

OVERWEIGHT



KLASIFIKASI RISIKO OBESITAS DENGAN JENIS KELAMIN, USIA, DAN VARIABEL GAYA HIDUP DAN FAKTOR KELUARGA

To classify obesity levels based on dietary, physical, and demographic attributes.

KELOMPOK 11



CHRISTAMA EZRA YUDIANTO

&



VERY FACHRUROZI

OBESITAS MERUPAKAN MASALAH KESEHATAN GLOBAL YANG DIPENGARUHI OLEH BERBAGAI FAKTOR, TERMASUK POLA MAKAN, AKTIVITAS FISIK, DAN GAYA HIDUP. MEMAHAMI FAKTOR-FAKTOR YANG BERKONTRIBUSI TERHADAP OBESITAS DAPAT MEMBANTU DALAM PENGAMBILAN KEPUTUSAN YANG LEBIH BAIK TERKAIT PENCEGAHAN DAN PENANGANANNYA.




PROBLEMS

4

- Pola makan dan aktivitas fisik berperan besar dalam risiko obesitas. Asupan energi berlebih dan kurangnya aktivitas fisik meningkatkan potensi obesitas. (<https://p2ptm.kemkes.go.id/infographic-p2ptm/obesitas/page/pola-makan-dan-pola-aktivitas-fisik-yang-menyebabkan-obesitas>)
- Deteksi obesitas masih manual dan sering terlambat. Teknologi machine learning dapat memanfaatkan data pola makan dan aktivitas fisik untuk prediksi lebih dini. (https://jurnal.unigal.ac.id/JKP/article/view/16884?utm_source)
- Penelitian menunjukkan pola makan tidak sehat dan kurangnya aktivitas fisik sebagai faktor utama obesitas, terutama di perkotaan. (https://www.mendeley.com/catalogue/cff6835f-f568-3bf8-8f36-ac7f127b852c/?utm_source)

DATASHEET

5

 ARAVINDPCODER · UPDATED A YEAR AGO

▲ 166 <> Code

Obesity or CVD risk (Classify/Regressor/Cluster)

The dataset consists of 17 attributes and 2111 records exploring CVD's

[Data Card](#) [Code \(414\)](#) [Discussion \(6\)](#) [Suggestions \(0\)](#)

About Dataset

The data consist of the estimation of obesity levels in people from the countries of Mexico, Peru and Colombia, with ages between 14 and 61 and diverse eating habits and physical condition, data was collected using a web platform with a survey where anonymous users answered each question, then the information was processed obtaining 17 attributes and 2111 records.

The attributes related with eating habits are: Frequent consumption of high caloric food (FAVC), Frequency of consumption of vegetables (FCVC), Number of main meals (NCP), Consumption of food between meals (CAEC), Consumption of water daily (CH20), and Consumption of alcohol (CALC). The attributes related with the physical condition are: Calories consumption monitoring (SCC), Physical activity frequency (FAF), Time using technology devices (TUE), Transportation used (MTRANS)

Dataset ini mencakup data dari individu di Meksiko, Peru, dan Kolombia (usia 14–61 tahun), yang dikumpulkan melalui survei berbasis web.

APA SAJA ATRIBUTNYA ?

POLA MAKAN

7

- FREKUENSI MAKANAN BERLEMAK
- FREKUENSI MAKANAN DALAM SEHARI
- MEROKOK
- FREKUENSI MAKANAN CAMILAN
- KONSUMSI MINUMAN AIR PUTIH SEHARI
- PENYAKIT TERKAIT KONSUMSI MINUMAN KALORI
- KONSUMSI ALKOHOL SEHARI

INTERNAL

- GENDER
- BERAT BADAN
- TINGGI BADAN
- UMUR

LAIN - LAIN

- MODAL TRANSPORTASI
- FAMILY OBESITAS
- AKTIVITAS FISIK

SIAPA TAGETNYA?

8

Membantu dokter dan tenaga medis

Memberikan wawasan personal

GOALS

Meningkatkan Akurasi dan Efisiensi Diagnosis Obesitas/CVD

Memungkinkan Deteksi Dini dan Intervensi Proaktif

Menyediakan Rekomendasi Personalisasi Berbasis Data

EXPLORATORY DATA ANALYSIS

KONDISI DATASHEET

10

Tidak terdapat nilai
yang null

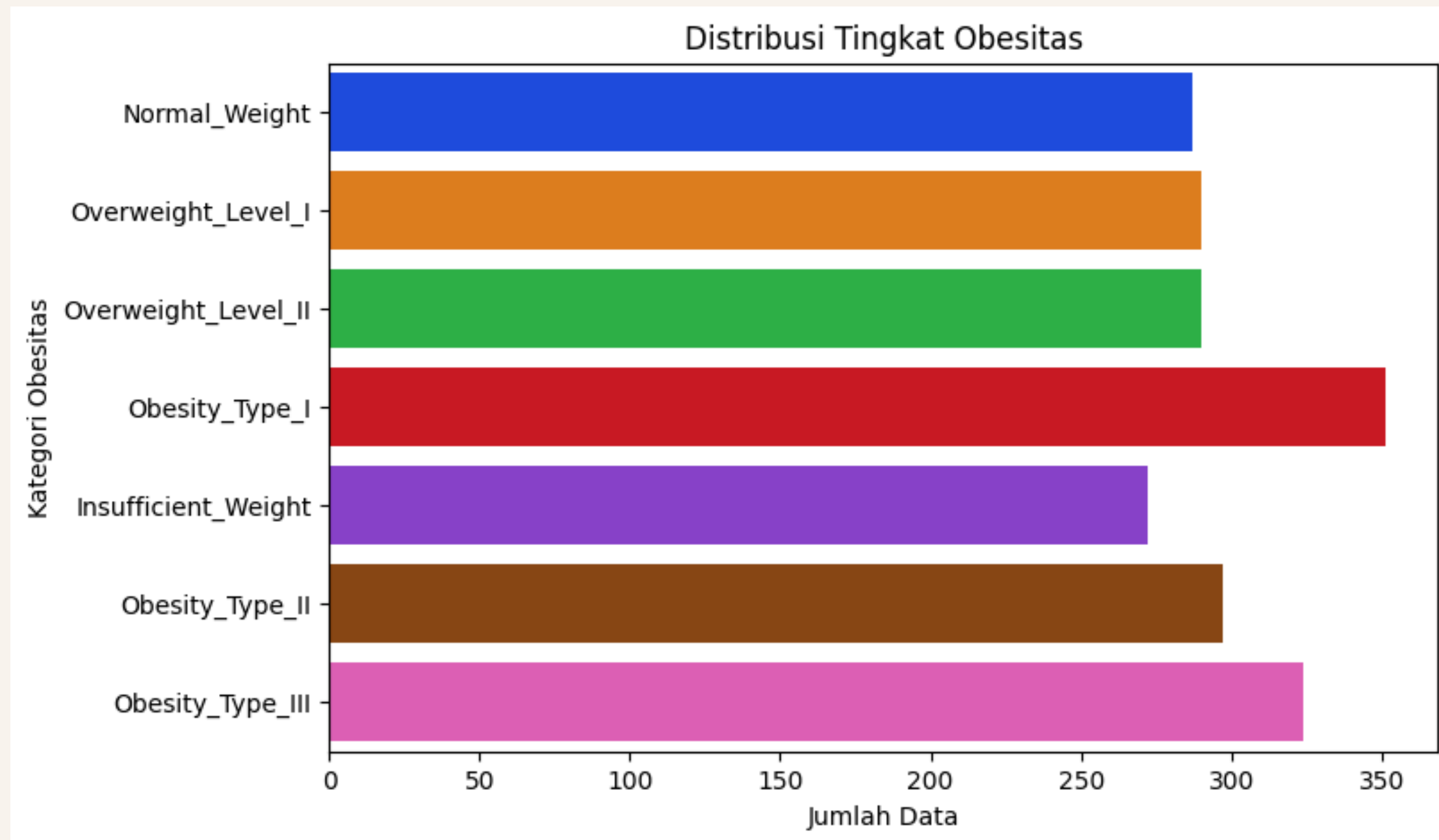
Tidak terdapat
terdapat data yang
anomali

Terdapat Feature
yang Imbalance

Feature terbagi menjadi 2
kategorikal dan
numerikal

APA YANG DIANALISIS ?

DISTRIBUS TARGET OBESITAS

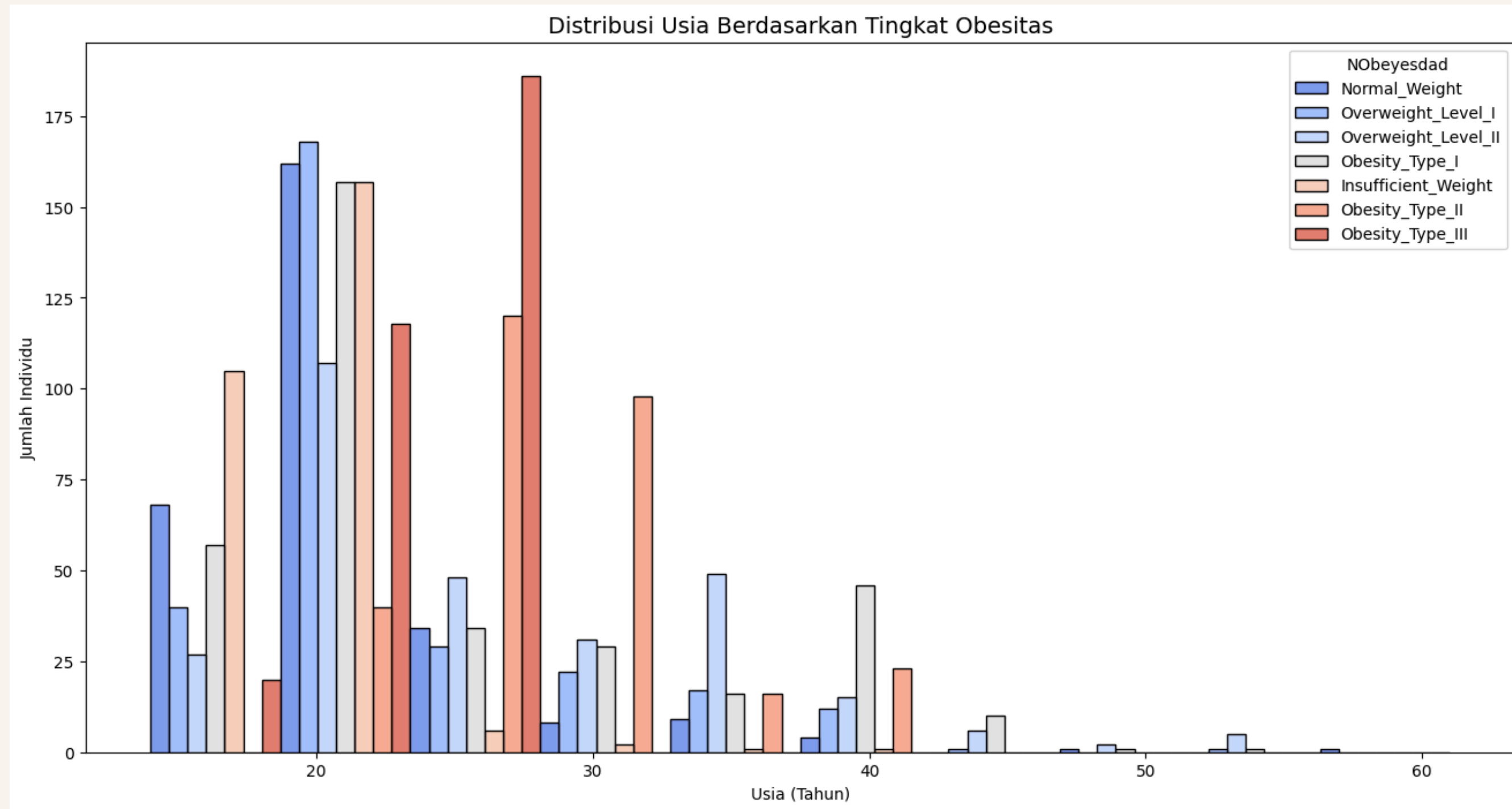


TOP 3

1. Obesity_Type_I
2. Obesity_Type_III
3. Obesity_Type_II

DISTRIBUSI USIA BERDASARKAN TINGKAT OBESITAS

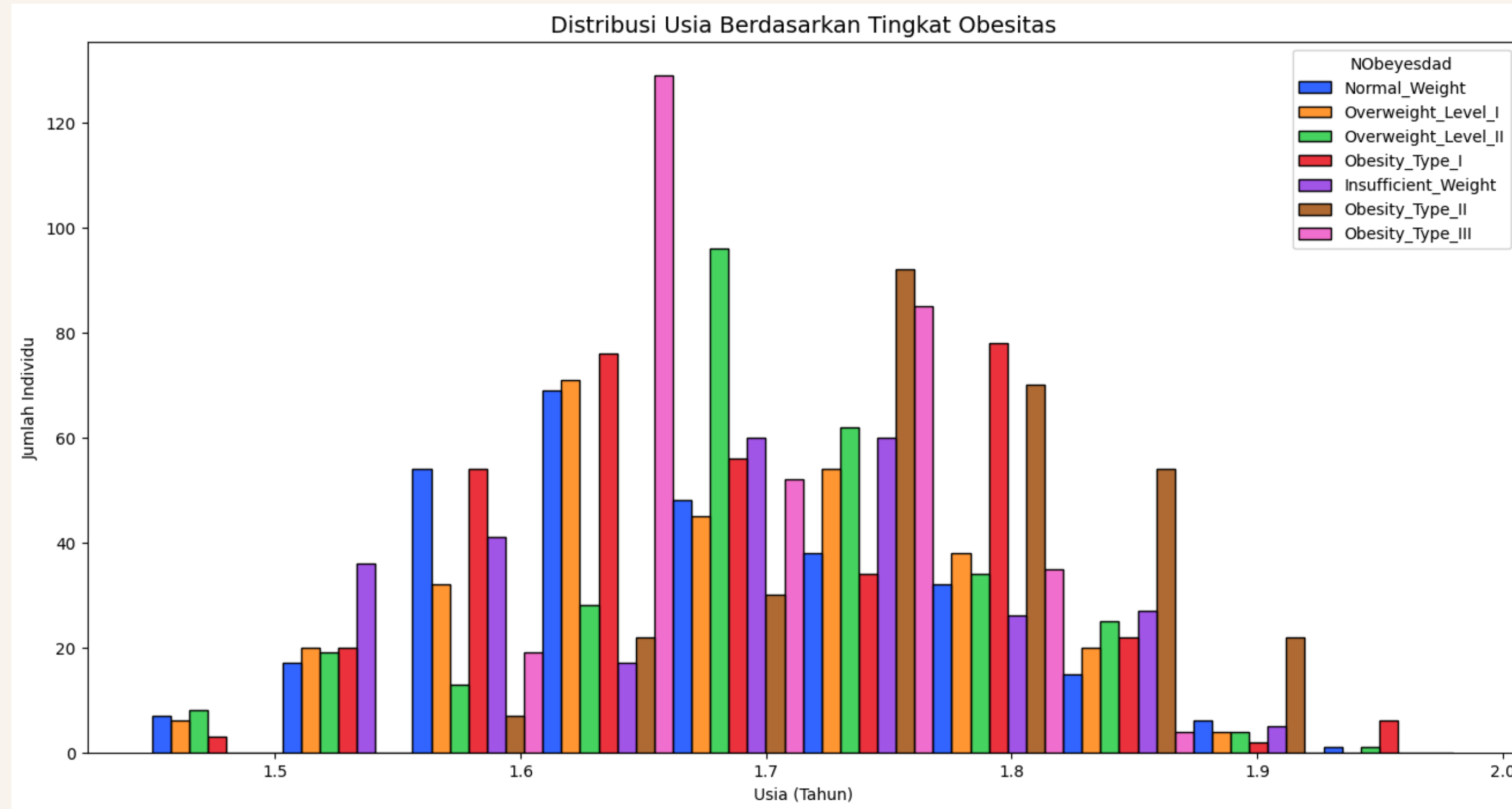
12



distribusi obesitas terbesar Obesity_Type_II pada usia **25 – 30 tahun**

DISTRIBUSI TINGGI BERDASARKAN TINGKAT OBESITAS

13



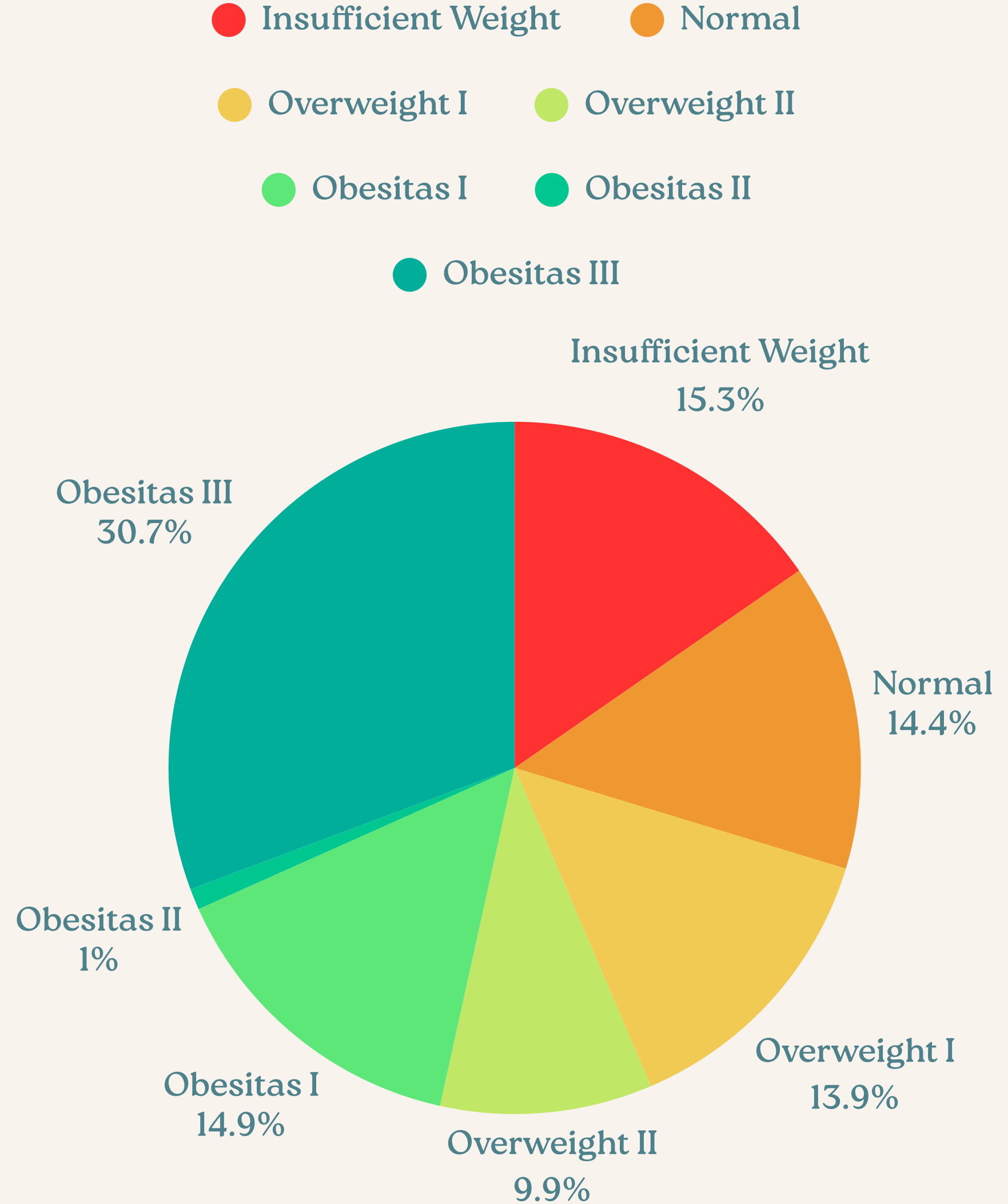
Distribusi obesitas terbesar Obesity_Type_III ada pada tinggi

1.63 - 1.67 m

DISTRIBUSI GENDER FEMALE TERHADAP TINGKAT OBESITAS

14

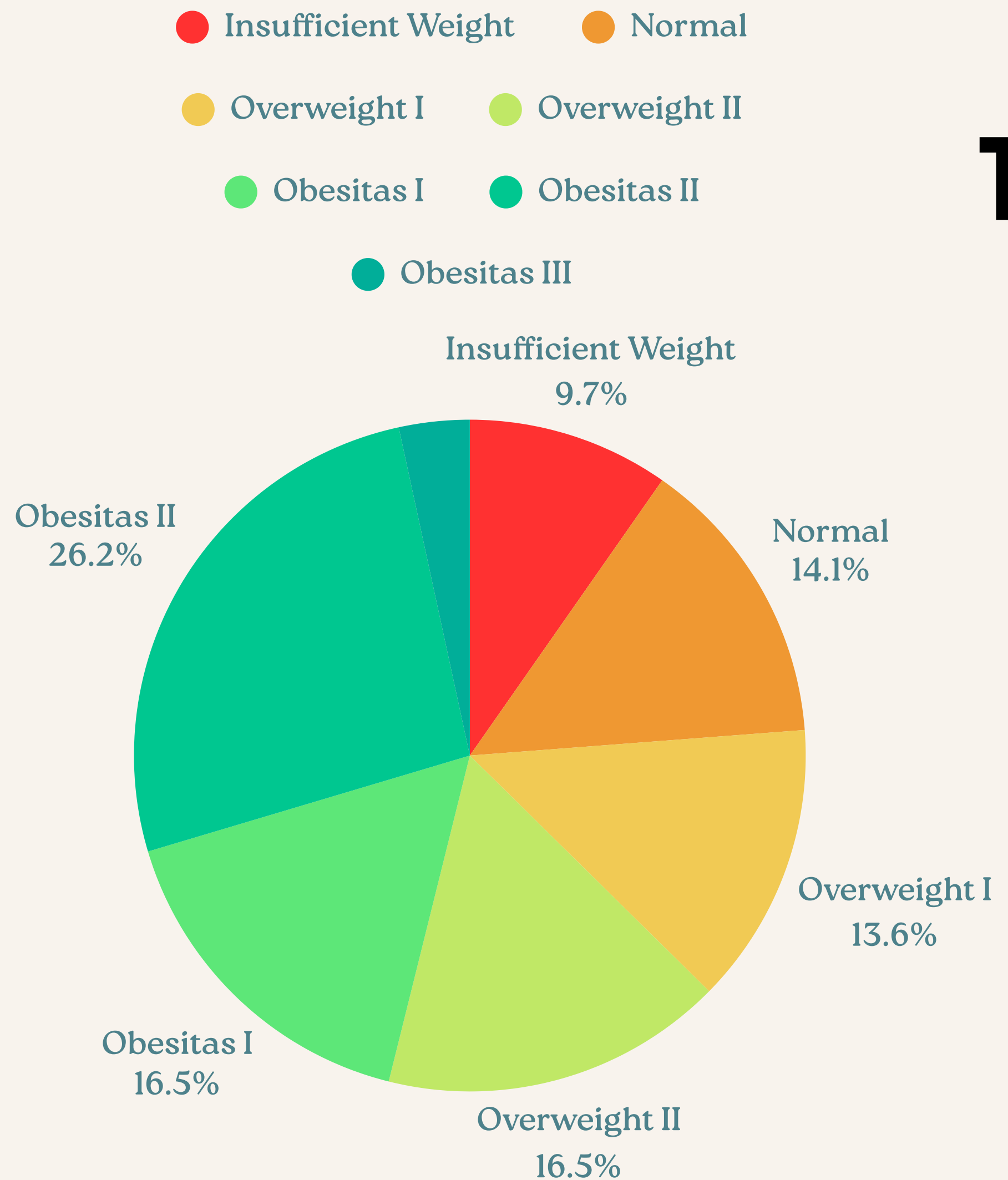
Tingkat obesitas paling tinggi pada **WANITA** yaitu berada pada tingkat obesitas **tipe 3**



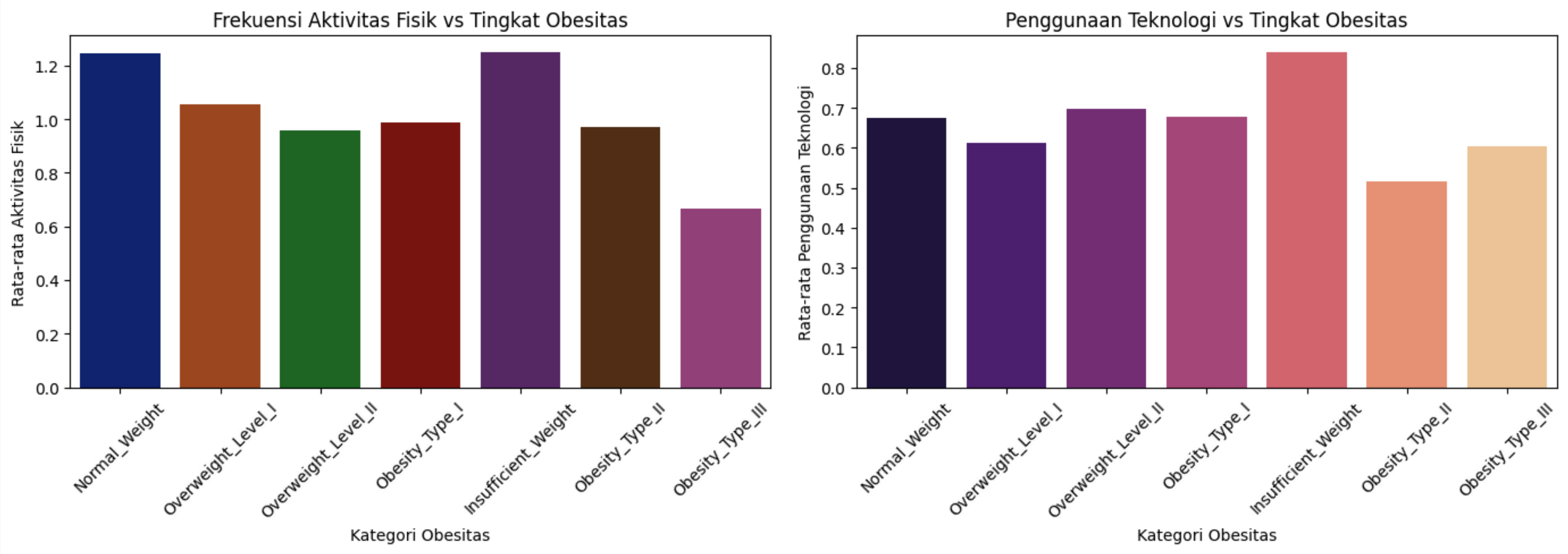
DISTRIBUSI GENDER MALE TERHADAP TINGKAT OBESITAS

Tingkat obesitas
paling tinggi pada
PRIA yaitu berada
tingkat obesitas tipe

2

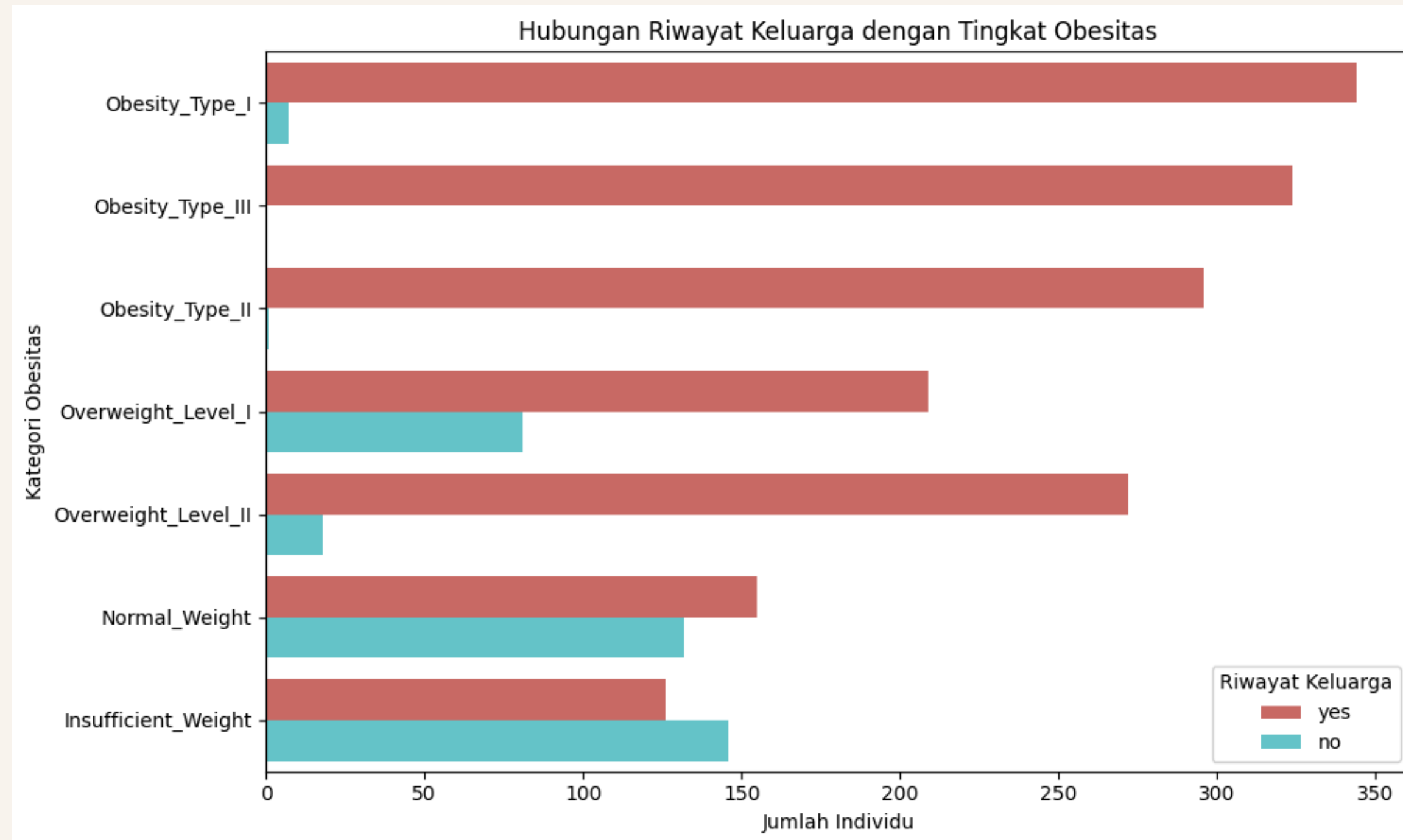


DISTRIBUSI AKTIVITAS FISIK TERHADAP TINGKAT OBESITAS16

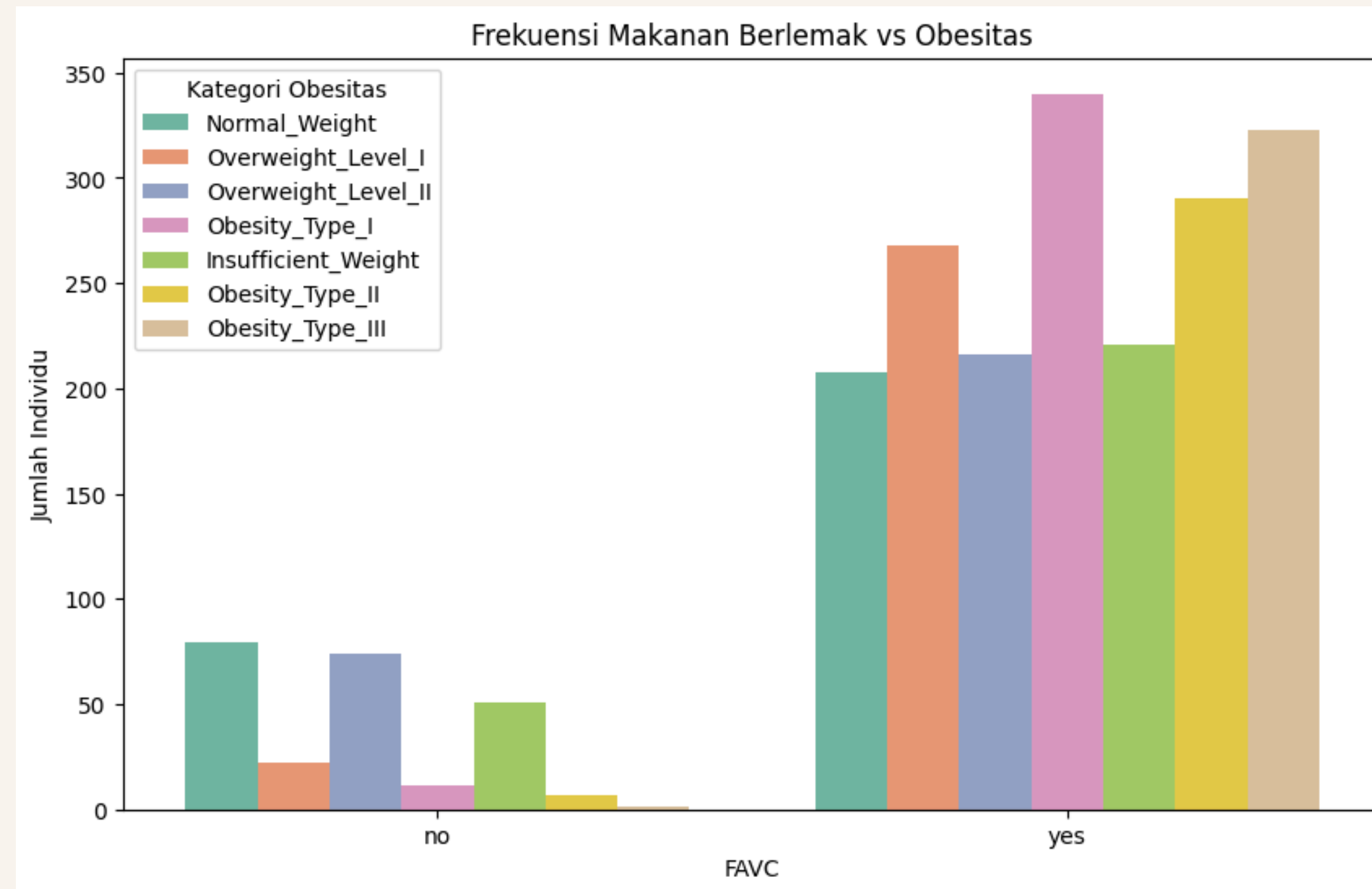


DISTRIBUSI HUBUNGAN KELUARGA BERDASARKAN TINGKAT OBESITAS

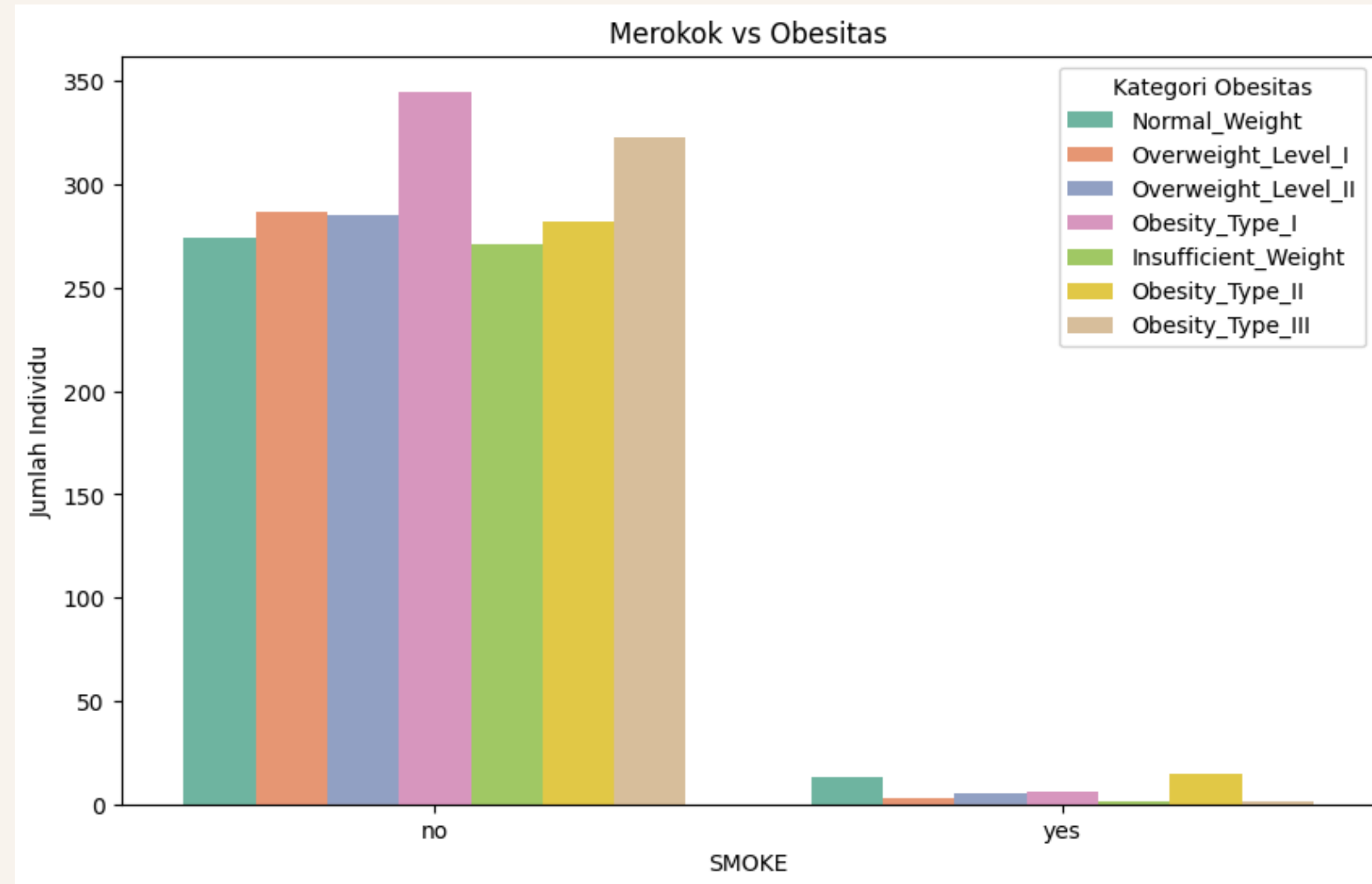
17



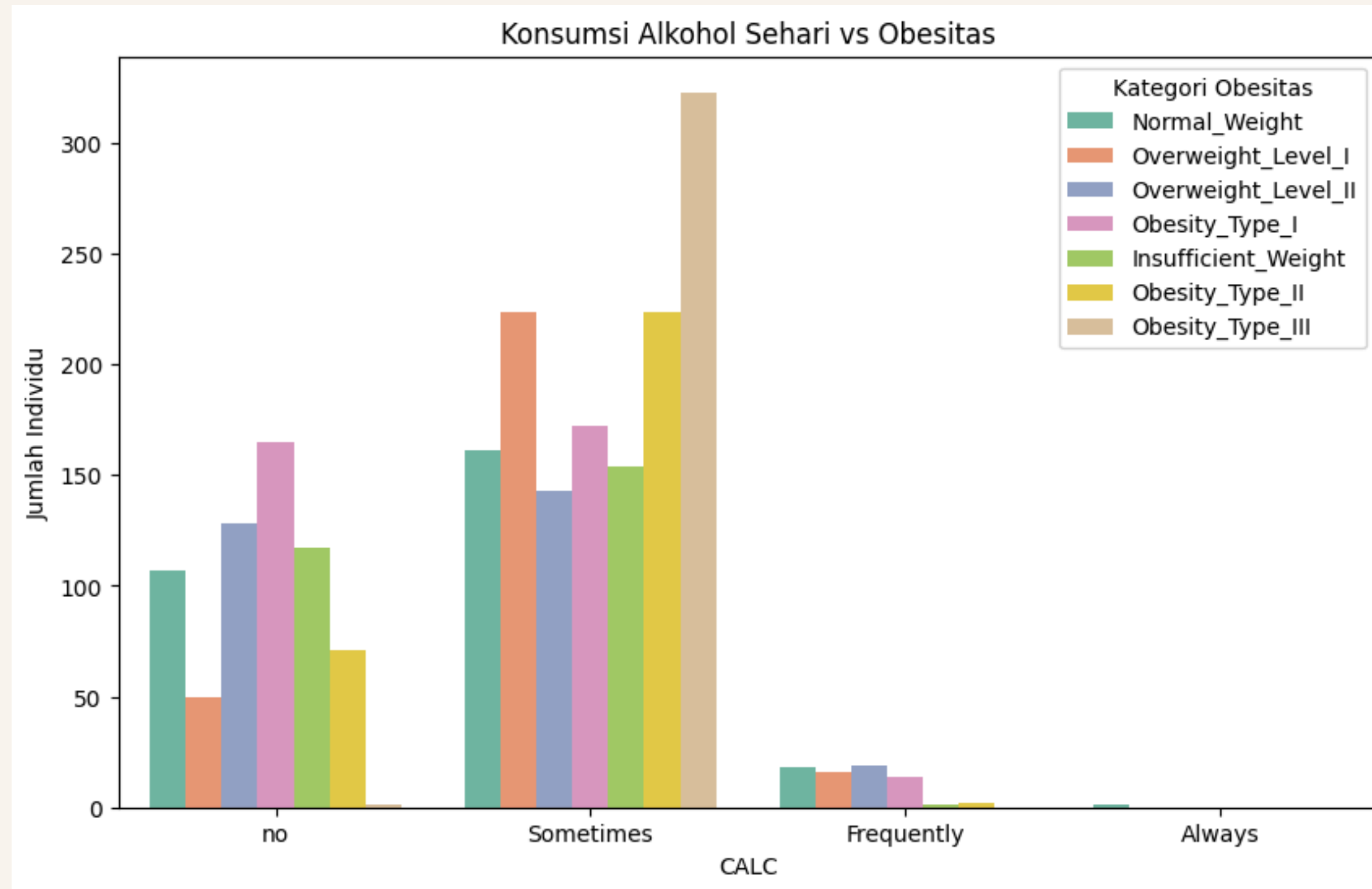
MAYORITAS INDIVIDU DENGAN RIWAYAT KELUARGA OBESITAS CENDERUNG MEMILIKI TINGKAT OBESITAS YANG LEBIH TINGGI.



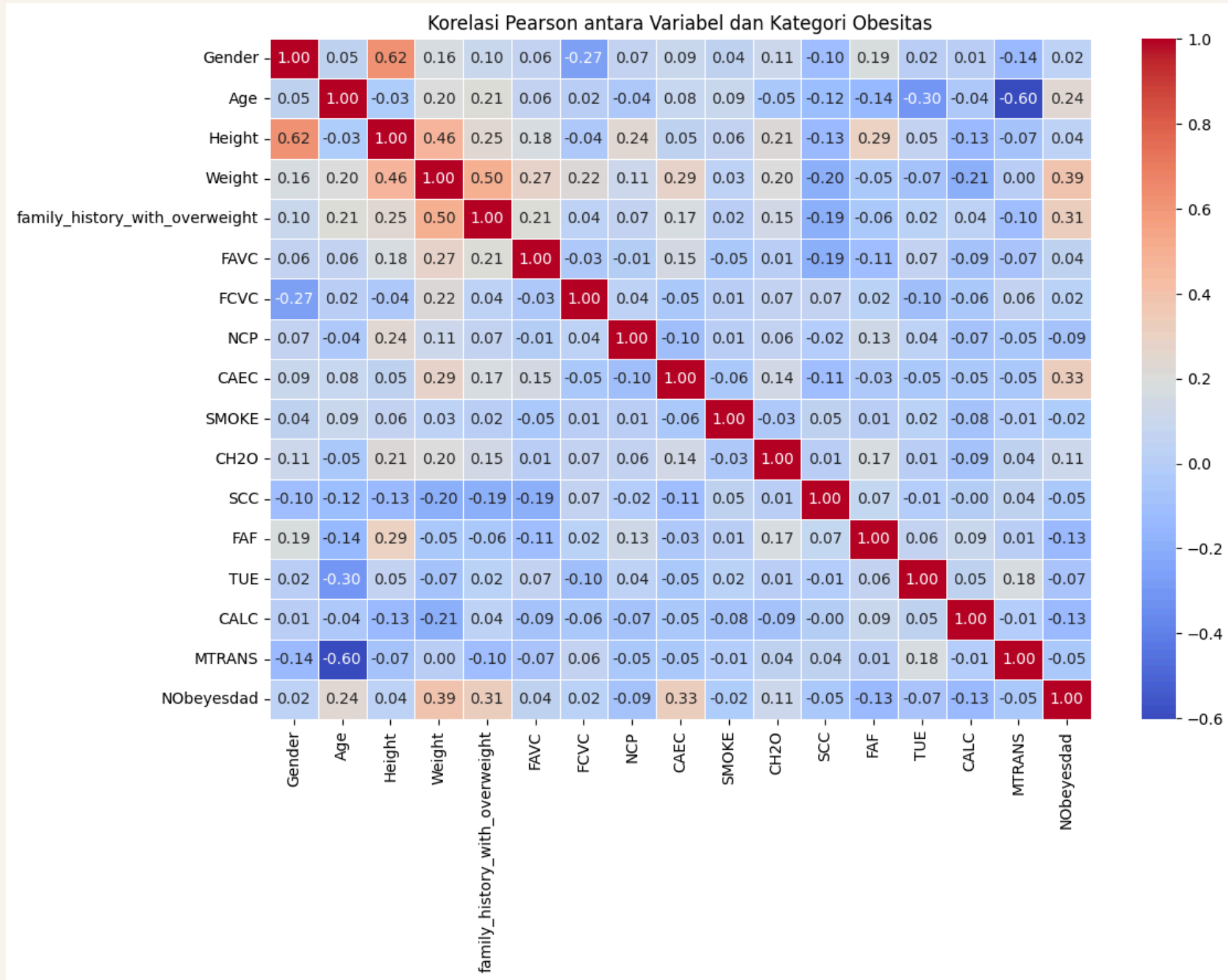
LEBIH BANYAK INDIVIDU DENGAN OBESITAS DITEMUKAN PADA KELOMPOK YANG MENGONSUMSI MAKANAN BERLEMAK.



MAYORITAS INDIVIDU, BAIK DENGAN BERAT BADAN NORMAL MAUPUN OBESITAS, ADALAH NON-PEROKOK



MAYORITAS INDIVIDU MENGONSUMSI ALKOHOL SESEKALI, TERUTAMA PADA KATEGORI OBESITAS TIPE 3.



**BERAT BADAN ($R = 0.39$),
POLA MAKAN CAMILAN ($R = 0.33$), DAN RIWAYAT
KELUARGA OBESITAS ($R = 0.31$) LEBIH RELEVAN,
SEDANGKAN MEROKOK ($R \approx 0$) DAN KONSUMSI ALKOHOL
($R \approx 0$) BERPENGARUH
KECIL.**

CATEGORIKAL FEATURE DENGAN TINGKAT OBESITAS

22

Feature	Chi-Square Score
Gender	324.978359
SCC	117.429254
family_history_with_overweight	113.435378
MTRANS	102.780885
SMOKE	31.467977
FAVC	27.081298
CALC	21.819606

TOP 3

- 1. GENDER (SKOR: 324.978359)
- 2. SCC (SKOR: 117.429254)
- 3. FAMILY_HISTORY_WITH_OVERWEIGHT (SKOR: 113.435378)

NUMERIKAL FEATURE DENGAN TINGKAT OBESITAS

23

Feature	ANOVA Score
Weight	1966.518018
FCVC	112.315462
Age	77.954154
Height	38.432313
NCP	26.811662
FAF	17.484200
CH20	16.171142
TUE	7.876656

TOP 3

1. WEIGHT (SKOR: 1966.518018)
2. FCVC (SKOR: 112.315462)
3. AGE (SKOR: 77.954154)

DATA PREPROCESSING

DATA TRANSFORMATION - KATEGORIKAL

APA YANG DIPAKAI ?

- **Sifat Data yang Berurutan (Ordinal Data)**
- **Efisiensi**
- **Cocok untuk Model yang Mendukung Data Berurutan**

DATA TRANSFORMATION - KATEGORIKAL

25

Feature	Kategori Asli	Encoded Value
family_history_with_overweight	yes / no	1 / 0
FAVC	yes / no	1 / 0
SMOKE	yes / no	1 / 0
SCC	yes / no	1 / 0
CALC	Never, Sometimes, Frequently	0, 1, 2
MTRANS	Public_Transportation, Automobile, Bike, Walking	0, 1, 2, 3

DATA TRANSFORMATION - NUMERIKAL

APA YANG DIPAKAI ?

Standardization

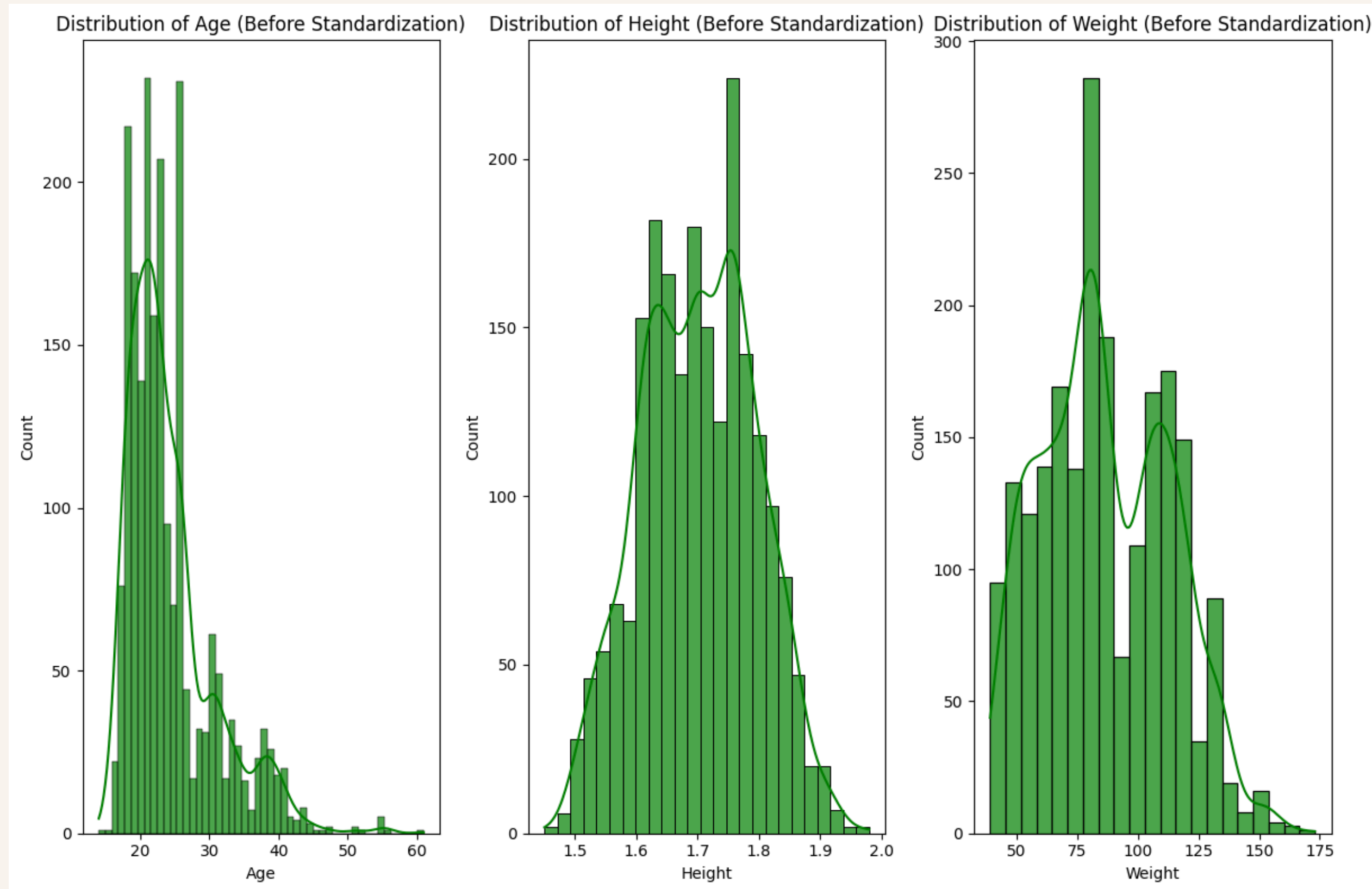
- **Age**
- **Height**
- **Weight**

Normalization

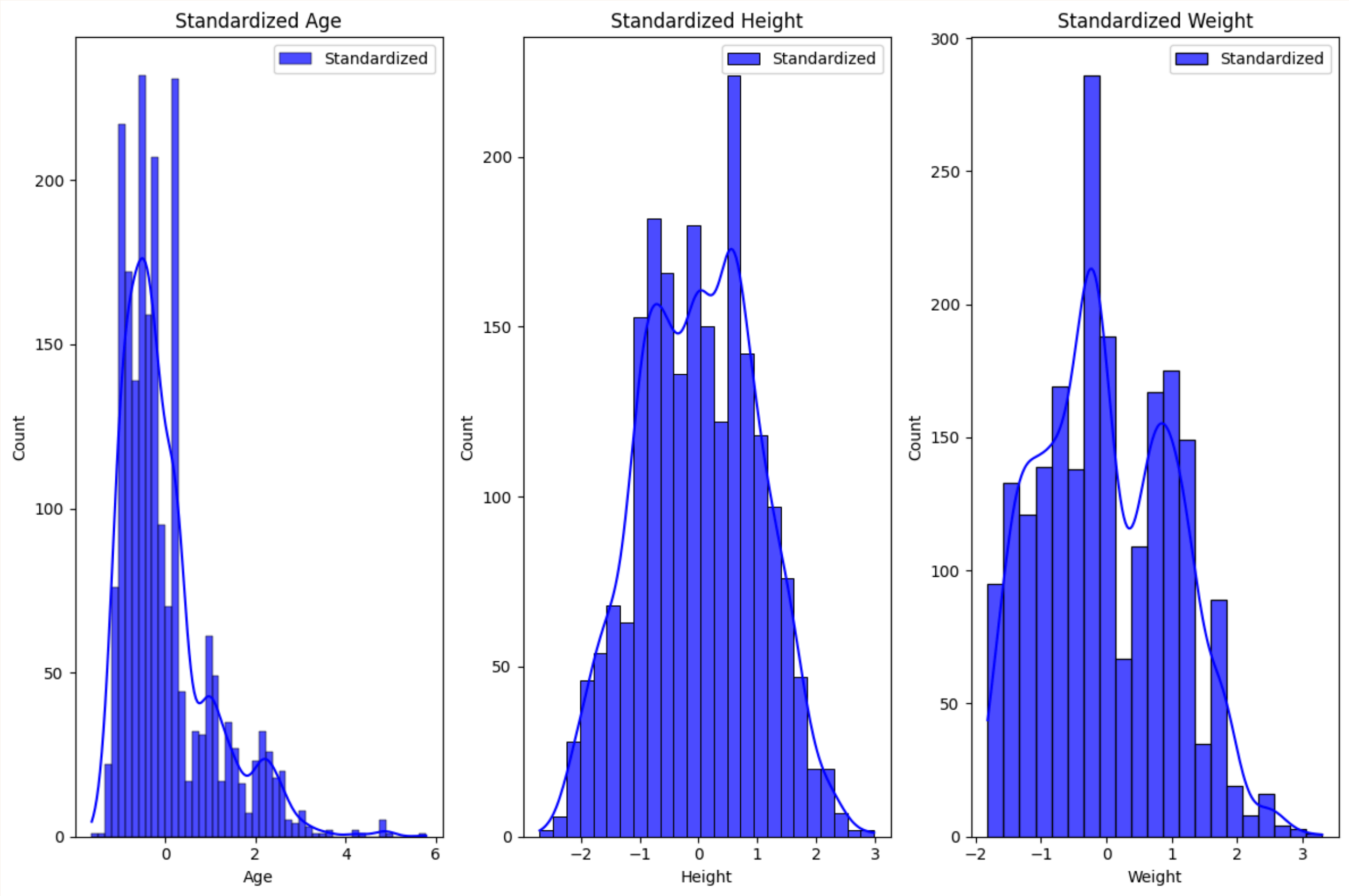
- **FCVC**
- **NCP**
- **CH2O**
- **FAF**
- **TUE**

DISTRIBUSI SEBELUM FEATURE SCALING- STANDARDIZATION

27

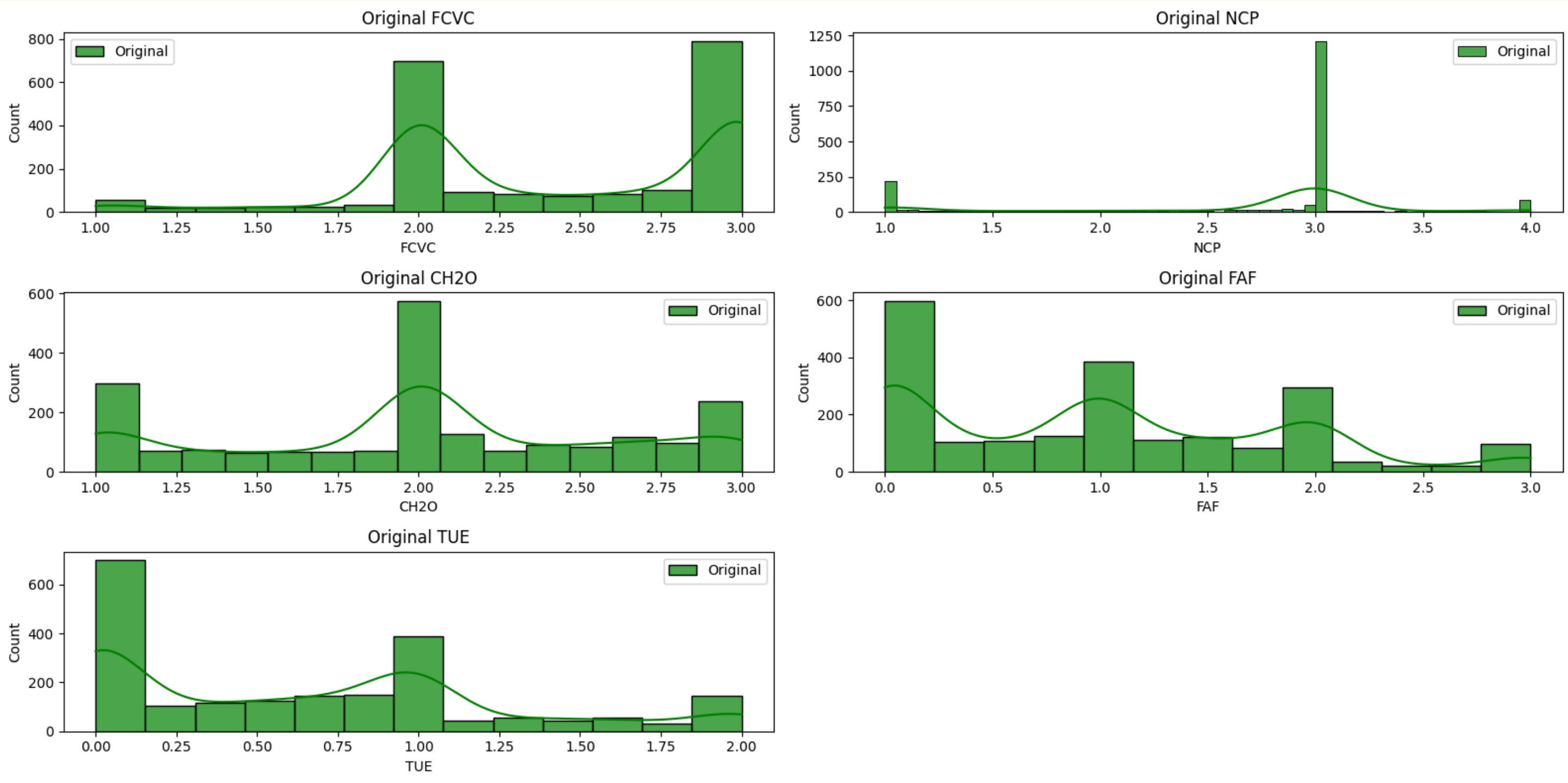


DISTRIBUSI SESUDAH FEATURE SCALING- STANDARDIZATION



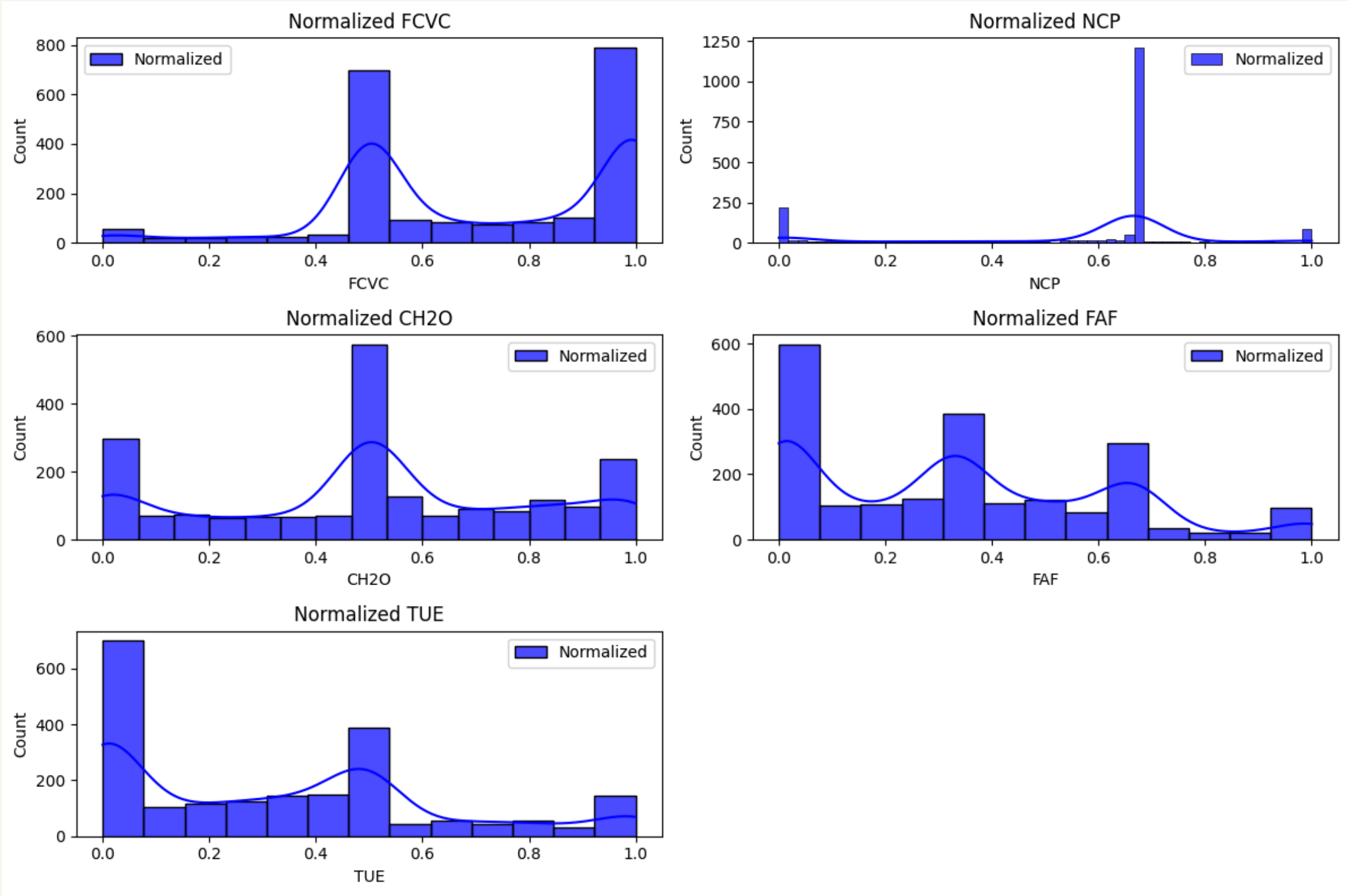
DISTRIBUSI SEBELUM FEATURE SCALING- NORMALIZATION

29



DISTRIBUSI SESUDAH FEATURE SCALING- NORMALIZATION

30



FEATURE SELECTION

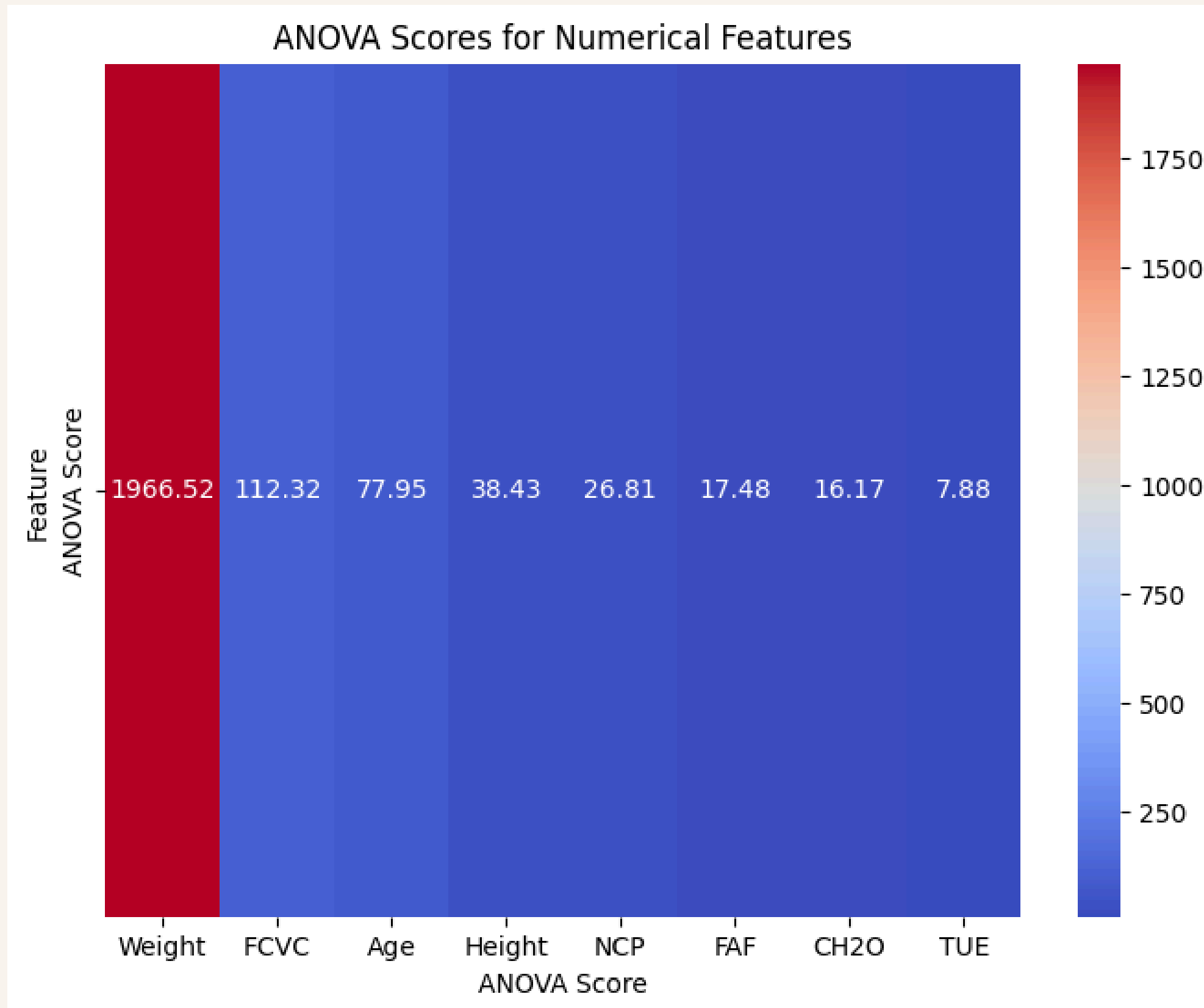
NUMERIKAL

ANOVA

31

- Mengidentifikasi fitur penting yang memiliki pengaruh signifikan terhadap target.
- Meningkatkan akurasi model dengan hanya memilih fitur yang relevan.
- Mengurangi dimensi data, sehingga model lebih efisien.
- Mengurangi overfitting dengan membuang fitur yang tidak berkontribusi.
- Mempercepat proses pelatihan dengan dataset yang lebih ringkas.

32

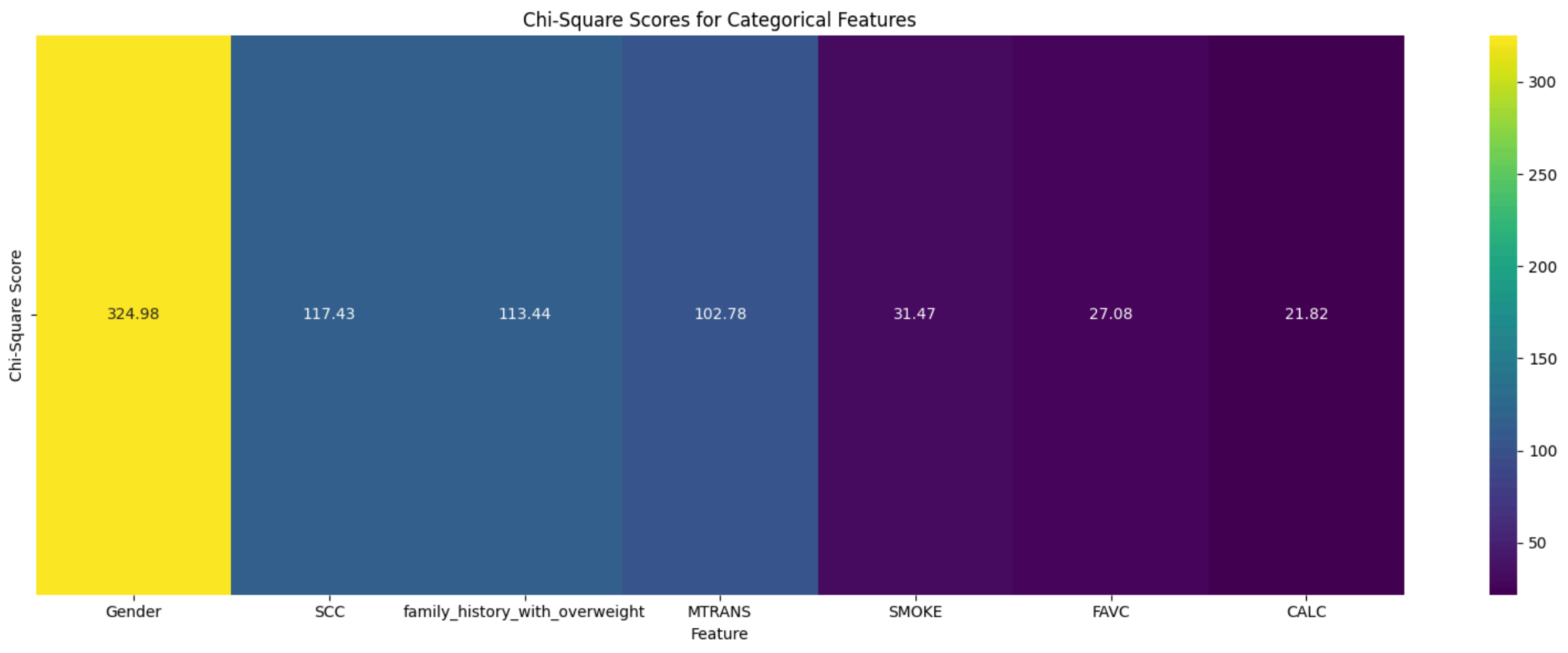


DROP FEATURE:

- TUE
- CH20
- FAF

KATEGORIKAL

- **Goodness of Fit:** Menguji kecocokan data dengan distribusi yang diharapkan.
- **Independence:** Menguji hubungan antar variabel kategorikal.
- **Hipotesis Nol:** Menguji apakah ada perbedaan atau hubungan.
- **Signifikansi:** Menilai perbedaan signifikan antara data yang diamati dan diharapkan.



DROP FEATURE:

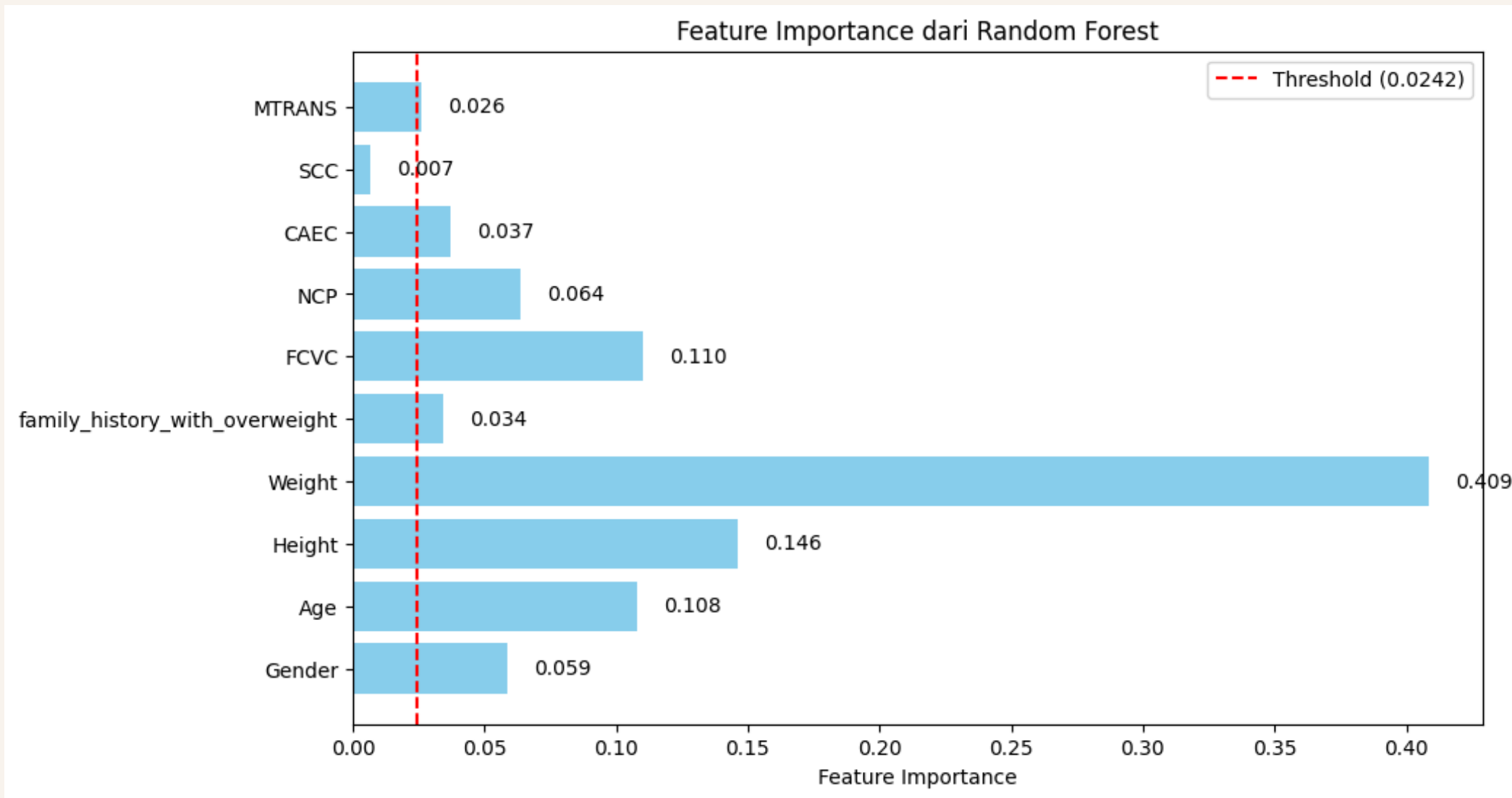
- SMOKE
- FAVC
- CALC

FEATURE IMPORTANCE DARI RANDOM FOREST

Feature Importance dari Random Forest

35

- **Seleksi Fitur:** Identifikasi fitur paling relevan untuk meningkatkan performa model.
- **Interpretasi Model:** Memahami faktor yang berpengaruh pada prediksi.
- **Efisiensi Model:** Mengurangi fitur tak penting untuk mempercepat pelatihan.
- **Deteksi Redundansi:** Menyoroti fitur dengan kontribusi serupa.



Bentuk dataset setelah penghapusan fitur:

X_train: (1688, 9)

X_test: (423, 9)

Akurasi setelah penghapusan fitur: 0.9574

Bentuk dataset sebelum penghapusan fitur:

X_train: (1688, 10)

X_test: (423, 10)

Total skor Feature Importance: 1.0000

Threshold berdasarkan persentil ke-10: 0.0242

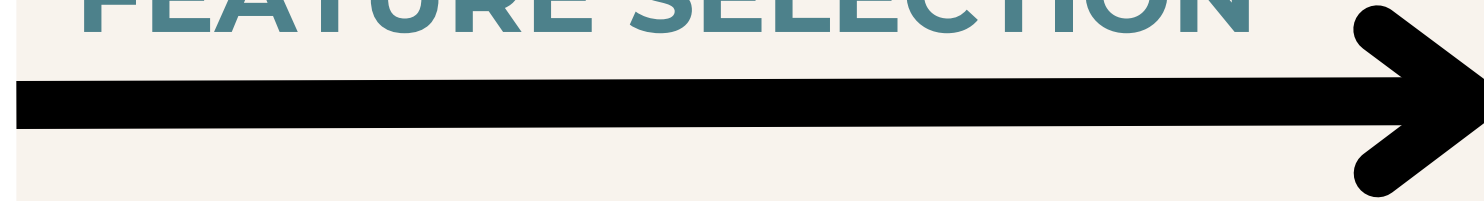
Fitur dengan skor rendah yang bisa dihapus (di bawah threshold 0.0242):

SCC: 0.0065

16 FEATURES

FEATURE SELECTION

9 FEATURES



MODELING

Random Forest

Mampu menangani data dengan fitur yang kompleks, baik numerik maupun kategori. Model ini juga lebih robust terhadap overfitting karena menggunakan banyak pohon keputusan, serta dapat mengidentifikasi fitur yang paling berpengaruh dalam klasifikasi obesitas

KNN

- Cocok untuk data dengan pola yang jelas dan mudah diperbarui tanpa perlu retraining.
- Algoritma yang sederhana, mudah dipahami, dan tidak memerlukan training eksplisit.

SVM

Efektik dalam dimensi tinggi:
SVM efektif dalam ruang fitur yang memiliki banyak dimensi, yang umumnya terjadi dalam dataset komplek

XGBOOSTS

- Mampu menangani data yang tidak seimbang dengan baik karena mampu menyesuaikan bobot kelas secara otomatis.
- Memiliki regularisasi yang kuat untuk mengurangi overfitting.
- Dapat menangani berbagai jenis fitur dan skala data.

F1-Score

Menyeimbangkan Precision dan Recall → F1-score adalah rata-rata harmonik antara precision dan recall, sehingga cocok untuk dataset yang mungkin memiliki ketidakseimbangan antar kelas.

Digunakan dalam Klasifikasi Multi-Kelas → Dalam masalah klasifikasi seperti ini (klasifikasi obesitas), F1-score membantu mengevaluasi model dengan lebih baik dibandingkan hanya menggunakan akurasi.

MATRIK EVALUASI YANG DIGUNAKAN

40

F1-Score dihitung menggunakan rumus berikut

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Presisi adalah jumlah prediksi positif yang benar dibagi dengan total prediksi positif yang dilakukan oleh model.

Recall (atau sensitivity) adalah jumlah prediksi positif yang benar dibagi dengan total instance kelas positif yang sebenarnya dalam dataset.

DEFAULT MODELING

Default KNN

41

```
KNN F1 Score: 0.8556361350565602
KNN Classification Report:
              precision    recall  f1-score   support

         0       0.79      0.94      0.86         86
         1       0.75      0.48      0.59         93
         2       0.88      0.95      0.92        102
         3       0.98      0.95      0.97         88
         4       0.96      1.00      0.98         98
         5       0.77      0.85      0.81         88
         6       0.89      0.85      0.87         79

 accuracy          0.86
 macro avg         0.86
weighted avg         0.86

KNN Confusion Matrix:
[[81  5  0  0  0  0  0]
 [21 45  2  0  0 20  5]
 [ 0  0 97  2  1  1  1]
 [ 0  0  1 84  3  0  0]
 [ 0  0  0  0 98  0  0]
 [ 1  8  2  0  0 75  2]
 [ 0  2  8  0  0  2 67]]
```

DEFAULT MODELING

Default SVM

42

SVM F1 Score: 0.5547439930901169					
SVM Classification Report:					
	precision	recall	f1-score	support	
0	0.69	0.83	0.75	86	
1	0.44	0.34	0.39	93	
2	0.58	0.33	0.42	102	
3	0.75	0.47	0.57	88	
4	0.63	1.00	0.77	98	
5	0.50	0.51	0.51	88	
6	0.42	0.53	0.47	79	
accuracy			0.57	634	
macro avg	0.57	0.57	0.55	634	
weighted avg	0.57	0.57	0.55	634	
SVM Confusion Matrix:					
[[71 15 0 0 0 0 0]					
[29 32 0 0 0 27 5]					
[0 0 34 14 19 2 33]					
[0 0 9 41 38 0 0]					
[0 0 0 0 98 0 0]					
[3 19 2 0 0 45 19]					
[0 7 14 0 0 16 42]]					

DEFAULT MODELING

Default Random Forest

43

```
Random Forest F1 Score: 0.9436939991173161
Random Forest Classification Report:
              precision    recall  f1-score   support

         0              0.99      0.95      0.97         86
         1              0.83      0.91      0.87         93
         2              0.98      0.96      0.97        102
         3              0.97      0.99      0.98         88
         4              1.00      0.99      0.99         98
         5              0.89      0.85      0.87         88
         6              0.96      0.94      0.95         79

 accuracy              0.94         634
  macro avg              0.94      0.94      0.94         634
weighted avg              0.95      0.94      0.94         634

Random Forest Confusion Matrix:
[[82  4  0  0  0  0  0]
 [ 1 85  0  0  0  6  1]
 [ 0  1 98  3  0  0  0]
 [ 0  0  1 87  0  0  0]
 [ 0  0  1  0 97  0  0]
 [ 0 11  0  0  0 75  2]
 [ 0  2  0  0  0  3 74]]
```


DEFAULT MODELING

Default XGBoost

XGBoost F1 Score: 0.9512437851762104					
XGBoost Classification Report:					
	precision	recall	f1-score	support	
0	0.94	0.97	0.95	86	
1	0.91	0.85	0.88	93	
2	0.99	0.95	0.97	102	
3	0.97	0.99	0.98	88	
4	1.00	0.99	0.99	98	
5	0.86	0.94	0.90	88	
6	1.00	0.97	0.99	79	
accuracy			0.95	634	
macro avg	0.95	0.95	0.95	634	
weighted avg	0.95	0.95	0.95	634	
XGBoost Confusion Matrix:					
[[83 2 0 0 0 1 0]					
[5 79 0 0 0 9 0]					
[0 0 97 2 0 3 0]					
[0 0 1 87 0 0 0]					
[0 0 0 1 97 0 0]					
[0 5 0 0 0 83 0]					
[0 1 0 0 0 1 77]]					

MODELING 2

46

Training Random Forest...

Accuracy for Random Forest: 0.9480

	precision	recall	f1-score	support
0	0.98	0.96	0.97	56
1	0.89	0.90	0.90	62
2	0.97	0.95	0.96	78
3	0.98	0.97	0.97	58
4	1.00	1.00	1.00	63
5	0.88	0.88	0.88	56
6	0.92	0.98	0.95	50
accuracy			0.95	423
macro avg	0.95	0.95	0.95	423
weighted avg	0.95	0.95	0.95	423

Training Support Vector Machine...

Accuracy for Support Vector Machine: 0.9125

	precision	recall	f1-score	support
0	0.89	0.98	0.93	56
1	0.92	0.76	0.83	62
2	0.99	0.91	0.95	78
3	0.92	0.98	0.95	58
4	1.00	1.00	1.00	63
5	0.82	0.84	0.83	56
6	0.82	0.92	0.87	50
accuracy			0.91	423
macro avg	0.91	0.91	0.91	423
weighted avg	0.92	0.91	0.91	423

MODELING 2

47

Training K-Nearest Neighbors...					
Accuracy for K-Nearest Neighbors: 0.8913					
	precision	recall	f1-score	support	
0	0.84	0.96	0.90	56	
1	0.79	0.68	0.73	62	
2	0.95	0.94	0.94	78	
3	0.97	0.98	0.97	58	
4	1.00	1.00	1.00	63	
5	0.84	0.77	0.80	56	
6	0.80	0.90	0.85	50	
accuracy			0.89	423	
macro avg	0.89	0.89	0.89	423	
weighted avg	0.89	0.89	0.89	423	

Accuracy for XGBoost: 0.9740					
	precision	recall	f1-score	support	
0	0.93	1.00	0.97	56	
1	0.98	0.89	0.93	62	
2	0.99	0.97	0.98	78	
3	0.98	0.98	0.98	58	
4	1.00	1.00	1.00	63	
5	0.95	0.98	0.96	56	
6	0.98	1.00	0.99	50	
accuracy			0.97	423	
macro avg	0.97	0.98	0.97	423	
weighted avg	0.97	0.97	0.97	423	

Presentase Prediksi	KNN	Random Forest	SVM	XGBoost
Normal_Weight	0.90	0.97	0.93	0.97
Overweight_Level_I	0.73	0.90	0.83	0.93
Overweight_Level_II	0.94	0.96	0.95	0.98
Obesity_Type_I	0.97	0.97	0.95	0.98
Insufficient_Weight	1.00	1.00	1.00	1.00
Obesity_Type_II	0.80	0.88	0.83	0.96
Obesity_Type_III	0.85	0.95	0.87	0.99

49

BEFORE MODEL TUNNING

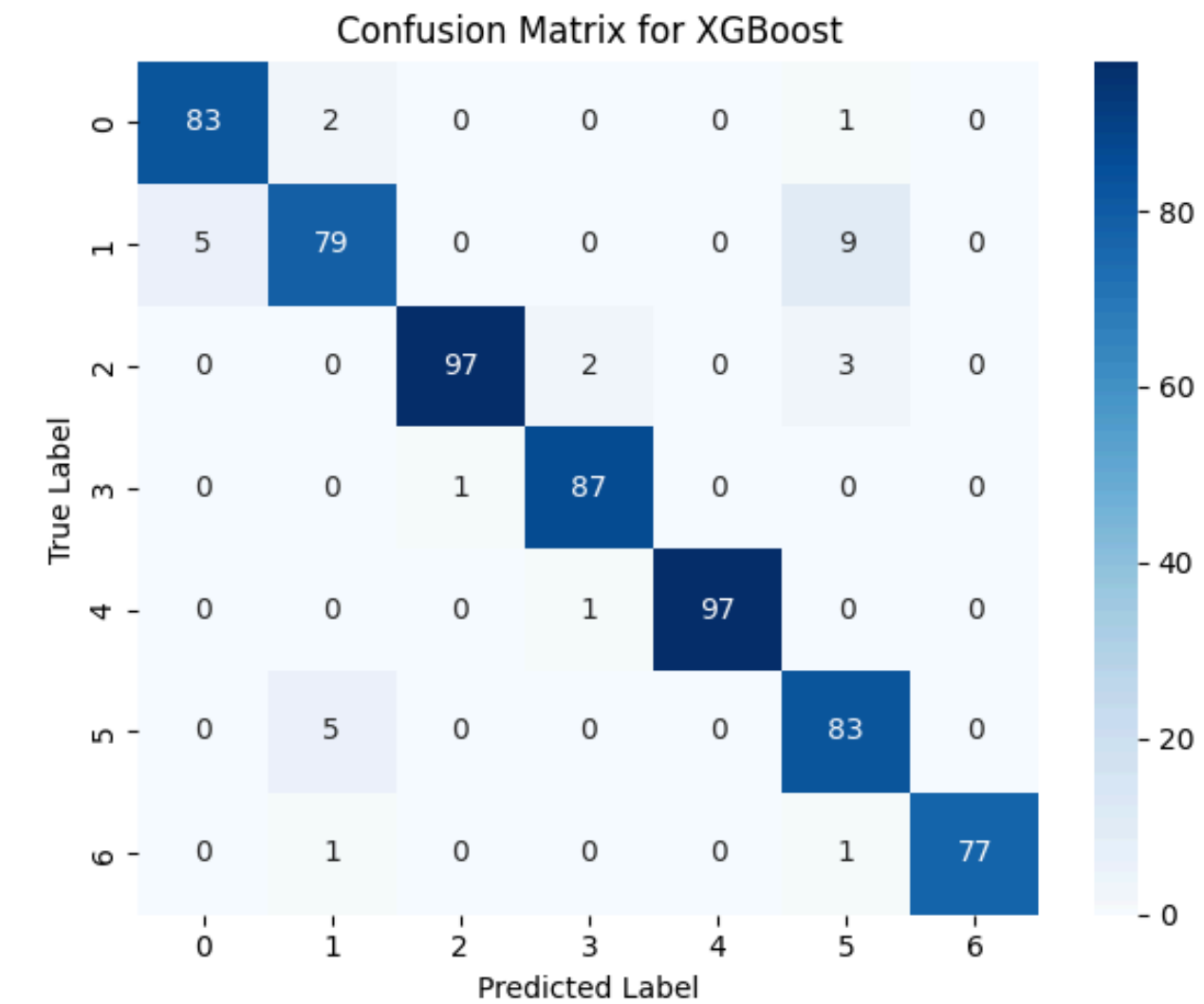
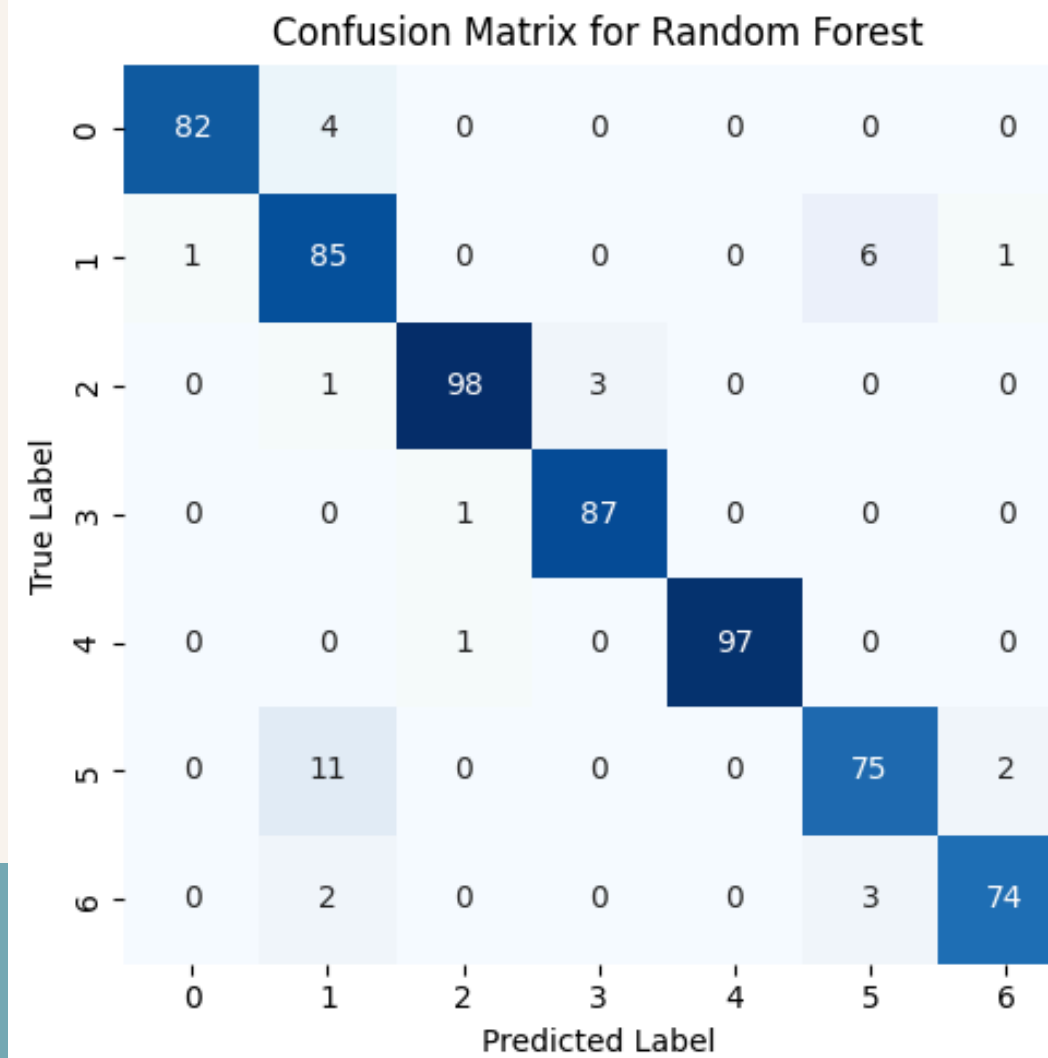
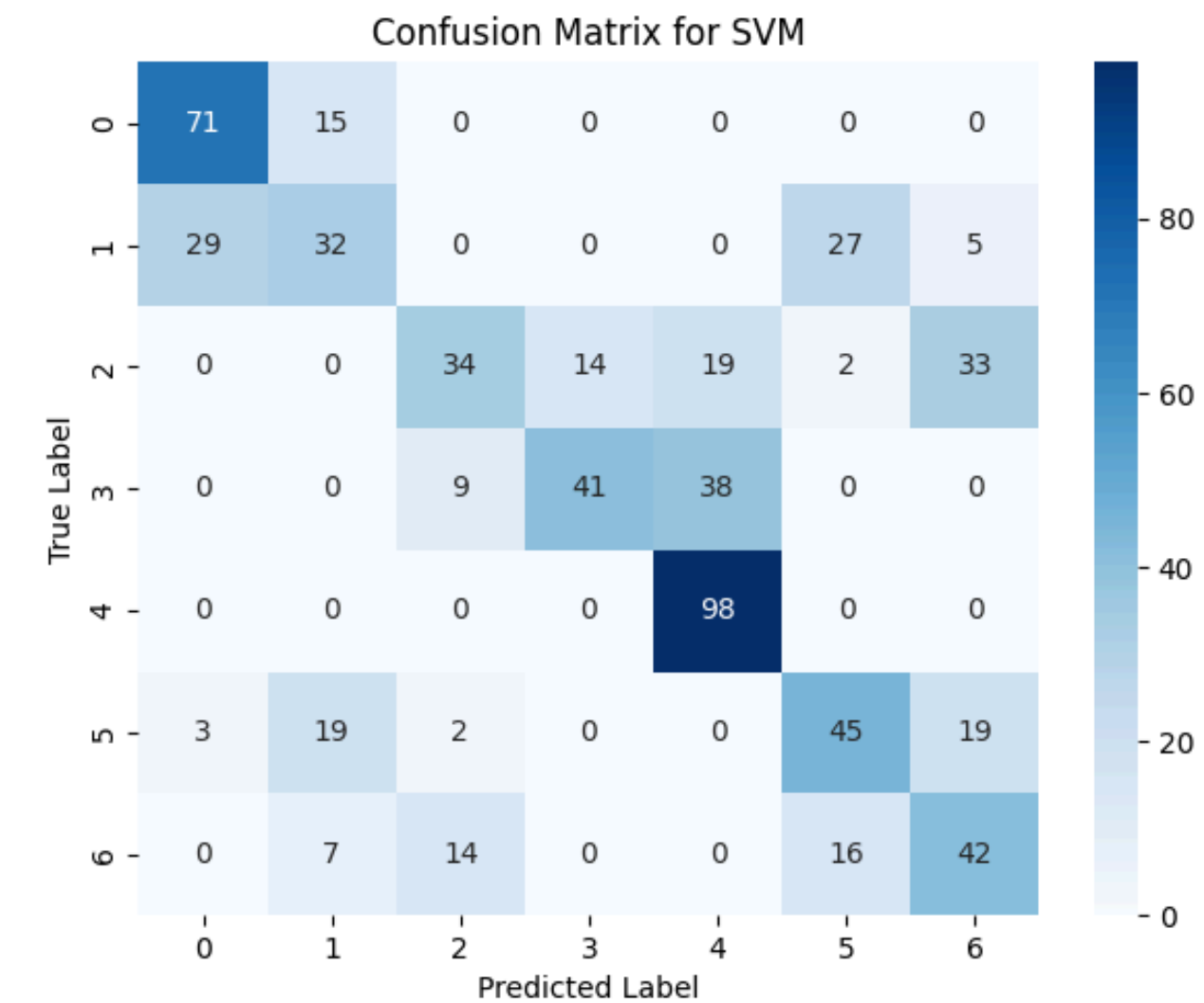
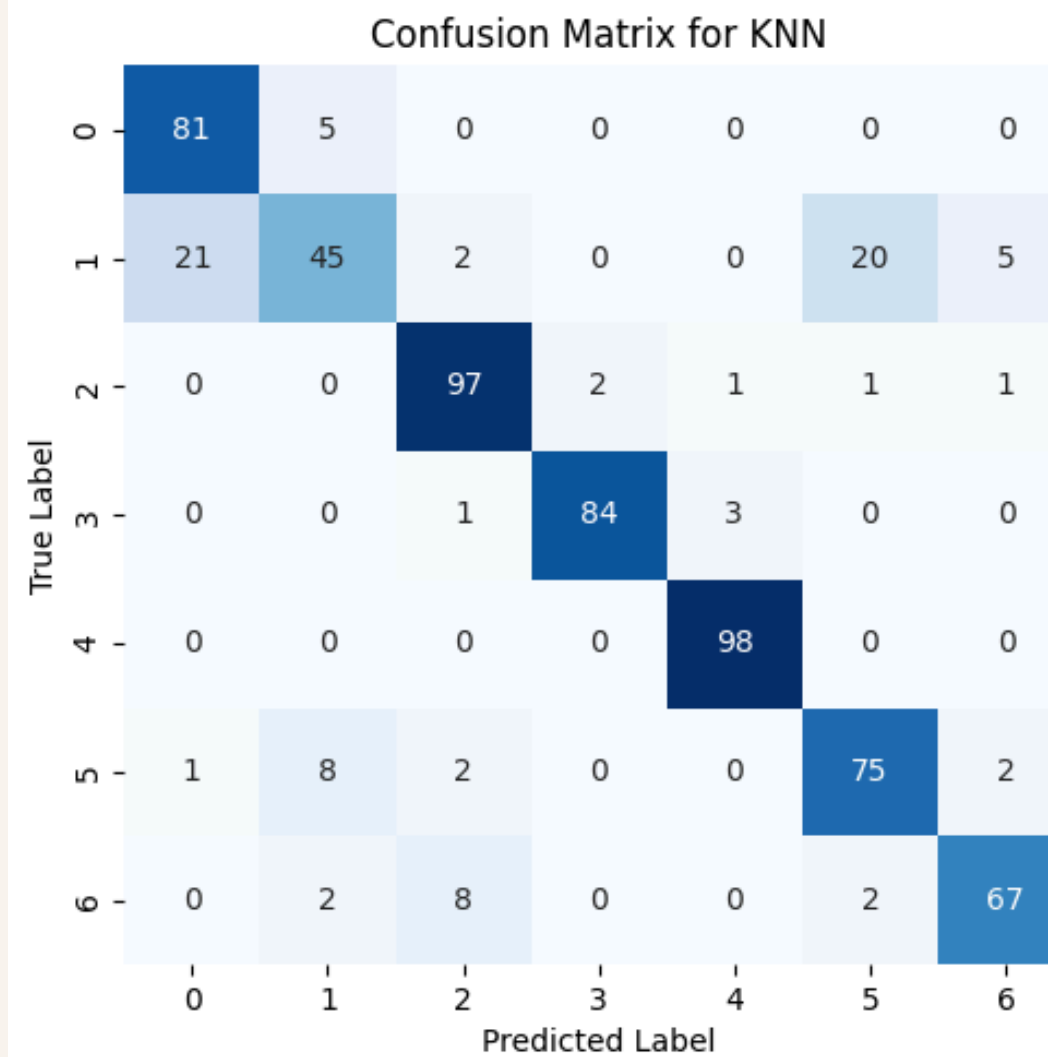
Model Accuracies:

KNN: 0.8628

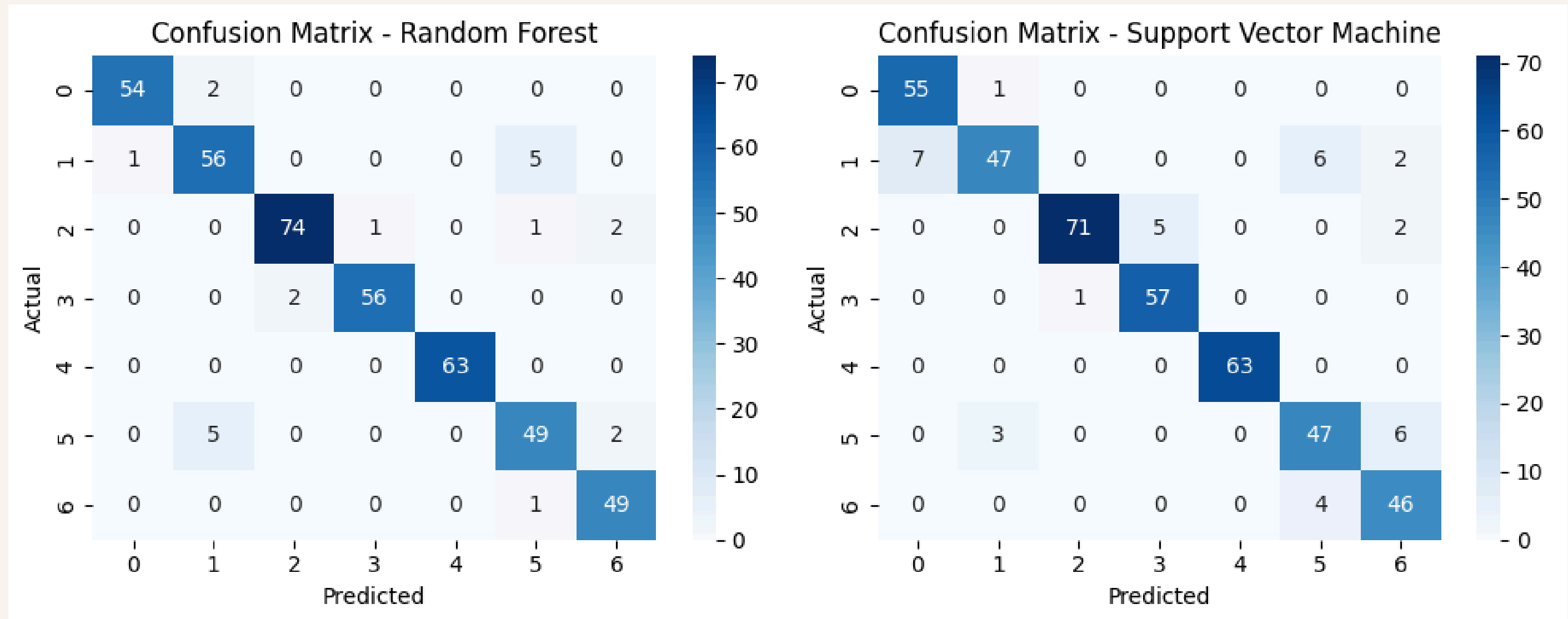
SVM: 0.5726

Random Forest: 0.9432

XGBoost: 0.9511



AFTER MODEL TUNNING



Model Accuracies:

KNN: 0.8913 (CV: 0.8596)

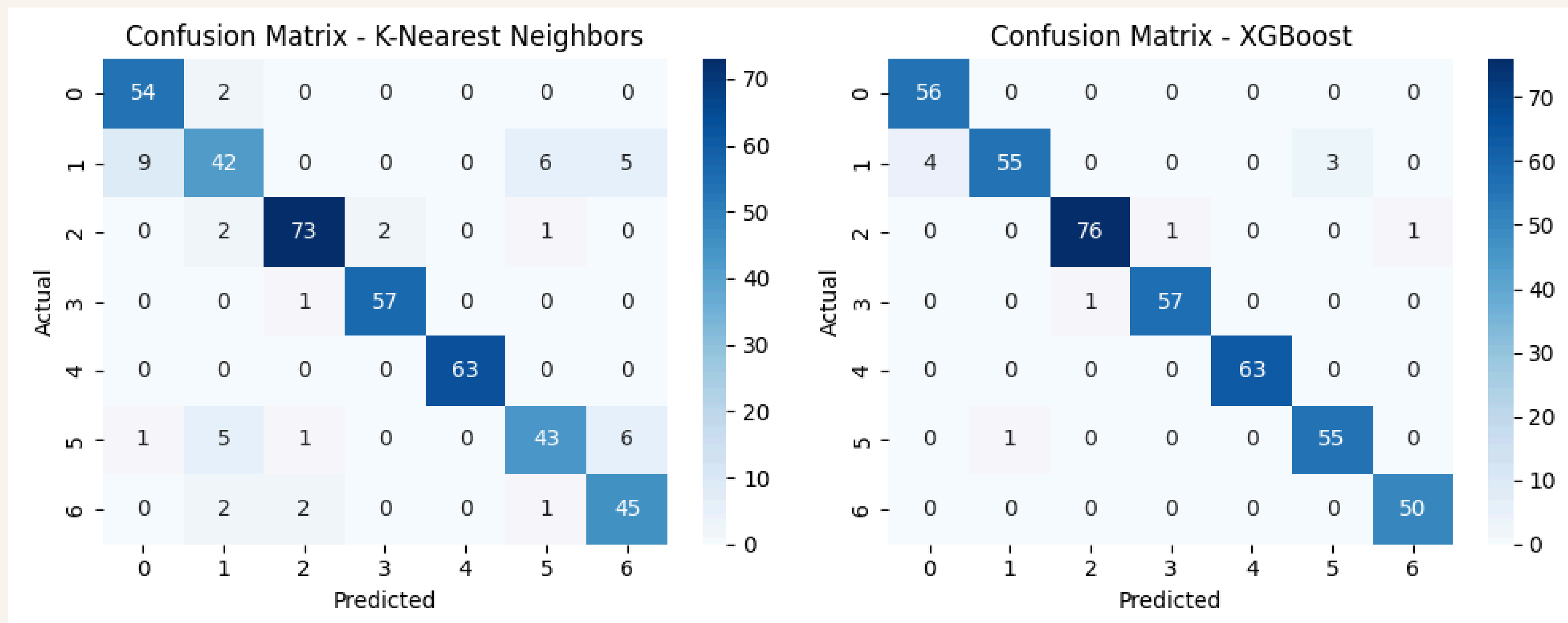
SVM: 0.9125 (CV: 0.8963)

Random Forest: 0.9480 (CV: 0.9491)

XGBoost: 0.9740 (CV: 0.9651)

AFTER MODEL TUNNING

51



Dari hasil modeling di atas, XGBoost dan Random Forest menunjukkan performa terbaik berdasarkan F1-score, sementara SVM memiliki hasil paling rendah. Confusion matrix menunjukkan misclassifications yang signifikan pada beberapa kelas, terutama di model SVM. Oleh karena itu diperlukan preprocessing lebih lanjut, seperti:

- Imbalance Handling
- Feature Selection
- Feature Importance dari Random Forest

COMPARE WITH TEST SCORE

Untuk menentukan apakah hasil modeling **overfitting**, kita perlu melihat perbedaan hasil scoring antara data training dan data testing.

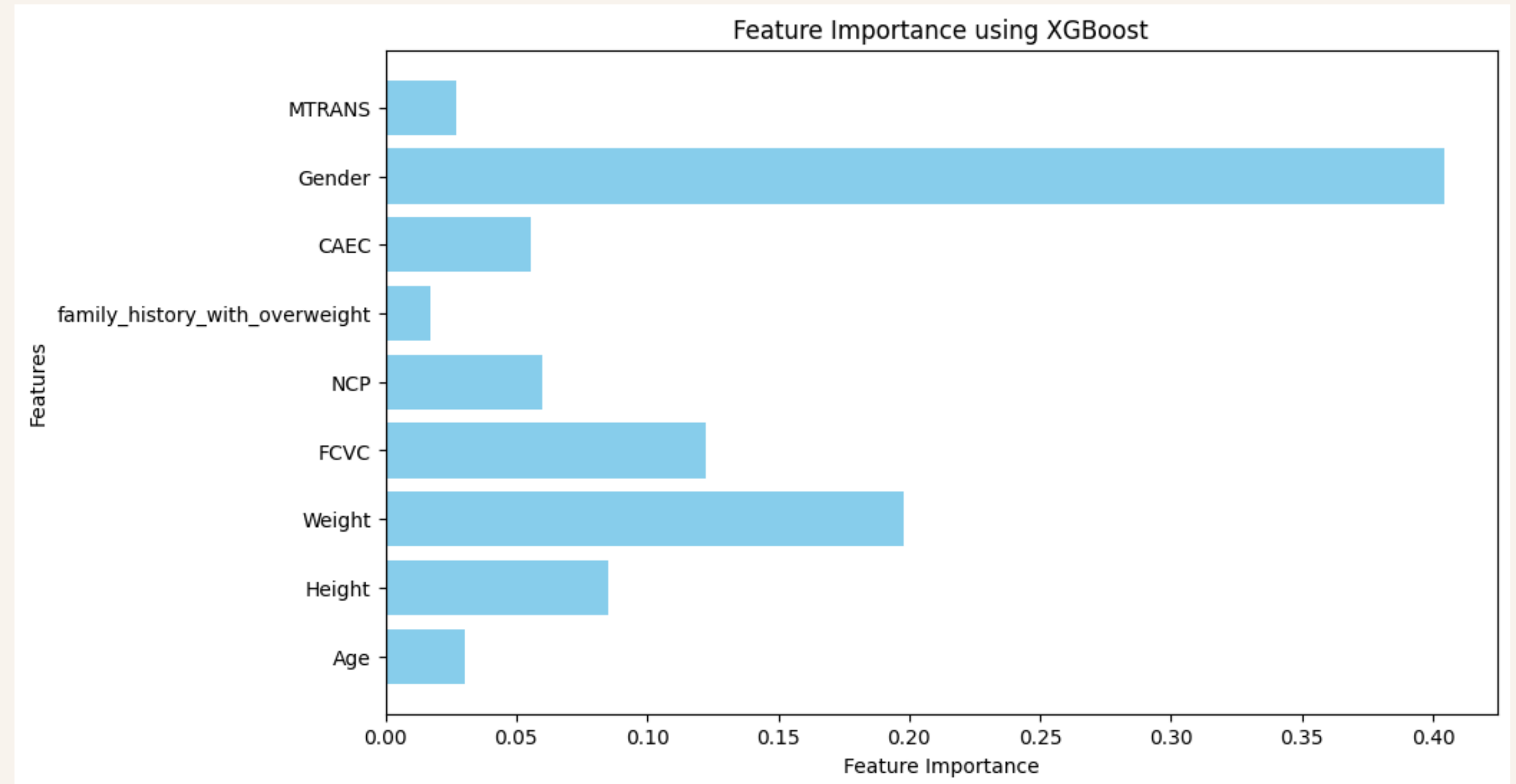
```
Model F1-Score Comparison (Train vs Test):  
KNN: Train = 0.9166, Test = 0.8892, CV = 0.8568  
SVM: Train = 0.9280, Test = 0.9118, CV = 0.8956  
Random Forest: Train = 1.0000, Test = 0.9481, CV = 0.9493  
XGBoost: Train = 1.0000, Test = 0.9737, CV = 0.9650  
  
Overfitting Check:  
KNN does not seem to be overfitting (Train: 0.9166, Test: 0.8892)  
SVM does not seem to be overfitting (Train: 0.9280, Test: 0.9118)  
Random Forest does not seem to be overfitting (Train: 1.0000, Test: 0.9481)  
XGBoost does not seem to be overfitting (Train: 1.0000, Test: 0.9737)
```

Overfitting biasanya ditandai dengan skor akurasi yang sangat tinggi pada data training, tetapi rendah pada data testing. Dapat dilihat bahwa ketika divalidasi dengan data test, **score semua model lebih baik**. Model yang kita gunakan tidak terjadi **Overfitting**

INTERPRETASI MODEL

53

Faktor utama dalam klasifikasi obesitas adalah **Gender**, diikuti oleh Weight dan **FCVC**. Faktor lain seperti **Height**, **CAEC**, dan **NCP** berpengaruh sedang, sementara **MTRANS**, dan **Age** memiliki dampak kecil.



CONCLUSION

54

Model yang dapat memberikan hasil terbaik dari prediksi Attrition dalam tujuan project kita yaitu dengan **Xgboost** dengan F1 Score sebesar **0.9740**

Pencegah obesitas pada individu, dapat memperhatikan faktor-faktor utama seperti **Gender, Weight, dan FCVC (Frekuensi konsumsi sayur)**. Selain itu, faktor lain seperti **Height, CAEC (Konsumsi makanan tinggi kalori), dan NCP (Jumlah makanan per hari)** juga berperan dalam risiko obesitas.

THANK YOU
