MEMPREDIKSI OBESITAS DENGAN MACHINE LEARNING

Moh. Nafis Husen Romadani

DAFTAR ISI

- Pengenalan Proyek
- Overview Data
- Exploratory Data Analysis
- Pembangunan Model
- Kesimpulan

PENGENALAN PROYEK

This project aims to analyze and classify a person's obesity level based on various health and lifestyle factors. Using a dataset with features such as weight, height, eating habits, exercise routines, and more, this project applies exploratory data analysis (EDA) and several machine learning algorithms to gain deeper insights into obesity.

OVERVIEW DATA

This dataset helps estimate obesity levels based on eating habits, family history, and physical condition. It includes data from individuals in Mexico, Peru, and Colombia, covering 16 lifestyle and health-related features with 2111 records. The labels classify obesity levels, ranging from underweight to different obesity types.

OVERVIEW DATA

	Gender	Age	Height	Weight	family_history_with_overweight	FAVC	FCVC	NCP	CAEC	SMOKE	CH2O	SCC	FAF
0	Female	21	1.62	64.00	yes	no	2.0	3.0	Sometimes	no	2.00	no	0.00
1	Female	21	1.52	56.00	yes	no	3.0	3.0	Sometimes	yes	3.00	yes	3.00
2	Male	23	1.80	77.00	yes	no	2.0	3.0	Sometimes	no	2.00	no	2.00
3	Male	27	1.80	87.00	no	no	3.0	3.0	Sometimes	no	2.00	no	2.00
4	Male	22	1.78	89.80	no	no	2.0	1.0	Sometimes	no	2.00	no	0.00
...	...	...	...	...	...	...	...	...	...	...	...	...	...
2106	Female	21	1.71	131.41	yes	yes	3.0	3.0	Sometimes	no	1.73	no	1.68
2107	Female	22	1.75	133.74	yes	yes	3.0	3.0	Sometimes	no	2.01	no	1.34
2108	Female	23	1.75	133.69	yes	yes	3.0	3.0	Sometimes	no	2.05	no	1.41
2109	Female	24	1.74	133.35	yes	yes	3.0	3.0	Sometimes	no	2.85	no	1.14
2110	Female	24	1.74	133.47	yes	yes	3.0	3.0	Sometimes	no	2.86	no	1.03

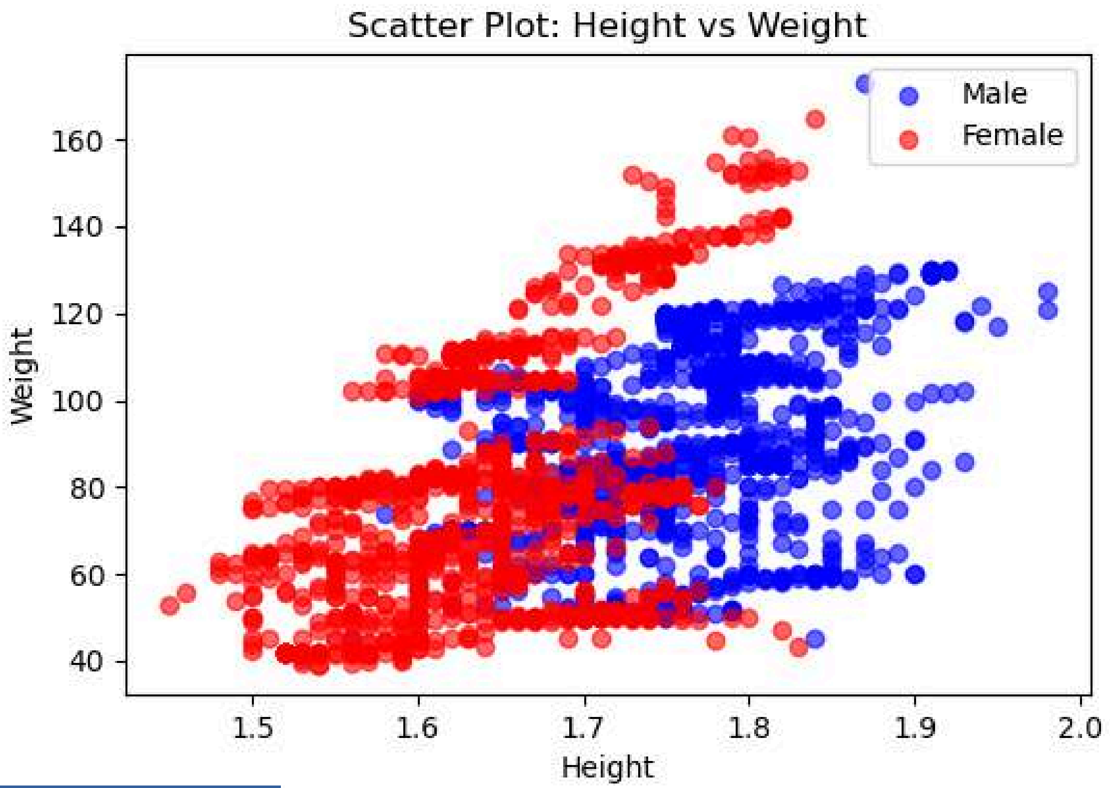
2111 rows × 17 columns

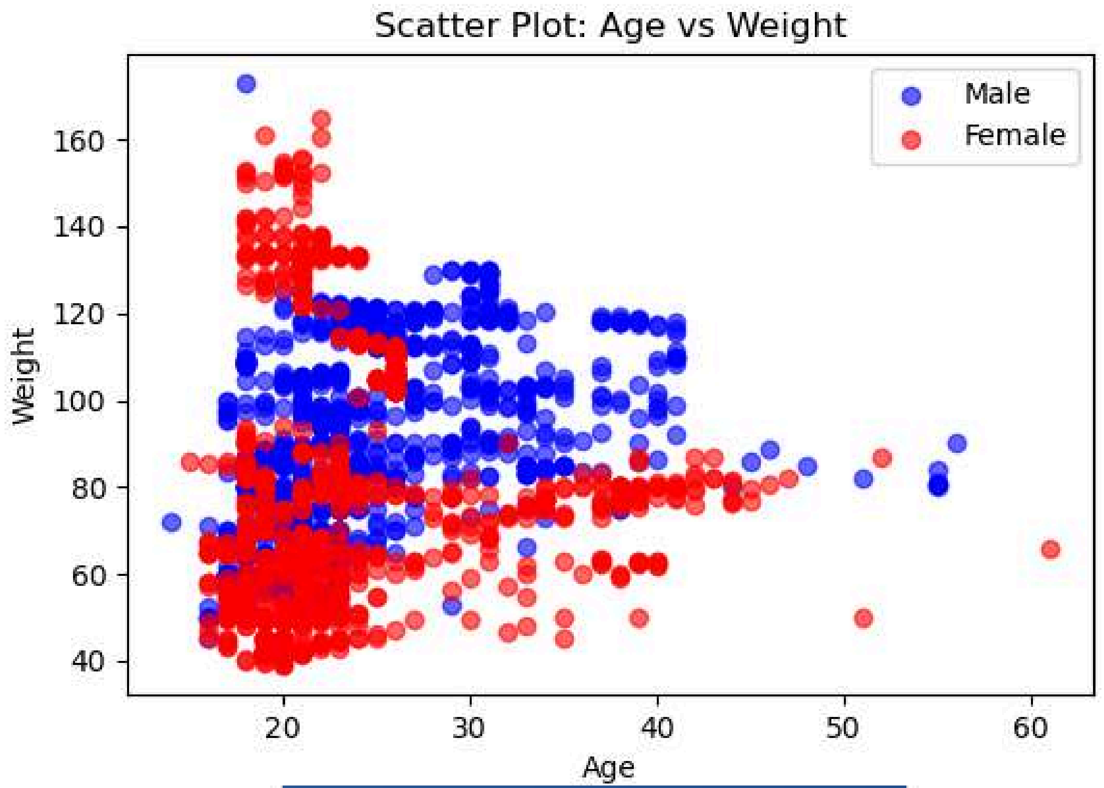
OVERVIEW DATA

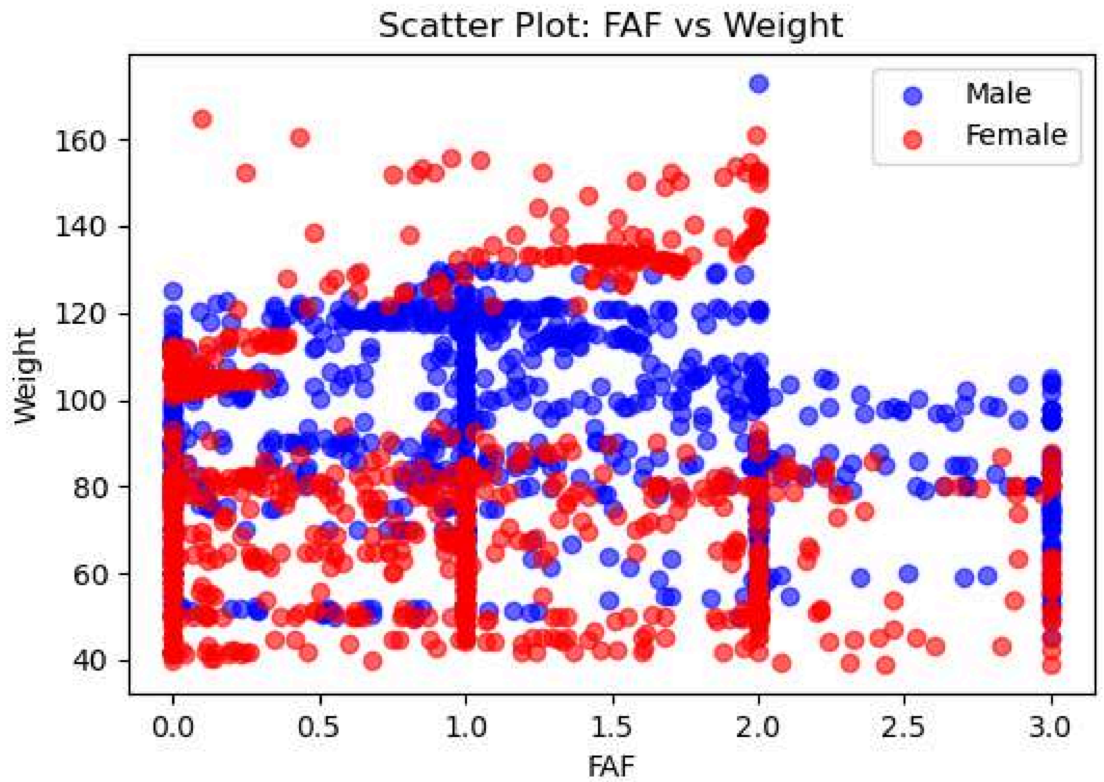
Age	Height	Weight	FCVC	NCP
2111.000000	2111.000000	2111.000000	2111.000000	2111.000000
24.315964	1.701620	86.586035	2.418986	2.685651
6.357078	0.093368	26.191163	0.533996	0.778079
14.000000	1.450000	39.000000	1.000000	1.000000
20.000000	1.630000	65.470000	2.000000	2.660000
23.000000	1.700000	83.000000	2.390000	3.000000
26.000000	1.770000	107.430000	3.000000	3.000000
61.000000	1.980000	173.000000	3.000000	4.000000

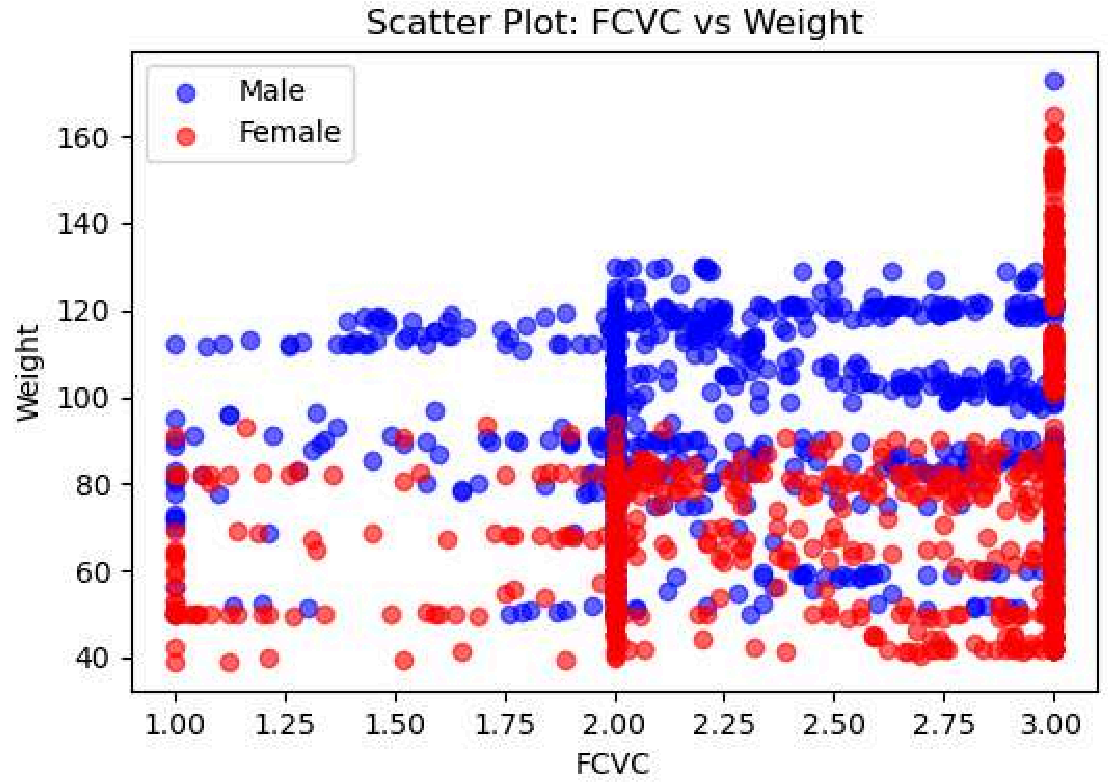
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2111 entries, 0 to 2110
Data columns (total 17 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   Gender                          2111 non-null   object
 1   Age                             2111 non-null   int64
 2   Height                          2111 non-null   float64
 3   Weight                          2111 non-null   float64
 4   family_history_with_overweight  2111 non-null   object
 5   FAVC                            2111 non-null   object
 6   FCVC                            2111 non-null   float64
 7   NCP                             2111 non-null   float64
 8   CAEC                            2111 non-null   object
 9   SMOKE                           2111 non-null   object
 10  CH2O                            2111 non-null   float64
 11  SCC                             2111 non-null   object
 12  FAF                             2111 non-null   float64
 13  TUE                             2111 non-null   float64
 14  CALC                            2111 non-null   object
 15  MTRANS                          2111 non-null   object
 16  NObeyesdad                      2111 non-null   object
dtypes: float64(7), int64(1), object(9)
memory usage: 280.5+ KB
```

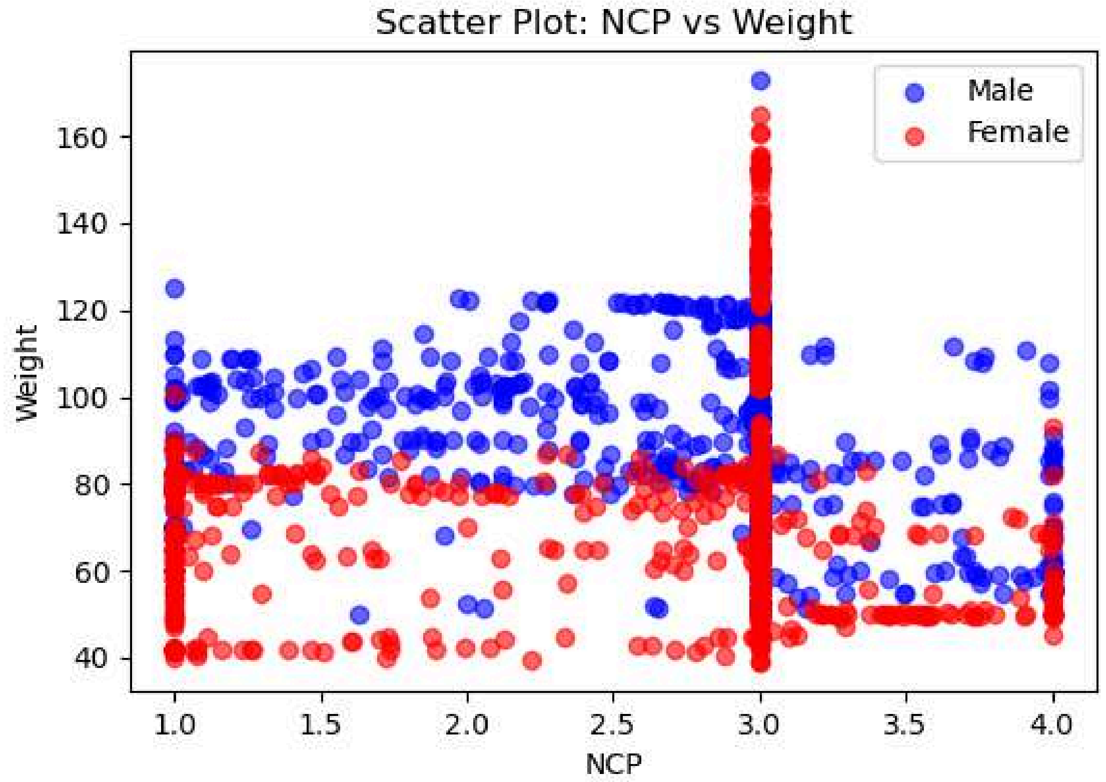
VISUALISASI DATA



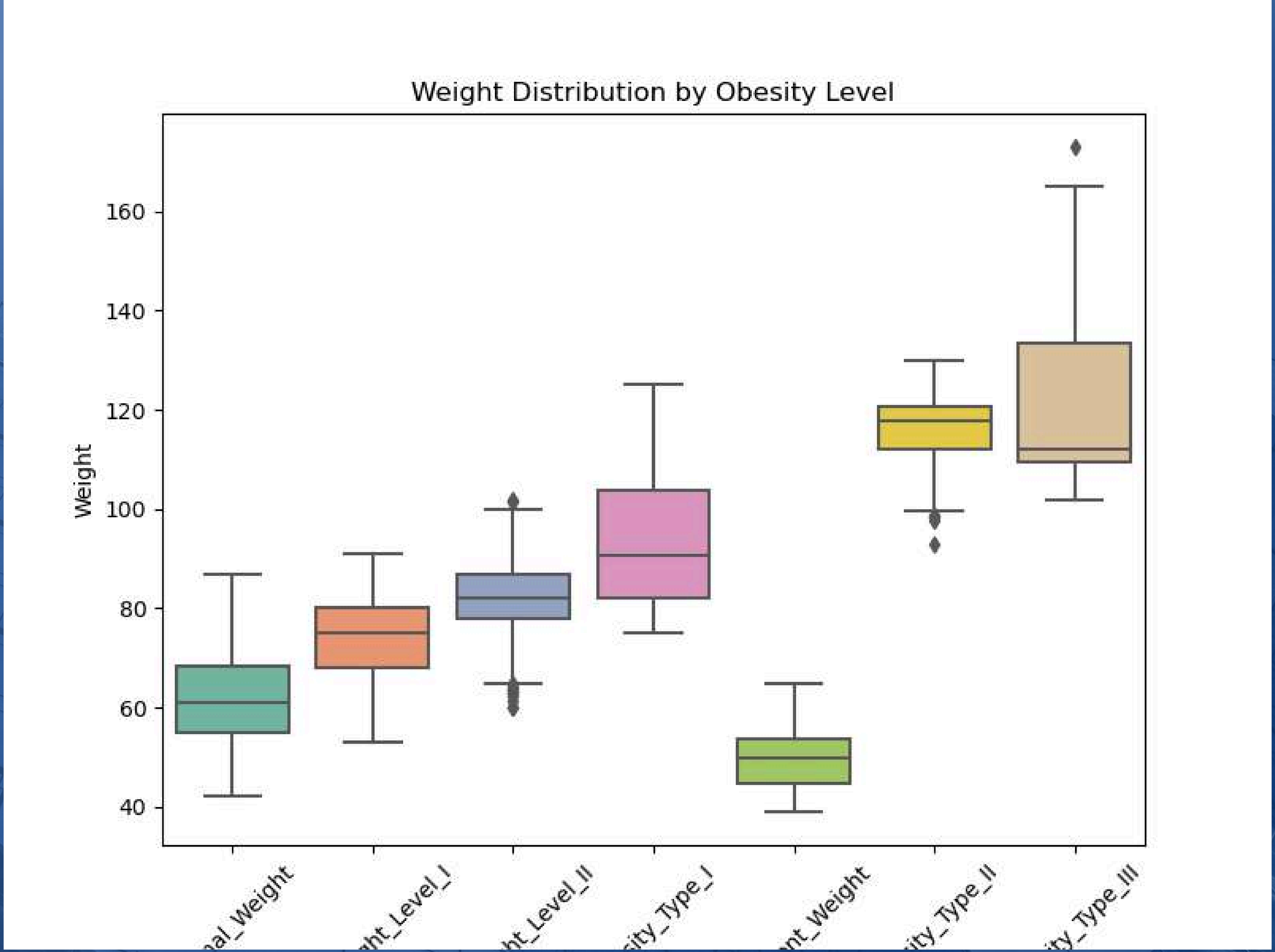






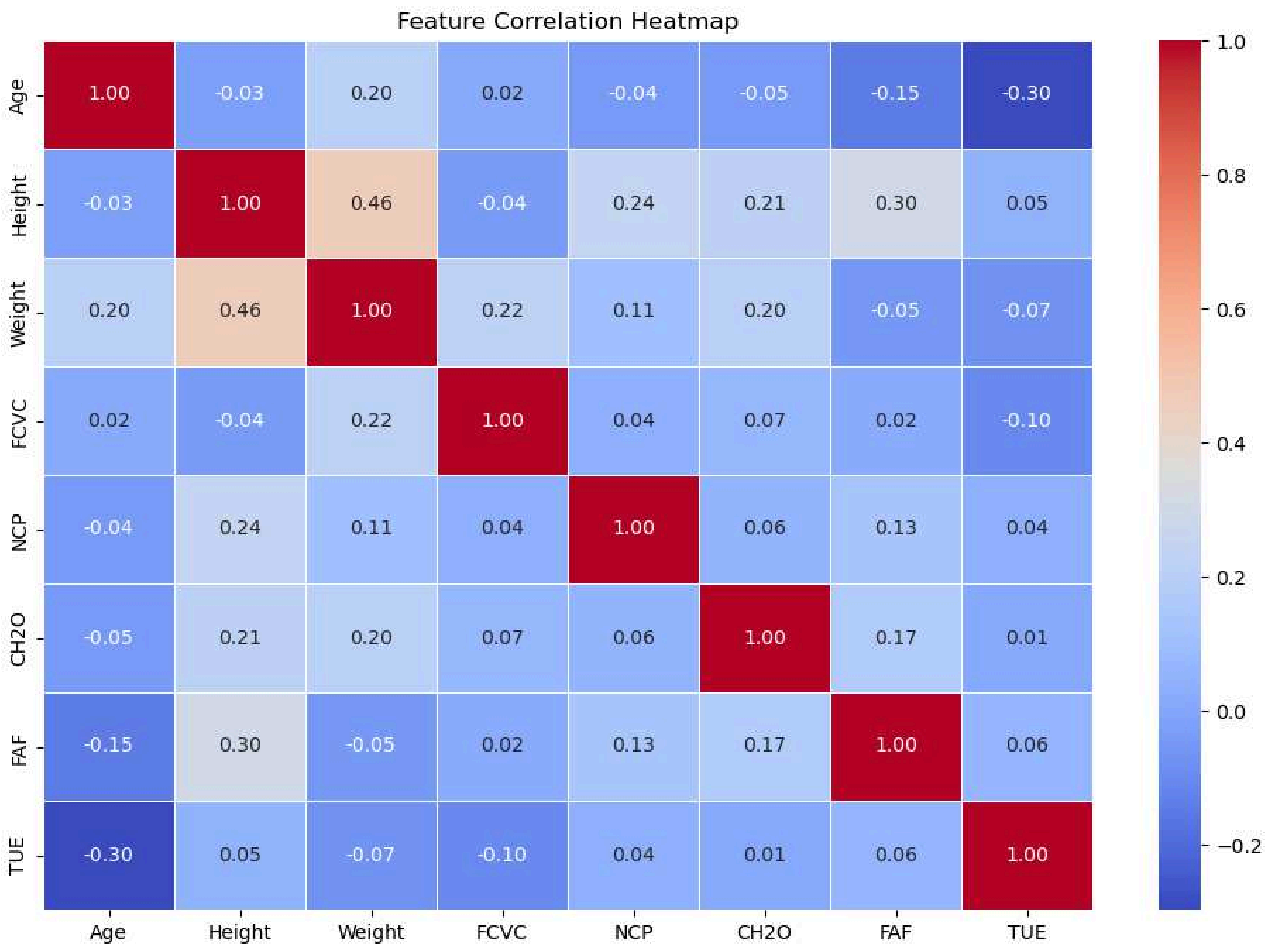


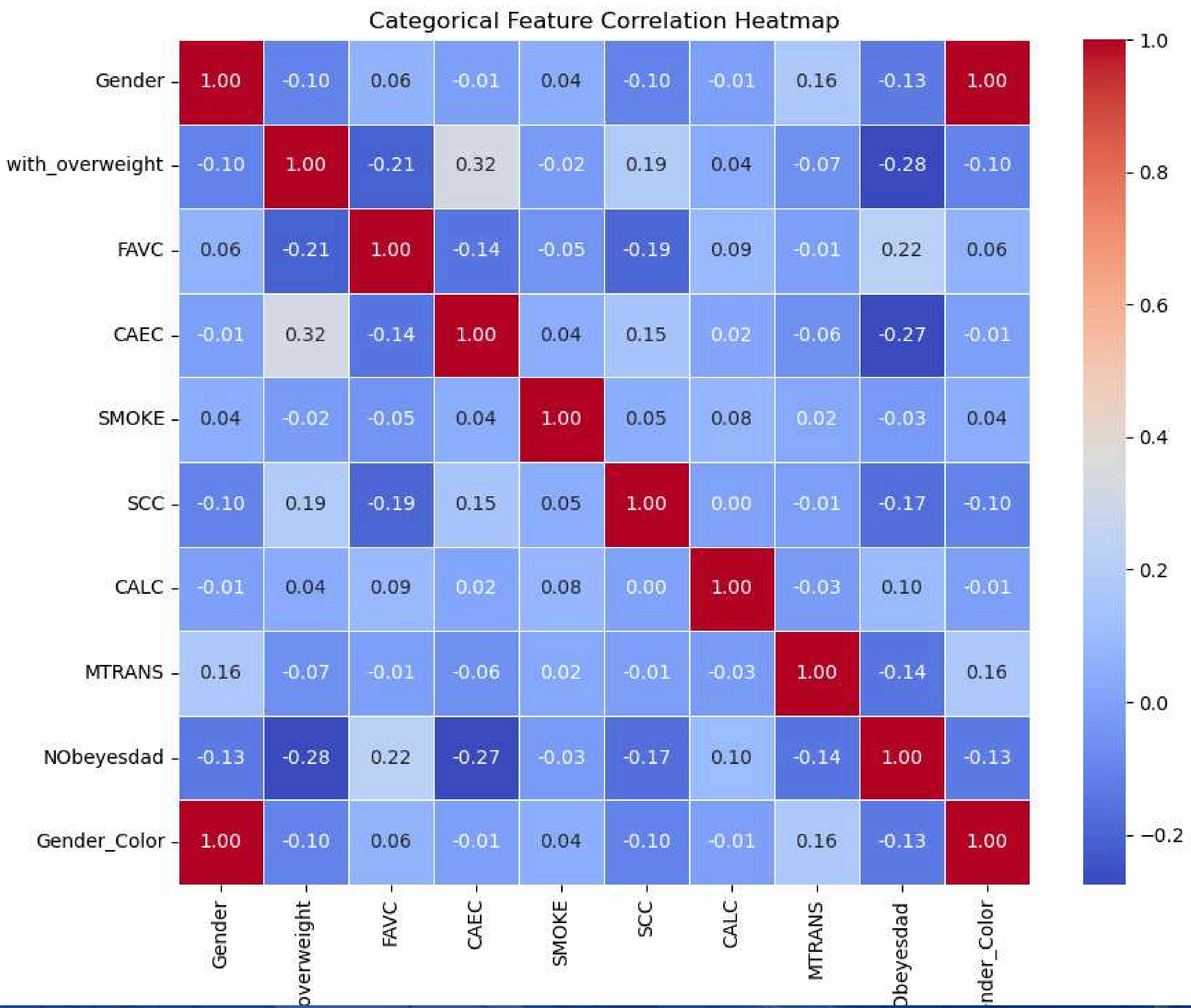
The image shows the relationship between weight and various variables (height, age, physical activity, vegetable consumption, and meal frequency) with gender differences. Height has a positive correlation with weight, where males tend to be taller and heavier than females. Meanwhile, age, physical activity frequency, vegetable consumption, and meal frequency do not show a clear pattern concerning weight. The data distribution indicates that these factors do not directly determine a person's weight.



Weight Distribution by Obesity Level
Weight
160
140
120
100
80
60
40

The boxplot illustrates weight distribution across different obesity levels. Normal weight individuals have a lower weight range (around 40 to 85), while overweight (Weight Level I & II) categories show a higher median and wider spread. Obesity (Type I, II, III) exhibits a significant increase in weight, with Obesity Type III having the widest range and outliers exceeding 160. In contrast, the underweight category has the lowest weight distribution. Overall, the trend indicates that as obesity levels increase, both median weight and variability rise, highlighting a strong correlation between obesity and weight gain.


Feature Correlation Heatmap
Age 1.00 -0.03 0.20 0.02 -0.04 -0.05 -0.15 -0.30
Height -0.03 1.00 0.46 -0.04 0.24 0.21 0.30 0.05
Weight 0.20 0.46 1.00 0.22 0.11 0.20 -0.05 -0.07
FCVC 0.02 -0.04 0.22 1.00 0.04 0.07 0.02 -0.10
NCP -0.04 0.24 0.11 0.04 1.00 0.06 0.13 0.04
CH2O -0.05 0.21 0.20 0.07 0.06 1.00 0.17 0.01
FAF -0.15 0.30 -0.05 0.02 0.13 0.17 1.00 0.06
TUE -0.30 0.05 -0.07 -0.10 0.04 0.01 0.06 1.00
Age Height Weight FCVC NCP CH2O FAF TUE
1.0 0.8 0.6 0.4 0.2 0.0 −0.2


Categorical Feature Correlation Heatmap
Gender 1.00 -0.10 0.06 -0.01 0.04 -0.10 -0.01 0.16 -0.13 1.00
with_overweight -0.10 1.00 -0.21 0.32 -0.02 0.19 0.04 -0.07 -0.28 -0.10
FAVC 0.06 -0.21 1.00 -0.14 -0.05 -0.19 0.09 -0.01 0.22 0.06
CAEC -0.01 0.32 -0.14 1.00 0.04 0.15 0.02 -0.06 -0.27 -0.01
SMOKE 0.04 -0.02 -0.05 0.04 1.00 0.05 0.08 0.02 -0.03 0.04
SCC -0.10 0.19 -0.19 0.15 0.05 1.00 0.00 -0.01 -0.17 -0.10
CALC -0.01 0.04 0.09 0.02 0.08 0.00 1.00 -0.03 0.10 -0.01
MTRANS 0.16 -0.07 -0.01 -0.06 0.02 -0.01 -0.03 1.00 -0.14 0.16
NObeyesdad -0.13 -0.28 0.22 -0.27 -0.03 -0.17 0.10 -0.14 1.00 -0.13
Gender_Color 1.00 -0.10 0.06 -0.01 0.04 -0.10 -0.01 0.16 -0.13 1.00
Gender FAVC CAEC SMOKE SCC CALC MTRANS
1.0 0.8 0.6 0.4 0.2 0.0 −0.2

The heatmaps show correlations between numerical and categorical features related to weight and obesity. Height and weight have a moderate positive correlation (0.46), indicating that taller individuals tend to weigh more. Other numerical features, such as age and physical activity (FAF), have weak correlations with weight, suggesting they are not strong predictors. In the categorical heatmap, alcohol consumption (CAEC) has a moderate correlation (0.32) with overweight, while having no family history of obesity (Noobeyesdad) shows a negative correlation (-0.28) with being overweight. Frequent consumption of high-calorie food (FAVC) has a weak correlation (-0.21) with overweight, implying that weight gain is influenced by multiple factors rather than a single lifestyle choice. Overall, the data suggests that obesity and overweight conditions result from a combination of genetic, dietary, and lifestyle factors rather than one dominant variable.

PEMBANGUNAN MODEL

01 Logistic Regression

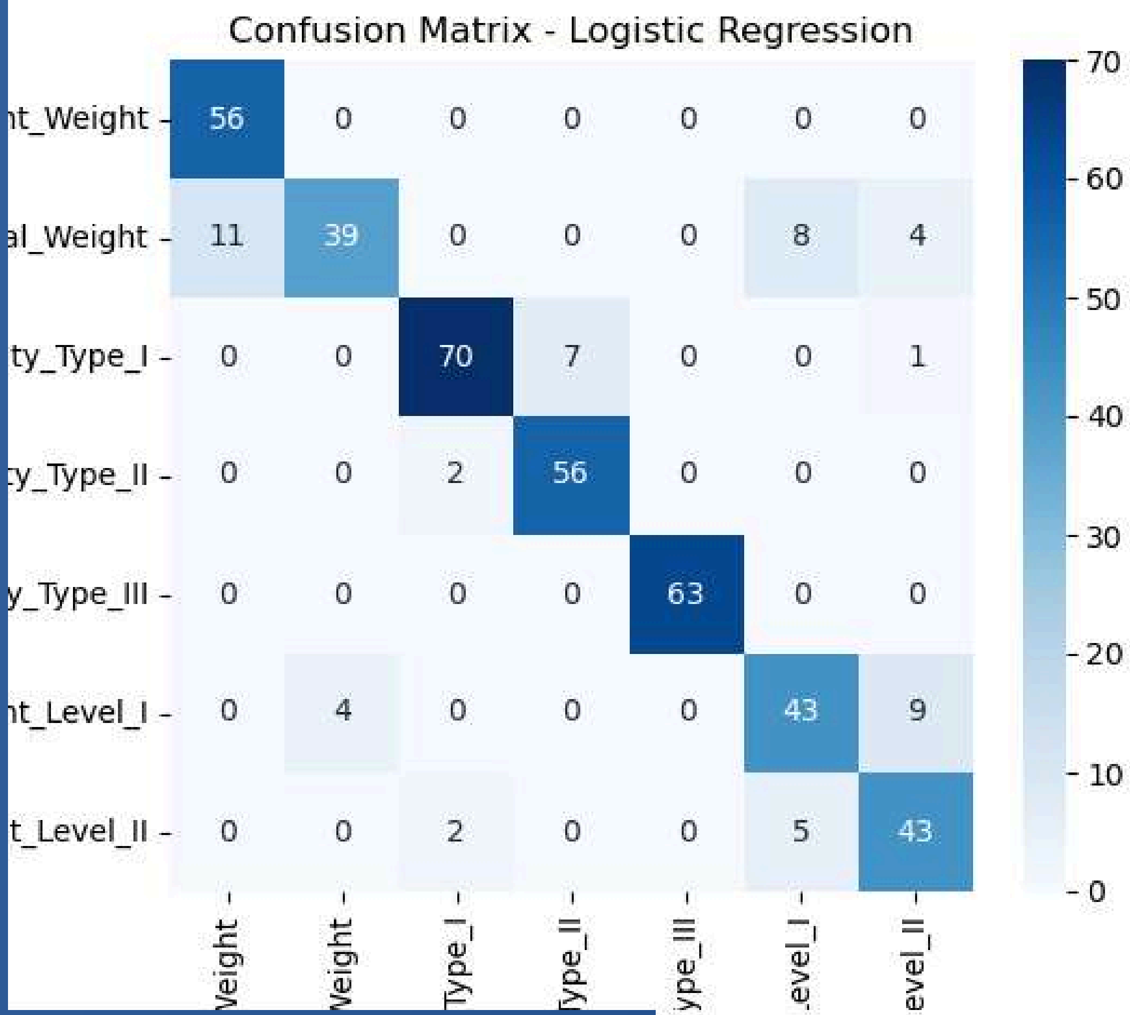
Chosen for its simplicity and interpretability. It is effective for binary and multiclass classification problems, making it a good baseline model.

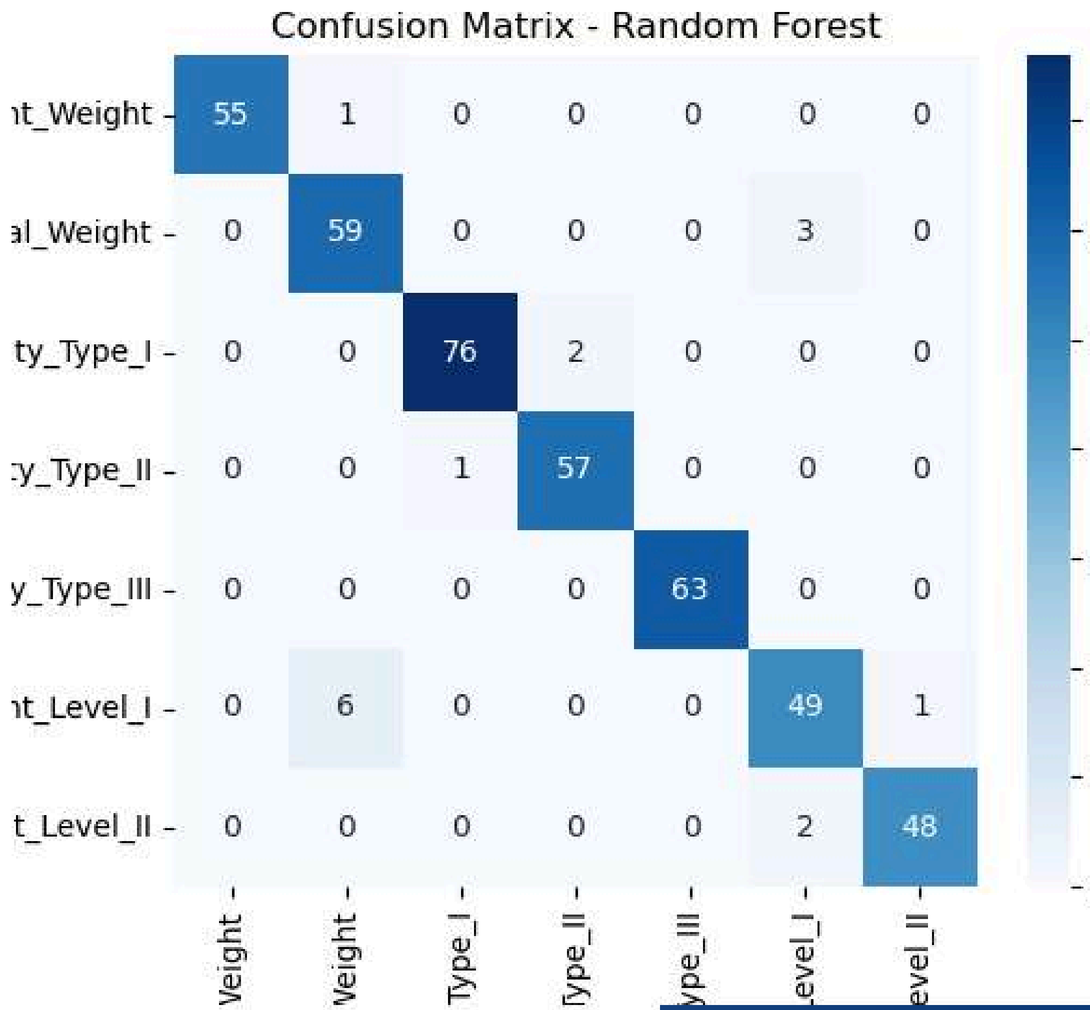
02 Random Forest

Selected for its ability to handle non-linearity, reduce overfitting through ensembling, and provide feature importance insights.

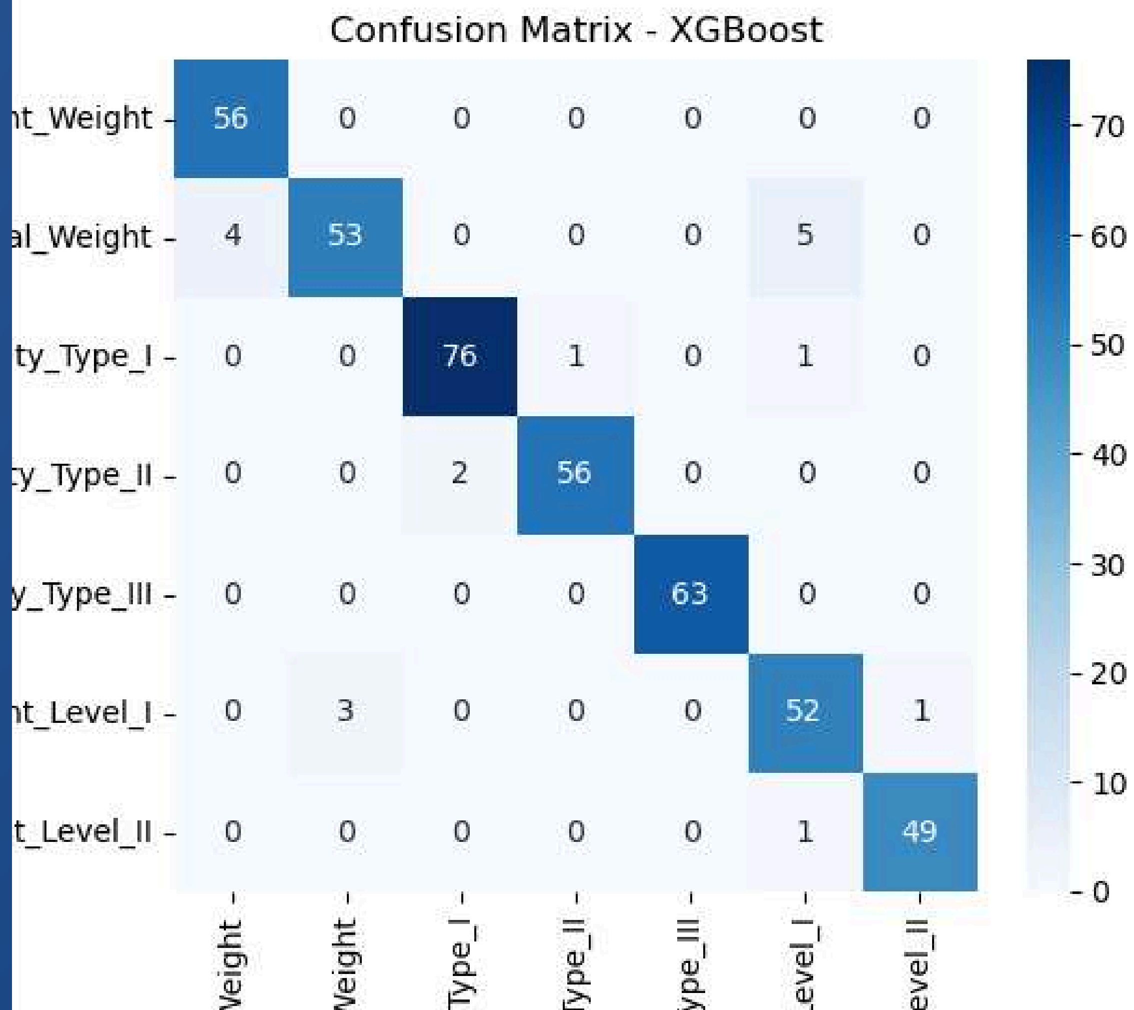
03 XGBoost

Used due to its high performance, efficiency, and ability to handle imbalanced datasets, making it an excellent choice for improving accuracy and robustness.





PERBANDINGAN MODEL



The image presents confusion matrices comparing the performance of Logistic Regression, Random Forest, and XGBoost models in classifying weight categories. The Random Forest and XGBoost models show better classification accuracy, with minimal misclassification compared to Logistic Regression. The Logistic Regression matrix exhibits more misclassifications, particularly in the "Normal Weight" and "Overweight Level I" categories, where some samples are incorrectly classified into neighboring classes. Random Forest and XGBoost display stronger classification capabilities, especially in handling obesity types, with fewer misclassified instances. XGBoost appears to perform slightly better than Random Forest in maintaining correct classifications across all categories, indicating its effectiveness in distinguishing between different weight levels.

THANK YOU