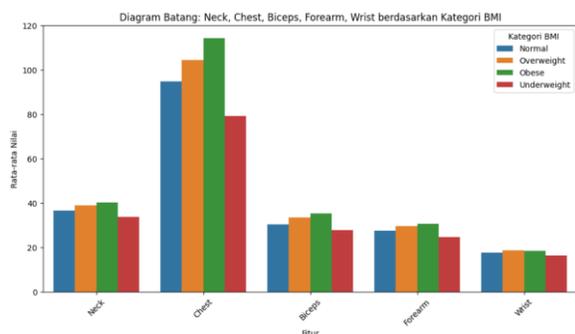


Gambar 5. Distribusi Kategori BMI Dataset Lemak Tubuh

Pada Gambar 5 menunjukkan distribusi frekuensi dari kategori Indeks Massa Tubuh (BMI) pada dataset yang digambarkan dalam bentuk bar chart dan pie chart. Terlihat bahwa kategori Normal memiliki frekuensi tertinggi, diikuti oleh *Overweight* dan *Obese*. Kategori *Underweight* memiliki frekuensi terendah. Dengan menggunakan pie chart dapat memberikan gambaran singkat tentang proporsi individu dalam setiap kategori BMI dalam dataset.

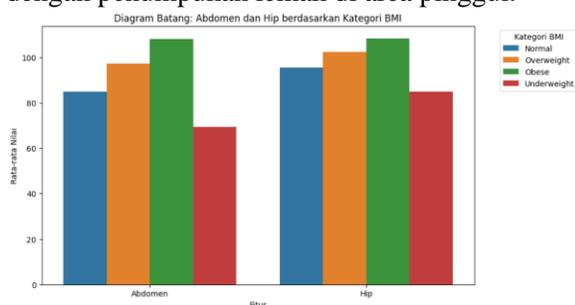
## 2) Analisis Multivariat

Analisis Multivariat digunakan untuk menganalisis hubungan antara lebih dari dua variabel. Kita akan melakukan analisis hubungan antara lingkaran tubuh bagian atas (*Neck*, *Chest*, *Biceps*, *Forearm*, *Wrist*), lingkaran tubuh bagian tengah (*Abdomen*, *Hips*) dan lingkaran tubuh bagian bawah (*Thigh*, *Knee*, *Ankle*) sebagian variabel dependen dengan variabel kategori BMI sebagai variabel independen. Hubungan antara lingkaran tubuh bagian atas dan kategori BMI dapat dilihat pada Gambar 6. Secara umum, orang dalam kategori BMI yang lebih tinggi (*Overweight* dan *Obese*) cenderung memiliki ukuran lingkaran tubuh yang lebih besar dibandingkan dengan orang dalam kategori BMI yang lebih rendah (Normal dan *Underweight*). Hal ini menunjukkan bahwa peningkatan BMI sering kali dikaitkan dengan ukuran lingkaran tubuh bagian atas. Terdapat sedikit persamaan pada fitur *Wrist* pada kategori BMI *Overweight* dan *Obese*.



Gambar 5.6 Diagram Batang Lingkaran Tubuh Bagian Atas dengan Kategori BMI

Analisis kedua dilakukan pada kategori fitur lingkaran tubuh bagian tengah dengan kategori BMI. Diagram batang pada Gambar 7 menunjukkan rata-rata lingkaran *Abdomen* (perut) dan *Hip* (pinggul) untuk setiap kategori BMI (*Underweight*, *Normal*, *Overweight*, dan *Obese*). Perbedaan lingkaran perut antara kategori BMI *Obese* dan *Underweight* lebih besar daripada perbedaan lingkaran pinggul. Individu dalam kategori BMI *Obese* memiliki lingkaran perut dan pinggul yang paling besar, diikuti oleh kategori *Overweight*, *Normal*, dan *Underweight*. Peningkatan lingkaran perut rata-rata antara kategori BMI *Obese* dan *Underweight* lebih besar daripada perbedaan lingkaran pinggul rata-rata. Hasil ini menunjukkan bahwa penumpukan lemak di area perut merupakan faktor yang lebih dominan dalam peningkatan BMI dibandingkan dengan penumpukan lemak di area pinggul.

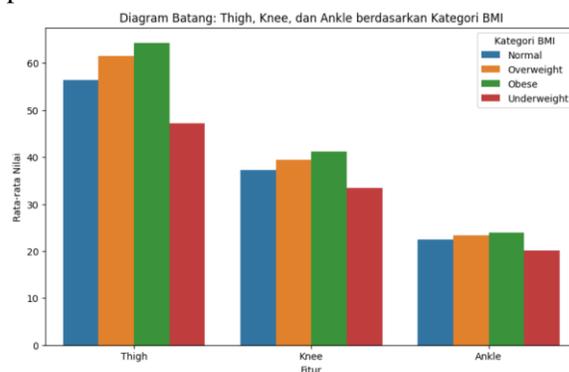


Gambar 7. Diagram Batang Lingkaran Tubuh Bagian Tengah dengan Kategori BMI

Analisis ketiga dilakukan pada kategori fitur lingkaran tubuh bagian bawah dengan kategori BMI. Pada Gambar 8 menunjukkan rata-rata lingkaran *Thigh* (paha), *Knee* (lutut), dan *Ankle* (pergelangan kaki) untuk setiap kategori BMI (*Underweight*, *Normal*, *Overweight*, dan *Obese*). Rata-rata lingkaran *Thigh*, *Knee*, dan *Ankle*

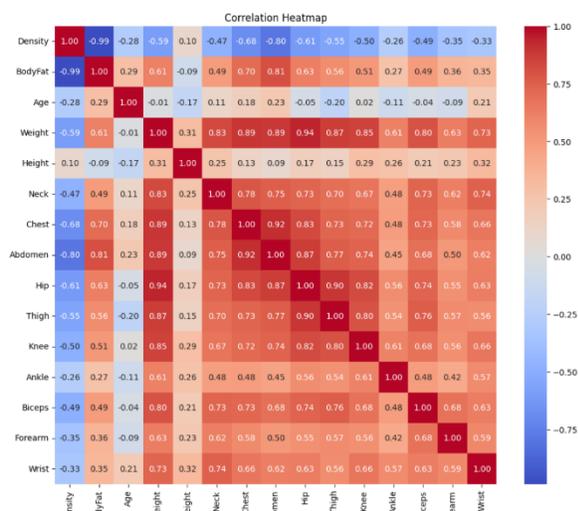
meningkat seiring dengan peningkatan kategori BMI. Individu dengan kategori BMI *Obese* memiliki rata-rata lingkaran paha, lutut, dan pergelangan kaki yang paling besar, diikuti oleh *Overweight*, *Normal*, dan terakhir *Underweight*. Peningkatan lingkaran paha paling signifikan di antara ketiga fitur tersebut. Perbedaan rata-rata lingkaran paha antara kategori *Obese* dan *Underweight* lebih besar dibandingkan dengan perbedaan rata-rata lingkaran lutut dan pergelangan kaki.

Lingkaran pergelangan kaki menunjukkan peningkatan yang relatif kecil antar kategori BMI. Individu dengan kategori BMI yang lebih tinggi cenderung memiliki ukuran lingkaran paha, lutut, dan pergelangan kaki yang lebih besar. Namun, peningkatan paling signifikan terlihat pada lingkaran paha, yang mengindikasikan bahwa area tersebut mungkin lebih sensitif terhadap perubahan BMI.



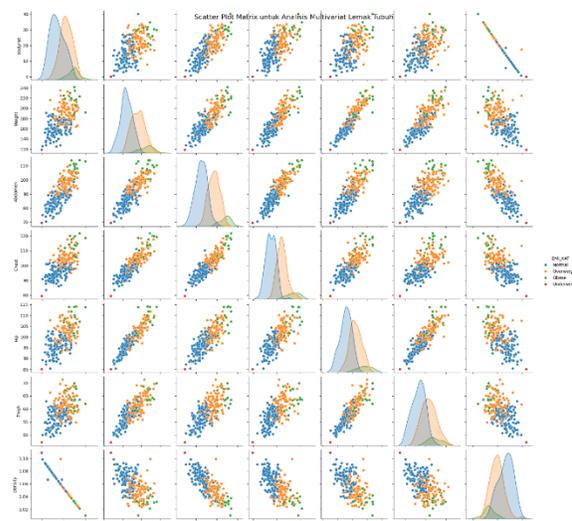
Gambar 8. Diagram Batang Lingkaran Tubuh Bagian Bawah dengan Kategori BMI

Selanjutnya akan dilakukan analisis korelasi antar fitur-fitur yang ada pada dataset menggunakan diagram Heatmap. Diagram ini dapat digunakan untuk memvisualisasikan matriks korelasi, yang menunjukkan kekuatan dan arah hubungan linear antara variabel numerik pada dataset.



Gambar 9. Heatmap Matrik Korelasi Fitur Dataset Lemak Tubuh

Warna pada heatmap menunjukkan kekuatan dan arah korelasi. Angka di dalam setiap sel heatmap adalah koefisien korelasi. Koefisien korelasi berkisar antara -1 hingga +1. *BodyFat* memiliki korelasi positif yang kuat dengan *Weight*, *Abdomen*, *Chest*, *Hip*, dan *Thigh*. Ini menunjukkan bahwa orang dengan berat badan, lingkaran perut, dada, pinggul, dan paha yang lebih besar cenderung memiliki persentase lemak tubuh yang lebih tinggi. Ada juga korelasi positif kuat antara variabel lingkaran tubuh (misalnya, *Abdomen* dan *Chest*, *Hip* dan *Thigh*). Ini menunjukkan bahwa ukuran lingkaran tubuh cenderung meningkat bersamaan. Dengan memahami korelasi ini, kita dapat memperoleh wawasan tentang faktor-faktor yang memengaruhi persentase lemak tubuh.



Gambar 10. Scatter Plot Matrik Fitur Dataset Lemak Tubuh

Scatter plot matrix pada Gambar 5.10 menampilkan hubungan antar variabel dalam dataset lemak tubuh. Variabel yang divisualisasikan meliputi *BodyFat*, *Weight*, *Abdomen*, *Chest*, *Hip*, *Thigh*, dan *Density*, dengan warna titik menunjukkan kategori BMI (*BMI\_KAT*). Secara umum, terlihat bahwa *BodyFat* memiliki korelasi positif yang kuat dengan *Weight*, *Abdomen*, *Chest*, *Hip*, dan *Thigh*. Artinya, semakin tinggi nilai variabel-variabel tersebut, semakin tinggi pula persentase lemak tubuh. Hal ini terlihat dari pola titik-titik yang cenderung naik dari kiri bawah ke kanan atas pada scatter plot yang melibatkan *BodyFat* dan variabel-variabel tersebut. Sebaliknya, *BodyFat* memiliki korelasi negatif yang kuat dengan *Density*. Artinya, semakin tinggi kepadatan tubuh, semakin rendah persentase lemak tubuh. Pola titik-titik yang cenderung turun dari kiri atas ke kanan bawah pada scatter plot *BodyFat* vs. *Density* mengonfirmasi hal ini.

### 3.4 Kesimpulan dan Rekomendasi

Sebagian besar individu dalam dataset termasuk dalam kategori BMI normal, diikuti oleh *overweight*, *obese*, dan terakhir *underweight*. Hal ini menunjukkan bahwa dataset cenderung merepresentasikan populasi umum dengan proporsi individu yang sehat dan obesitas. Terdapat hubungan yang kuat antara persentase lemak tubuh (*body fat*) dengan berat

badan (*weight*), lingkaran perut (*abdomen*), lingkaran dada (*chest*), lingkaran pinggul (*hip*), dan lingkaran paha (*thigh*). Individu dengan berat badan dan lingkaran tubuh yang lebih besar cenderung memiliki persentase lemak tubuh yang lebih tinggi. Kepadatan tubuh (*density*) memiliki hubungan yang kuat dan berlawanan dengan persentase lemak tubuh (*body fat*). Semakin tinggi kepadatan tubuh, semakin rendah persentase lemak tubuh.

Penting untuk mempertimbangkan variabel-variabel seperti berat badan, lingkaran perut, dada, pinggul, dan paha dalam strategi pencegahan dan penanganan obesitas. Variabel-variabel ini memiliki korelasi kuat dengan persentase lemak tubuh dan dapat menjadi target intervensi. Pengukuran BMI dapat menjadi indikator awal untuk mengidentifikasi individu yang berisiko mengalami obesitas dan masalah kesehatan terkait.

Untuk rekomendasi penerapan machine learning dapat melakukan pengembangan untuk model regresi untuk memprediksi persentase lemak tubuh berdasarkan variabel-variabel seperti berat badan, lingkaran tubuh, usia, dan tinggi badan. Serta model klasifikasi untuk mengklasifikasikan individu ke dalam kategori BMI (*underweight*, normal, *overweight*, *obese*) berdasarkan variabel-variabel yang relevan.

#### 4. Kesimpulan dan Saran

Hasil dari implementasi EDA untuk dataset Lemak Tubuh dapat menjadi contoh penerapan EDA yang komprehensif. Dimulai dengan pemahaman data yang baik, termasuk deskripsi variabel dan rekayasa fitur seperti BMI dan kategori BMI. Anda telah menunjukkan analisis univariat untuk setiap variabel, dan analisis multivariat untuk memahami hubungan antar variabel. Visualisasi informatif seperti diagram batang, diagram lingkaran, boxplot, heatmap, dan scatter plot digunakan untuk menyajikan informasi dengan jelas.

Kualitas dari dataset lemak tubuh semakin meningkat. Dengan memastikan tidak ada *missing values* dalam dataset, dan merekayasa fitur baru (BMI dan kategori BMI) untuk meningkatkan nilai informasi. Penanganan

outlier menggunakan metode IQR juga berperan penting dalam meningkatkan kualitas analisis. Dengan menggunakan EDA yang menyeluruh, yang mencakup analisis univariat dan multivariat, pola dan hubungan antara persentase lemak tubuh dan variabel lain seperti berat badan, tinggi badan, usia, dan lingkaran tubuh telah ditemukan.

Selanjutnya untuk saran yang perlu dilakukan sebagai upaya perbaikan pada penelitian ini yaitu untuk rekomendasi penerapan machine learning dapat melakukan pengembangan untuk model regresi atau untuk memprediksi persentase lemak tubuh berdasarkan variabel-variabel seperti berat badan, lingkaran tubuh, usia, dan tinggi badan.

#### DAFTAR PUSTAKA

- Hedgepeth, Royster C. 2016. *An Exploratory Analysis*.
- Powell-Wiley, Tiffany M, Paul Poirier, Lora E Burke, Jean-Pierre Després, Penny Gordon-Larsen, Carl J Lavie, Scott A Lear, et al. 2021. "Obesity and Cardiovascular Disease: A Scientific Statement from the American Heart Association." *Circulation* 143 (21): e984--e1010.
- Rahmany, Mahathir, Abdullah Mohd Zin, and Elankovan A Sundararajan. 2020. "Comparing Tools Provided By Python And R For Exploratory Data Analysis." *IJISCS (International Journal of Information System and Computer Science)*. <https://api.semanticscholar.org/CorpusID:228895058>.
- Ramadhani, Ramadhani, Ramadhanu Ramadhanu, and Taufik Hidayat. 2024. "Exploratory Data Analysis (EDA) Untuk Mengetahui Distribusi Data Kualitas Susu Sapi." *Jurnal SAINTIKOM (Jurnal Sains Manajemen Informatika Dan Komputer)* 23 (1): 68. <https://doi.org/10.53513/jis.v23i1.9500>.
- Sagala, Noviyanti Tri Mareta, and Fonggi Yudi Aryatama. 2022. "Exploratory Data Analysis (EDA): A Study of Olympic Medallist Analisis Data Eksplorasi (EDA): Studi Peraih Medali Olimpiade." In . <https://api.semanticscholar.org/CorpusID:252697013>.

- Salim, Baiq Rissa Khaerawati, Desak Made Wihandani, and Ni Nyoman Ayu Dewi. 2021. "Obesitas Sebagai Faktor Risiko Terjadinya Peningkatan Kadar Trigliserida Dalam Darah: Tinjauan Pustaka." *Intisari Sains Medis*.  
<https://api.semanticscholar.org/CorpusID:246883041>.
- Siri, William E. 1956. "The Gross Composition of the Body1 1The Author's Investigations Were Supported by a Contract between the AEC and the University of California, Berkeley." In , edited by JOHN H LAWRENCE and CORNELIUS A TOBIAS, 4:239–80. *Advances in Biological and Medical Physics*. Elsevier.  
<https://doi.org/https://doi.org/10.1016/B978-1-4832-3110-5.50011-X>.
- Valis, D., L. Zak, and Z. VINTR. 2019. "Perspective Exploratory Methods for Multidimensional Data Analysis." *IEEE International Conference on Industrial Engineering and Engineering Management*, 426–30.  
<https://doi.org/10.1109/IEEM44572.2019.8978643>.
- Venables, W N, and Brian D Ripley. 2002. "Exploratory Multivariate Analysis." In , 301–30. Springer, New York, NY.  
[https://doi.org/10.1007/978-0-387-21706-2\\_11](https://doi.org/10.1007/978-0-387-21706-2_11).
- Widiastuti, Ida Ayu Eka, Seto Priyambodo, and Gede Wira Buanayuda. 2020. "Korelasi Antara Persentase Lemak Tubuh Dan Rasio Lingkar Pinggang-Lingkar Pinggul Dengan Kebugaran Fisik Pada Mahasiswa Fakultas Kedokteran Universitas Mataram." *Intisari Sains Medis* 11: 421.  
<https://api.semanticscholar.org/CorpusID:229112991>.