**Estimation of Obesity Levels**

**Based on Eating Habits**

**and**

**Physical Condition**

**By: Valerie Kivayiru Jarenga**

Table of Contents

[Estimation of Obesity Levels Based on Eating Habits and Physical Condition Project 5](#_Toc186647916)

[Problem statement 5](#_Toc186647917)

[Week 1: Data Importing and Cleaning 6](#_Toc186647918)

[Task 1: Import the dataset and inspect its structure. 6](#_Toc186647919)

[Task 2: Data Type Conversion and Encoding 9](#_Toc186647922)

[Task 3: Outlier Detection and Handling 11](#_Toc186647923)

[Task4: Normalization/Standardization 14](#_Toc186647924)

[Week2: Exploratory Data Analysis (EDA) 15](#_Toc186647925)

[Task 1: Summary Statistics 15](#_Toc186647926)

[Task 2: Distribution Analysis 17](#_Toc186647927)

[Task 3: Relationship Exploration 19](#_Toc186647928)

[Task4: Correlation Analysis 23](#_Toc186647929)

[Week3: Advanced Visualizations and Machine Learning 25](#_Toc186647930)

[Task 1: Advanced Visualizations 25](#_Toc186647931)

[Task 2: Feature Engineering and Scaling 29](#_Toc186647932)

[Task 3: Train-Test Split 29](#_Toc186647933)

[Task4: Machine Learning Model Implementation 30](#_Toc186647934)

[Task 5: Model Evaluation 30](#_Toc186647935)

[Conclusion 32](#_Toc186647936)

[Recommendations 32](#_Toc186647937)

[Github repository 33](#_Toc186647938)

**Table of figures**

[Figure 1 Data types of data set 6](#_Toc186646660)

[Figure 2 Identification of missing values 7](#_Toc186646661)

[Figure 3 Data type conversion and encoding 8](#_Toc186646662)

[Figure 4 Datatype conversion and encoding 9](#_Toc186646663)

[Figure 5 Boxplot for weight outlier detection and handling 10](#_Toc186646664)

[Figure 6 Boxplot for height outlier detection and handling 11](#_Toc186646665)

[Figure 7 Boxplot for age outlier detection and handling 12](#_Toc186646666)

[Figure 8 Normalization of Age, Weight and Height using MinMax scaling 13](#_Toc186646667)

[Figure 9 Summary statistics 14](#_Toc186646668)

[Figure 10 Histogram, KDE and Q-Q plot for the distribution of Age 16](#_Toc186646669)

[Figure 11Histogram, KDE and Q-Q plot for the distribution of Weight 17](#_Toc186646670)

[Figure 12Histogram, KDE and Q-Q plot for the distribution of Height 17](#_Toc186646671)

[Figure 13 Box plots to show relationship between weight, height and normal weight 18](#_Toc186646672)

[Figure 14 Boxplots to show relationship between Age, Frequency of physical activity and normal weight 19](#_Toc186646673)

[Figure 15Boxplots to show relationship between weight, height and overweight\_Level\_I 19](#_Toc186646674)

[Figure 16Boxplots to show relationship between Age, Frequency of physical activity and overweight\_Level\_I 20](#_Toc186646675)

[Figure 17Boxplots to show relationship between Weight and Height of NObeyesdad\_Obesity\_Type\_II 21](#_Toc186646676)

[Figure 18Boxplots to show relationship between Age, Frequency of physical activity and NObeyesdad\_Obesity\_Type\_II 21](#_Toc186646677)

[Figure 19Correlation Heatmap of Continuous Features 23](#_Toc186646678)

[Figure 20Pair plots 24](#_Toc186646679)

[Figure 21Confusion matrix 26](#_Toc186646680)

[Figure 22feature importance plot 28](#_Toc186646681)

[Figure 23Train-Test split 28](#_Toc186646682)

[Figure 24Logistic regression and random forest metrics 29](#_Toc186646683)

# Estimation of Obesity Levels Based on Eating Habits and Physical Condition Project

## Problem statement

This project aims to analyze a dataset comprising health and dietary information from individuals in Mexico, Peru, and Colombia to predict obesity levels. By leveraging Python for data analysis and machine learning, the project will explore relationships between physical conditions, eating habits, and obesity, ultimately developing predictive models to estimate obesity levels based on these factors.

Dataset description

This dataset includes data to estimate obesity levels. The records are labeled with a class variable NObeyesdad (Obesity Level), which categorizes individuals into several obesity levels:

**Insufficient Weight**

**Normal Weight**

**Overweight Level I**

**Overweight Level II**

**Obesity Type I**

**Obesity Type II**

**Obesity Type III**

Key Features:

Gender

Age: Age of the individual

Height: Height in meters

Weight: Weight in kilograms

family\_history\_with\_overweight: Whether the individual has a family member

suffering from overweight

FAVC: Whether the individual eats high-calorie food frequently

FCVC: Frequency of vegetable consumption in meals

NCP: Number of main meals consumed daily

CAEC: Food consumption between meals

SMOKE: Whether the individual smokes

CH2O: Amount of water consumed daily

SCC: Whether the individual monitors calorie intake

FAF: Frequency of physical activity

Target Feature:

NObeyesdad: Obesity level categorized into 7 classes.

# Week 1: Data Importing and Cleaning

## Task 1: Import the dataset and inspect its structure.

Data types

The dataset has 17 columns, 2111 rows

Out of the 17 features in the dataset, there are 8 float type and 9 object type variables.

8 columns are of type float64 (representing numerical data).

9 columns are of type object (which indicates string data).

### 

Figure 1 Data types of data set

Looking for any missing values

The figure below shows that there are no missing values.

## 

Figure 2 Identification of missing values

## Task 2: Data Type Conversion and Encoding

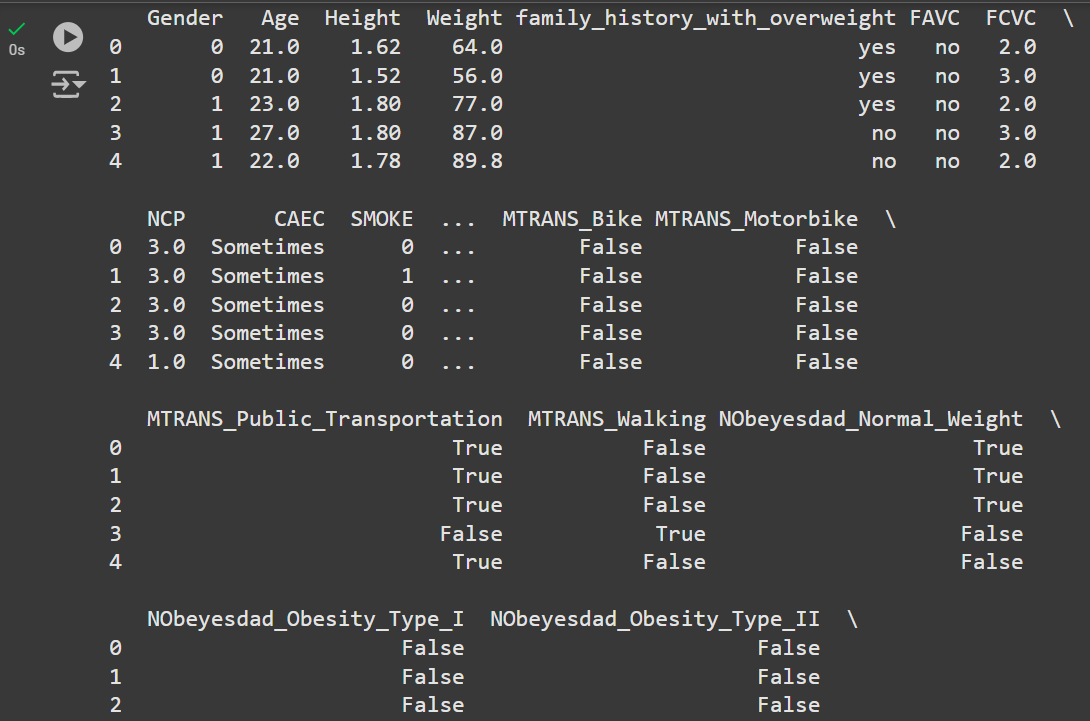


Figure 3 Data type conversion and encoding

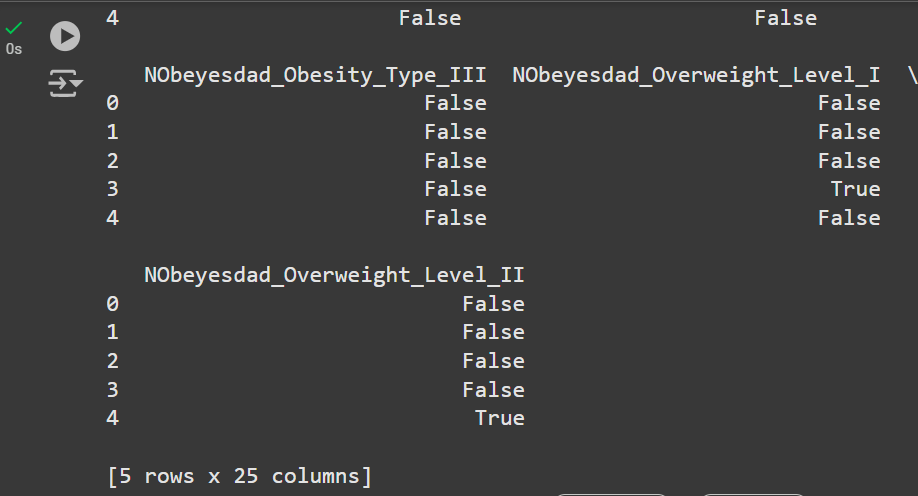


Figure 4 Datatype conversion and encoding

The tables above show label encoded binary variables like Gender, SMOKE, and one-hot encoded multi-class variables like MTRANS, NObeyesdad.

## Task 3: Outlier Detection and Handling

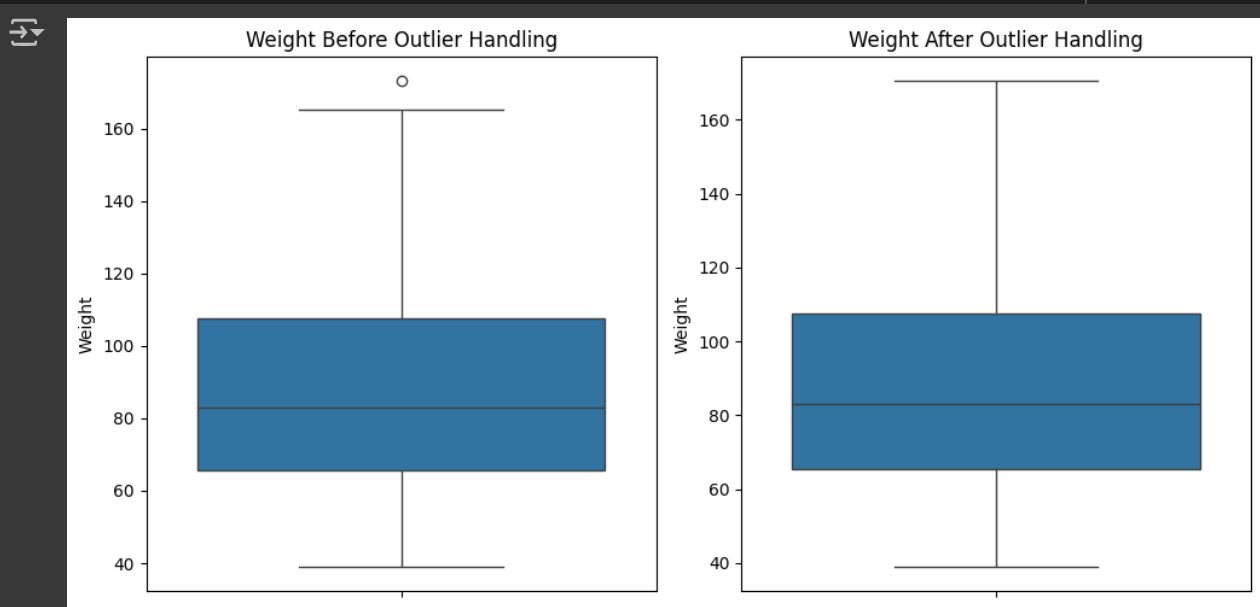


Figure 5 Boxplot for weight outlier detection and handling

The figure above represents weight boxplots

Before Outlier Handling (Left Plot):

The median is closer to the lower part of the box, indicating a slight skew in the data toward higher weights.

There is one outlier above the upper whisker (around 165+ kg), which represents an individual with an unusually high weight.

After Outlier Handling (Right Plot):

The outlier has been removed or capped, as the whisker now extends to the maximum valid weight below 165 kg.

The overall distribution of weights appears largely unchanged, meaning the outlier handling preserved the integrity of the data.

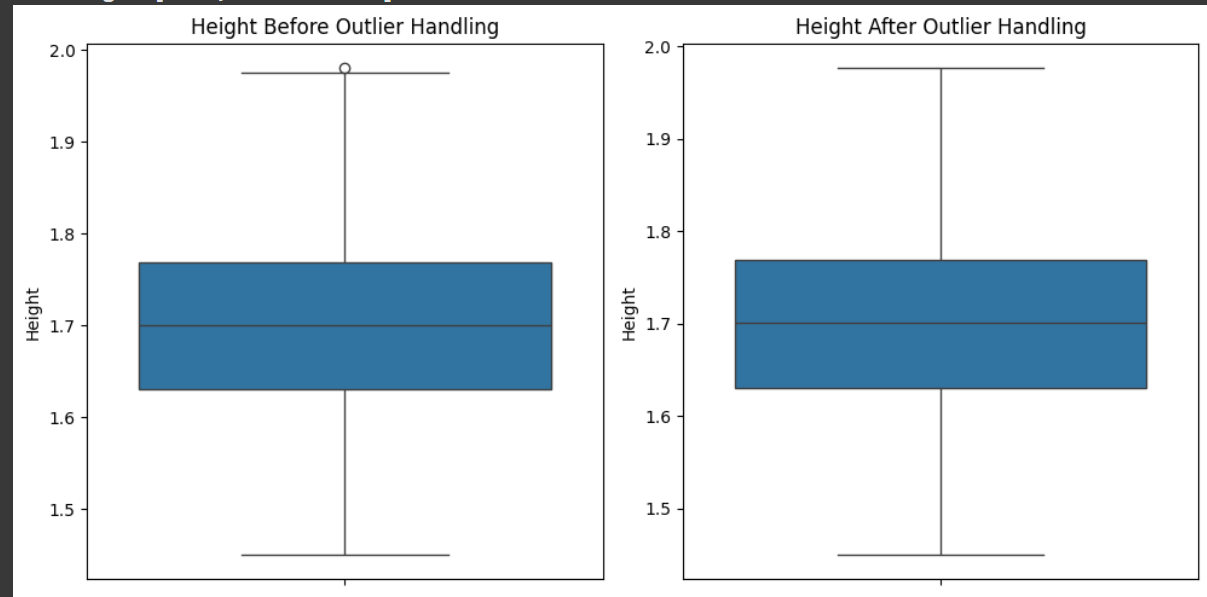


Figure 6 Boxplot for height outlier detection and handling

The figure above represents height boxplots

Before Outlier Handling (Left Plot):

The boxplot shows the interquartile range (IQR), where the central box represents the middle 50% of the data.

The median (central line) is close to the middle of the IQR, indicating a fairly symmetric distribution.

There is one outlier above the whisker (around 2.0 m), which likely represents an unusually tall individual.

After Outlier Handling (Right Plot):

The single outlier has been removed or capped, as the whisker now extends to the highest valid value below 2.0 m.

The distribution remains largely unchanged, indicating that the data-cleaning step was successful in addressing the outlier without distorting the overall data.

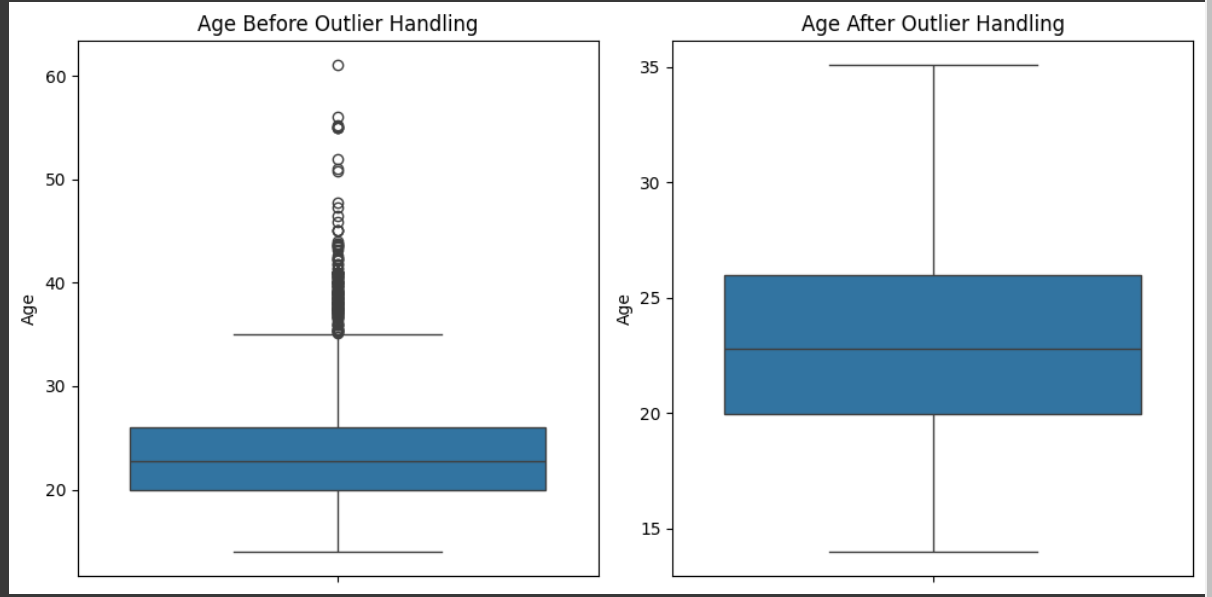


Figure 7 Boxplot for age outlier detection and handling

The figure above represents age boxplots

Before Outlier Handling (Left Plot):

The central box represents the middle 50% of the data (interquartile range, IQR), and the line inside the box indicates the median.

There are numerous outliers above the upper whisker (above ~35 years), with the maximum age extending up to 60.

The whiskers extend to the smallest and largest values within 1.5 \* IQR, which means values beyond this range are considered outliers.

The majority of the data is concentrated between 18 and 25 years, as indicated by the size of the box and position of the median.

After Outlier Handling (Right Plot):

The outliers have been removed, with the maximum age now around 35 years.

The distribution appears more balanced, with no extreme values influencing the visualization.

## Task4: Normalization/Standardization

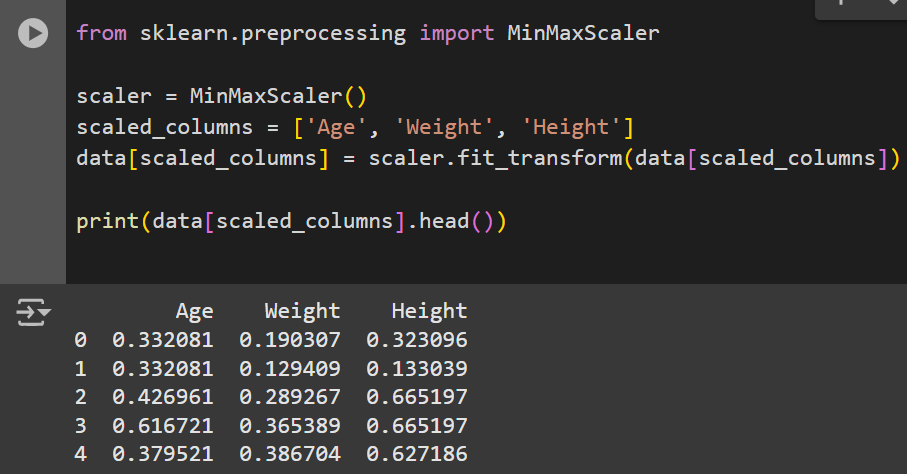


Figure 8 Normalization of Age, Weight and Height using MinMax scaling

In the figure above the 'Age', 'Weight', and 'Height' columns in the dataset have been scaled so that all values fall between 0 and 1. This is to improve the performance of machine learning models.

# Week2: Exploratory Data Analysis (EDA)

## Task 1: Summary Statistics

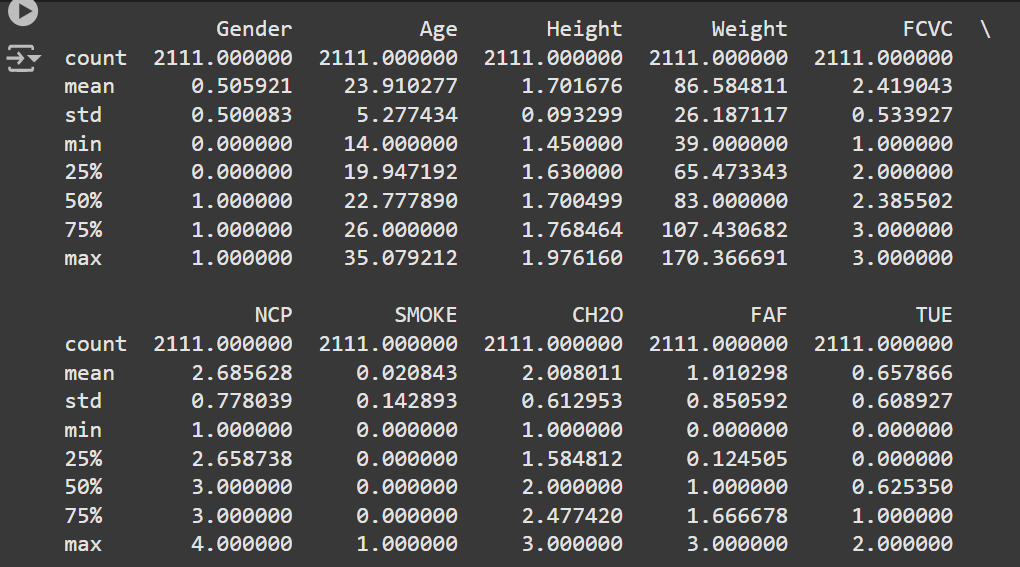


Figure 9 Summary statistics

The table above shows summary statistics of:

**Age:**

Range: The age of individuals ranges from 14 to 35 years.

Mean (Average): The average age is approximately 23.91 years, suggesting most individuals are relatively young.

Median (50%): The median is 22.78, which is slightly lower than the mean, indicating a right-skewed distribution (a few older individuals).

Standard Deviation (5.28): Indicates moderate variability in ages.

**Height (in meters):**

Range: Heights vary from 1.45 m to 1.98 m, representing a realistic range of human heights.

Mean: The average height is 1.70 m, typical for an adult population.

Standard Deviation (0.093): Heights are fairly consistent, with minor variability.

Median: 1.70 m, aligning closely with the mean, suggesting a symmetric distribution.

**Weight (in kg):**

Range: Weights span from 39 kg to 170 kg, covering a wide spectrum, including underweight and obese individuals.

Mean: The average weight is 86.58 kg, which seems higher than typical, possibly due to overweight/obese individuals in the dataset.

Standard Deviation (26.19): Indicates significant variability in weights.

Median (83 kg): Lower than the mean, indicating a right-skewed distribution (some individuals with very high weight).

**FCVC (Frequency of Consumption of Vegetables):**

Range: Values range from 1 to 3, where higher values likely indicate more frequent vegetable consumption.

Mean (2.42): On average, individuals consume vegetables slightly more than 2 times (possibly daily).

Standard Deviation (0.53): Moderate variability in vegetable consumption frequency.

**NCP (Number of Meals per Day):**

Range: From 1 to 4 meals per day.

Mean (2.68): Most people consume between 2 and 3 meals daily.

Median (3.0): Aligns with common dietary habits.

**CH2O (Water Consumption):**

Range: Water intake is between 1 and 3 units, likely referring to glasses or liters.

Mean (2.01): Average water consumption is around 2 units daily, which is moderate.

Standard Deviation (0.61): Moderate variation in water consumption.

**FAF (Physical Activity Frequency):**

Range: Physical activity frequency varies from 0 to 3, where 0 might indicate no activity and 3 represents frequent activity.

Mean (1.01): Average activity level is low.

Standard Deviation (0.85): Shows variability in activity levels, with some very active and others completely inactive.

**TUE (Time Using Technology):**

Range: Time spent on technology ranges from 0 to 2 (likely measured in hours or categories).

Mean (0.66): Average use is relatively low, less than 1 unit daily.

Standard Deviation (0.61): Indicates variability in technology usage habits.

## Task 2: Distribution Analysis

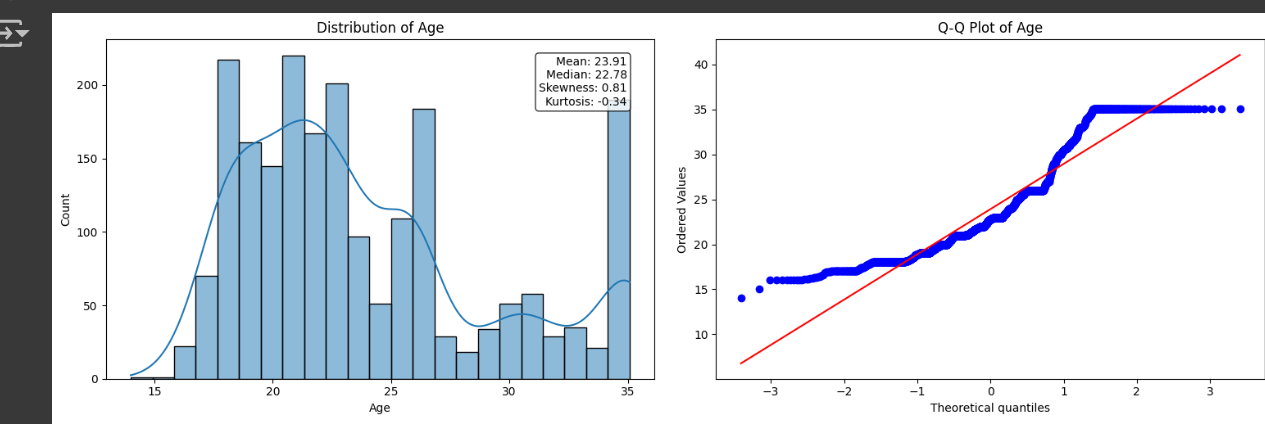


Figure 10 Histogram, KDE and Q-Q plot for the distribution of Age

The figure above represents age distribution:

Shows a multi-modal distribution (multiple peaks)

It has a higher kurtosis compared to the height and weight distribution graphs indicating it has more outliers

Positive skewness (0.81), indicating a right tail

The Q-Q plot shows deviation from normal, especially at the tails, confirming it's not normally distributed

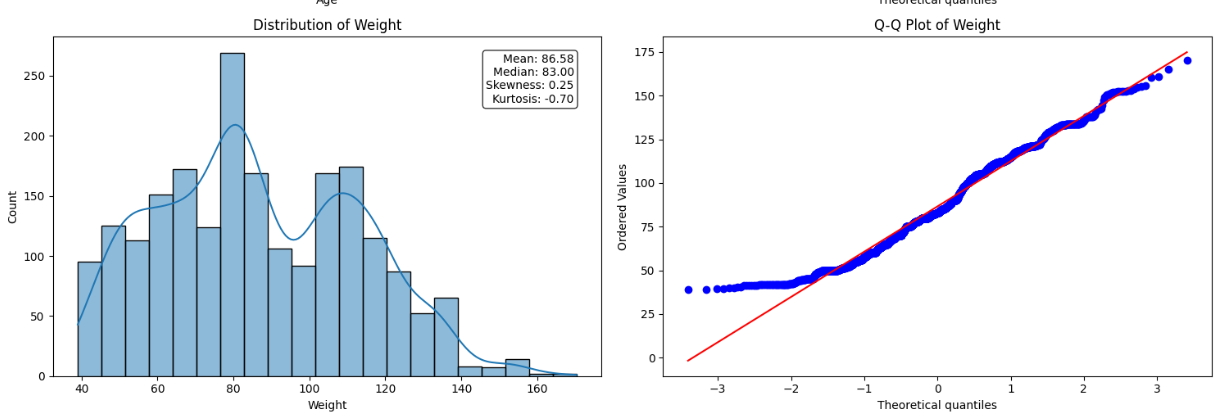


Figure 11Histogram, KDE and Q-Q plot for the distribution of Weight

The figure above represents weight distribution:

More spread-out distribution

Slight positive skewness (0.25)

The Q-Q plot follows the diagonal line more closely than age, suggesting it's closer to a normal distribution, though still with some deviations

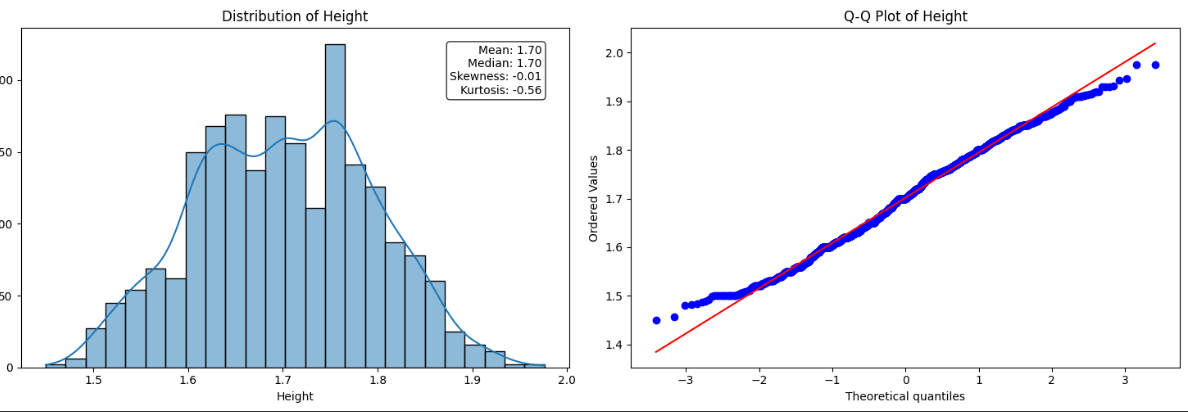


Figure 12Histogram, KDE and Q-Q plot for the distribution of Height

The figure above represents height distribution:

Most symmetric of the three variables

Nearly zero skewness (-0.01)

Bell-shaped curve, though slightly flatter than a normal distribution (negative kurtosis: -0.56)

The Q-Q plot shows the best fit to normality among all three variables, with points following the diagonal line quite closely

**Overall summary**

Height appears to be the most normally distributed variable, while Age shows the most deviation from normality with multiple distinct peaks. Weight falls somewhere in between, showing moderate deviation from normality but less extreme than Age.

## Task 3: Relationship Exploration

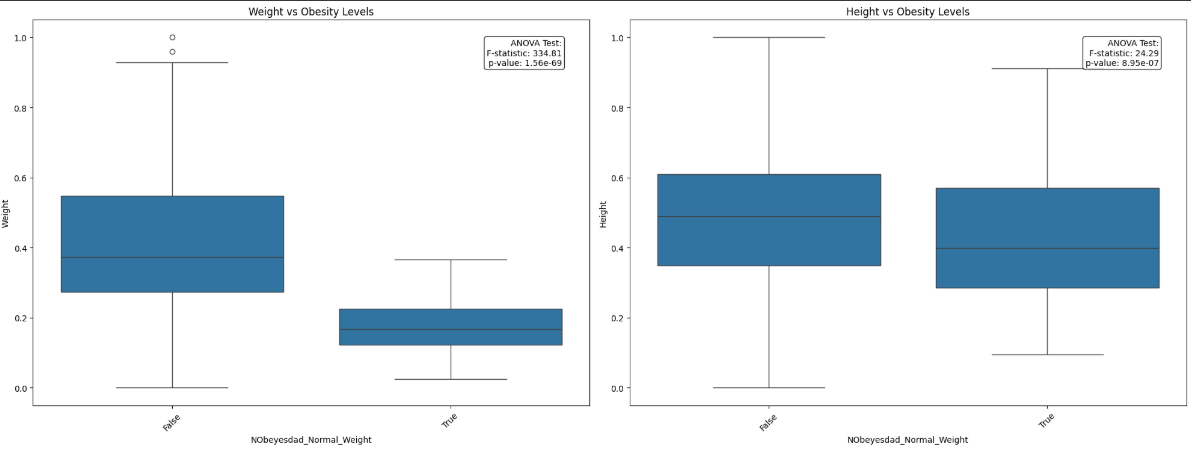


Figure 13 Box plots to show relationship between weight, height and normal weight

The figure above represents boxplots to show the relationship between weight of individuals who are of normal weight and height of individuals who are of normal weight. The (True) NObeyesdad\_Normal\_Weight group weigh less than the (False) NObeyesdad\_Normal\_Weight group. In the height, the difference between groups is less pronounced than with weight

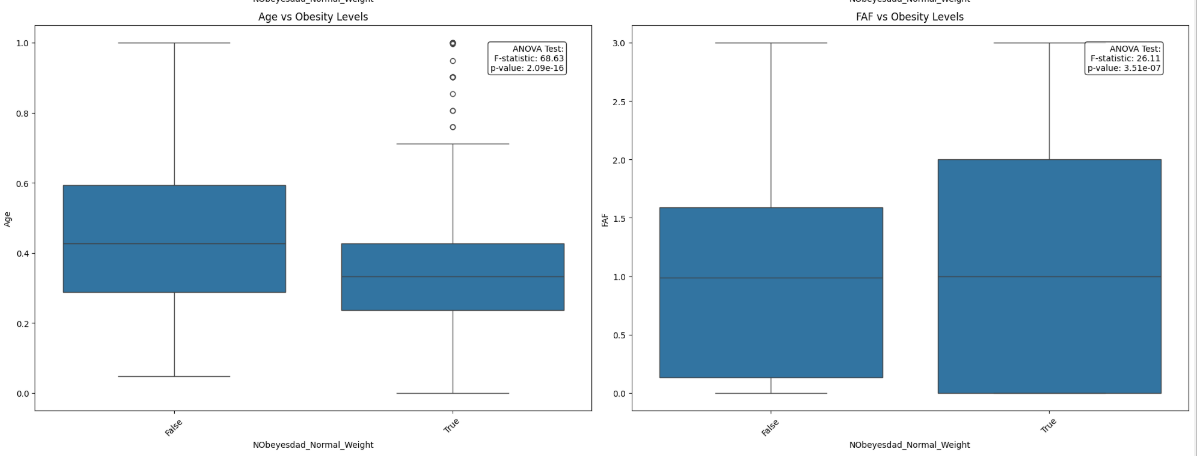


Figure 14 Boxplots to show relationship between Age, Frequency of physical activity and normal weight

In the figure above the (True) NObeyesdad\_Normal\_Weight group have a higher frequency in physical activity than the (False) NObeyesdad\_Normal\_Weightgroup and in terms of age several outliers are visible in the "True" group at higher age values.

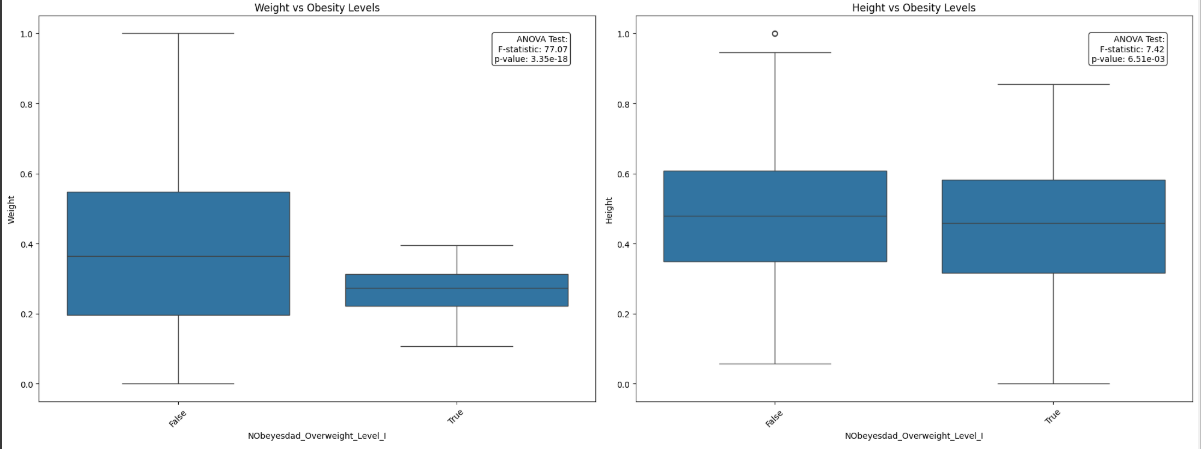


Figure 15Boxplots to show relationship between weight, height and overweight\_Level\_I

The figure above shows:

Weight vs NObeyesdad\_Overweight\_Level\_I:

There's a clear difference between the two groups;

The "True" (NObeyesdad\_Overweight\_Level\_I group has lower weight values (median around 0.25)

The "False" (NObeyesdad\_Overweight\_Level\_I) group has higher weight values (median around 0.35)

The ANOVA test shows a highly significant difference (p-value: 3.35e-19)

Height vs NObeyesdad\_Overweight\_Level\_I:

The difference between groups is less pronounced than with weight

There's one notable outlier in the "False" group

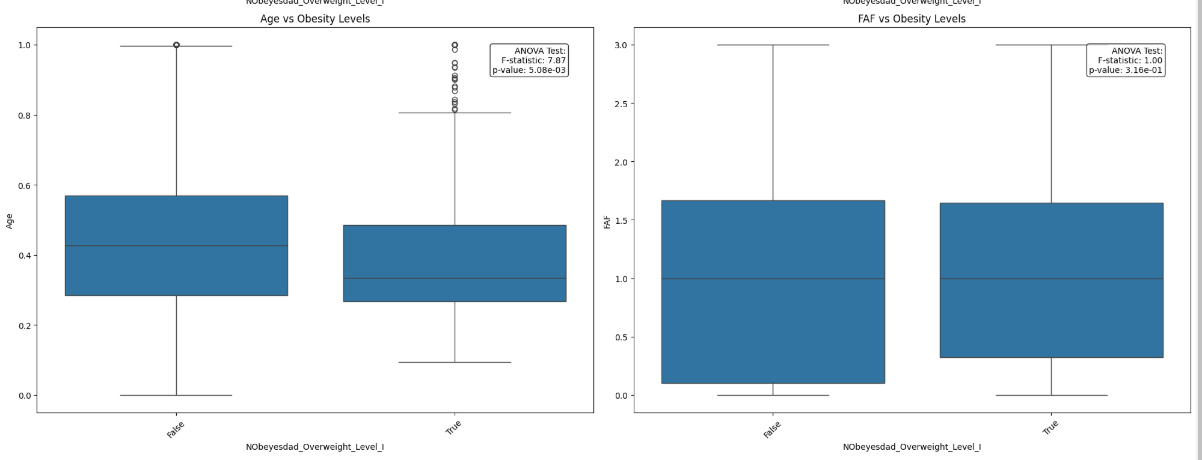


Figure 16Boxplots to show relationship between Age, Frequency of physical activity and overweight\_Level\_I

The boxplots above show:

Age vs NObeyesdad\_Overweight\_Level\_I:

The "True" group has a slightly lower median age

Several outliers are visible in the "True" group at higher age values

The ANOVA test shows a significant difference (p-value: 5.08e-03)

FAF vs NObeyesdad\_Overweight\_Level\_I:

Both groups have similar median values (around 1.0)

The spreads are relatively similar

The ANOVA test shows a non-significant difference (p-value: 3.16e-01)

FAF (Frequency of physical activity) is not strongly associated with obesity status

**Overall summary**

Weight shows the strongest relationship with obesity levels, while FAF shows the weakest. Height and Age show moderate relationships. The ANOVA test p-values confirm these visual interpretations, with Weight having the most significant difference between groups.

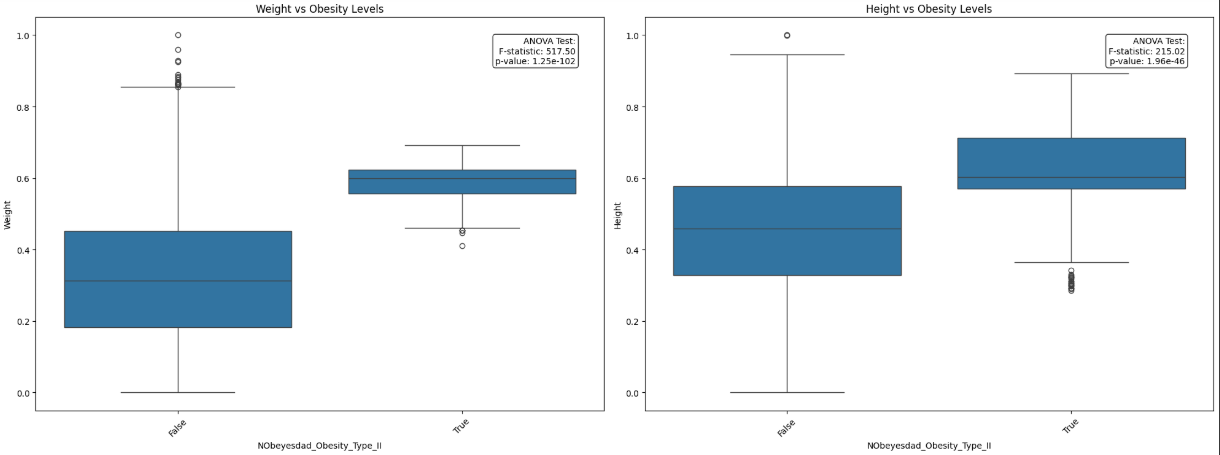


Figure 17Boxplots to show relationship between Weight and Height of NObeyesdad\_Obesity\_Type\_II

In the figure above the individuals in the obesity\_type\_II group weighed more and had higher height than individuals who are not in the obesity\_type\_II group

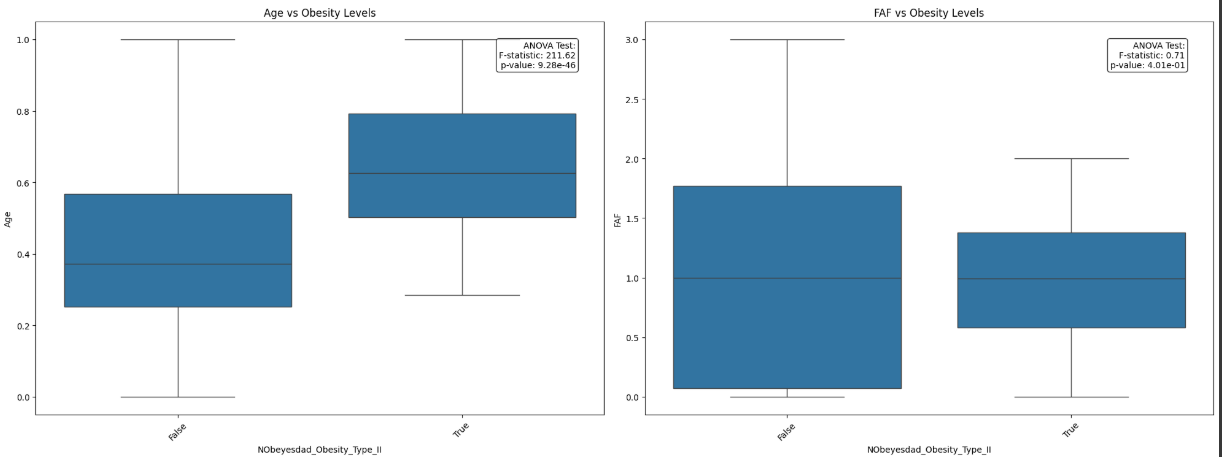


Figure 18Boxplots to show relationship between Age, Frequency of physical activity and NObeyesdad\_Obesity\_Type\_II

In the figure above the individuals in the obesity\_type\_II group majority of them were older in age than individuals who are not in the obesity\_type\_II group. A lesser number of individuals in the obesity\_type\_II group frequently engaged in physical activity than individuals who are not in the obesity\_type\_II group

## Task4: Correlation Analysis

The heatmap below:

Strongest Positive Correlations:

Height and Gender (0.62): Indicates a moderately strong relationship, suggesting taller heights are associated with one gender category

Height and Weight (0.46): Shows a moderate positive correlation, which is expected as taller people tend to weigh more

FAF and Height (0.29): Shows a weak positive correlation

Strongest Negative Correlations:

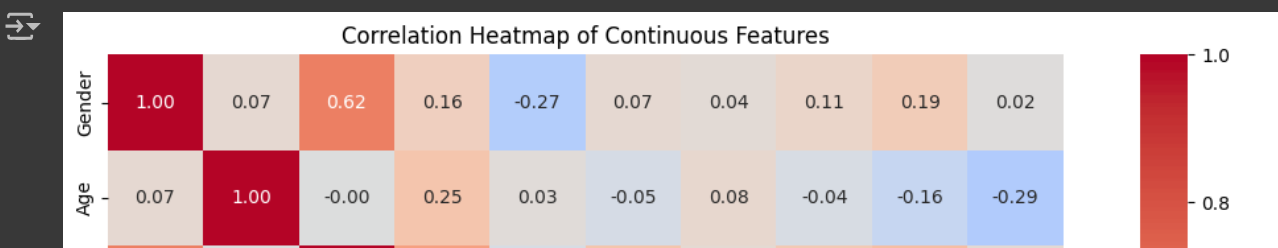
Age and TUE (-0.29): Shows a weak negative correlation

Gender and FCVC (-0.27): Shows a weak negative correlation

Age and FAF (-0.16): Shows a very weak negative correlation

**Overall summary**

The matrix is symmetrical Most variables show weak to moderate correlations with each other, suggesting these features are relatively independent. The strongest relationship is between Gender and Height, which aligns with typical biological patterns.



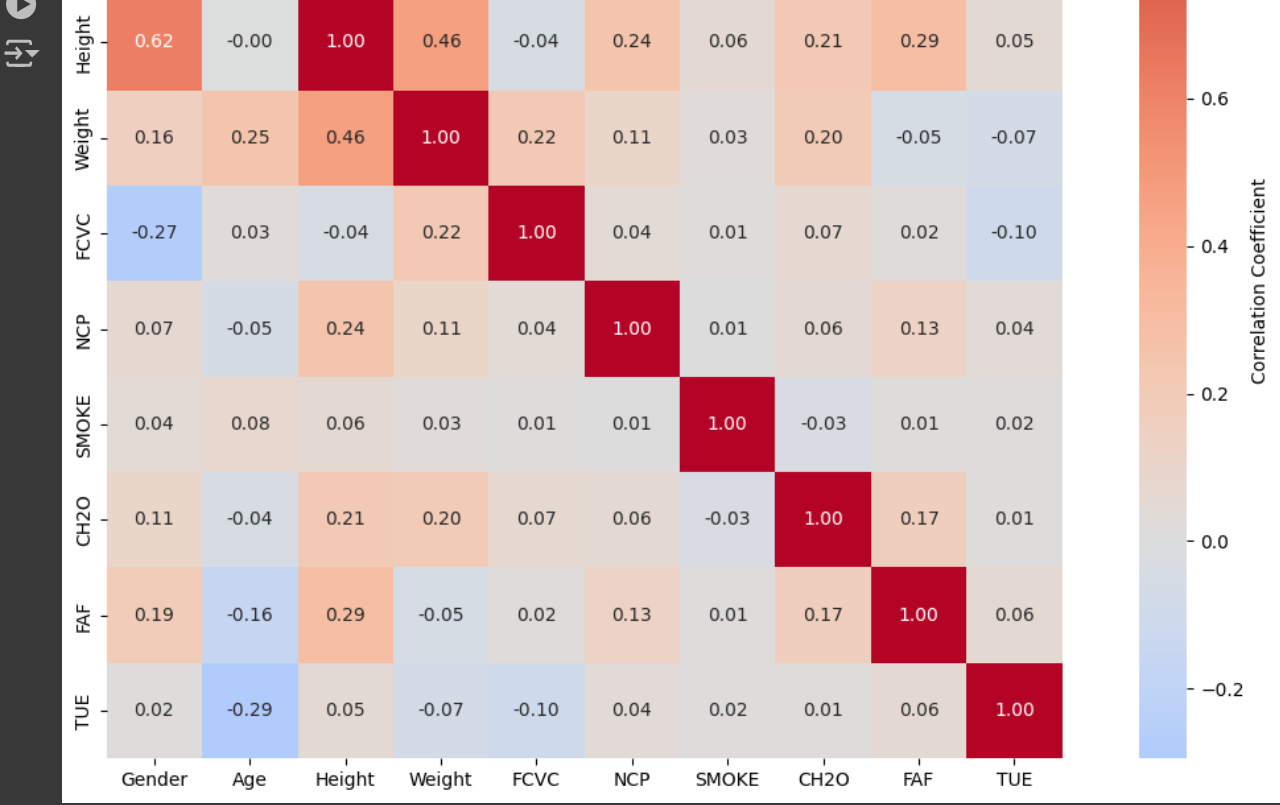
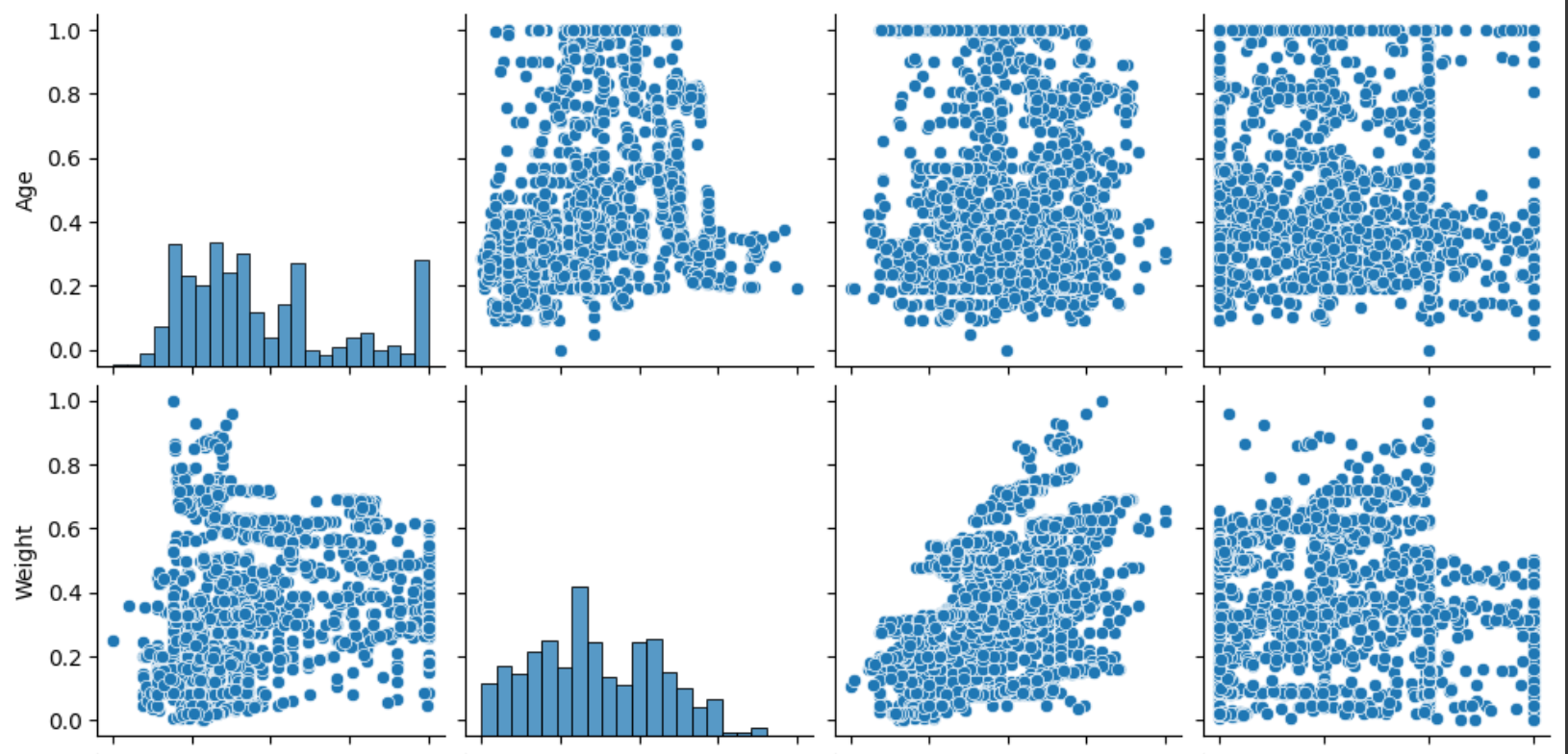


Figure 19Correlation Heatmap of Continuous Features

# Week3: Advanced Visualizations and Machine Learning

## Task 1: Advanced Visualizations



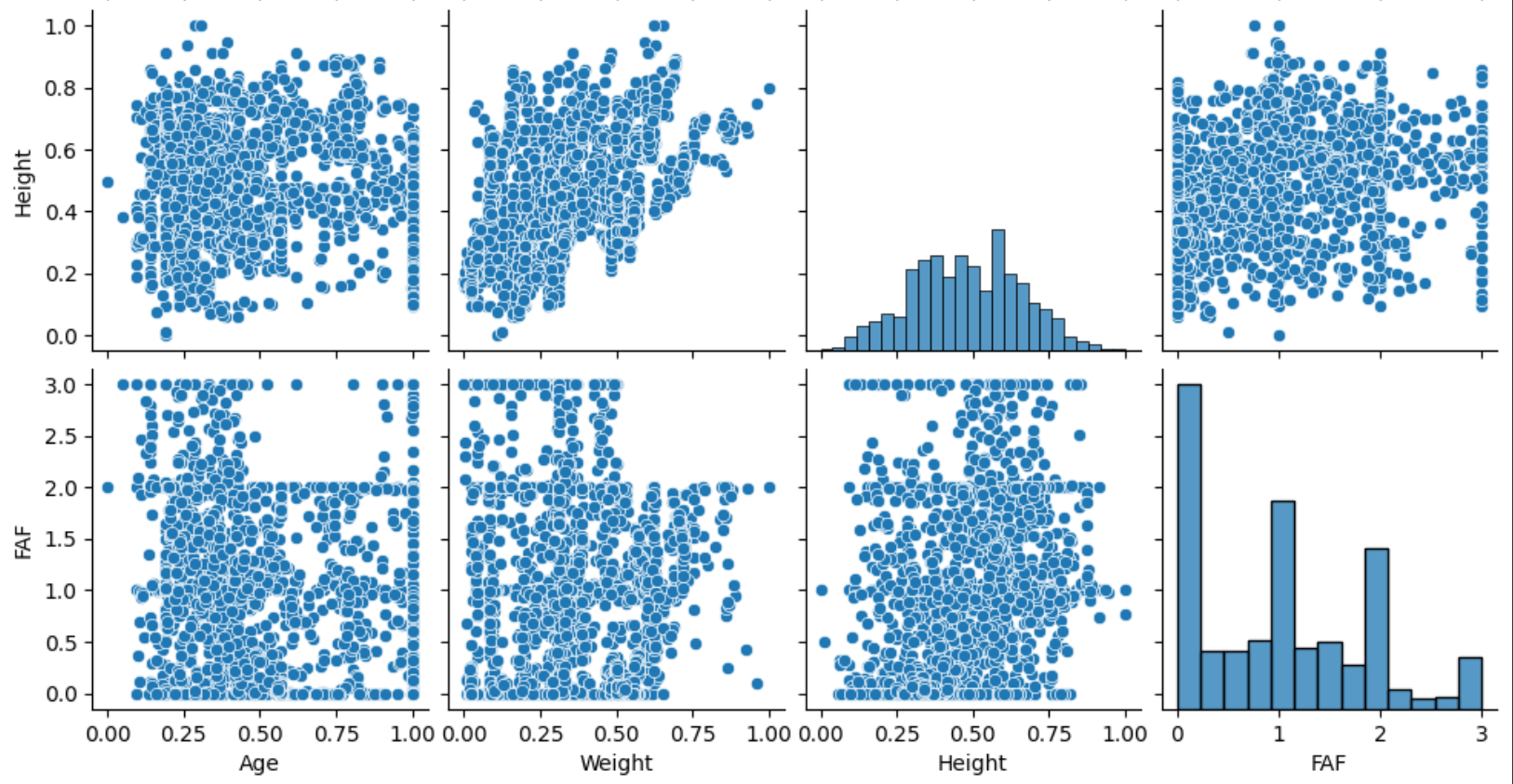


Figure 20Pair plots

The pairplot above shows relationships between Age, Weight, Height, and FAF (Physical Activity Frequency):

Diagonal (Distributions):

Age: Shows a multi-modal distribution with several peaks

Weight: Shows a slightly right-skewed distribution

Height: Appears roughly normal (bell-shaped) distribution

FAF: Shows a discrete distribution with clear peaks at certain values, suggesting this might be a rating scale

Relationships:

Age relationships:

Age vs Weight: Weak positive relationship with wide scatter

Age vs Height: No clear relationship, points fairly evenly distributed

Age vs FAF: Slight negative relationship, with younger ages showing more varied FAF values

Weight relationships:

Weight vs Height: Moderate positive correlation, as expected (taller people tend to weigh more)

Weight vs FAF: No strong pattern, though there's some clustering

Weight vs Age: Weak positive relationship

Height relationships:

Height vs FAF: Weak positive relationship

Height vs Age: No clear pattern

Height vs Weight: Moderate positive correlation (as noted above)

FAF (Physical Activity Frequency) relationships:

FAF vs Age: Slight negative trend, younger people showing more variety in activity levels

FAF vs Weight: No strong pattern

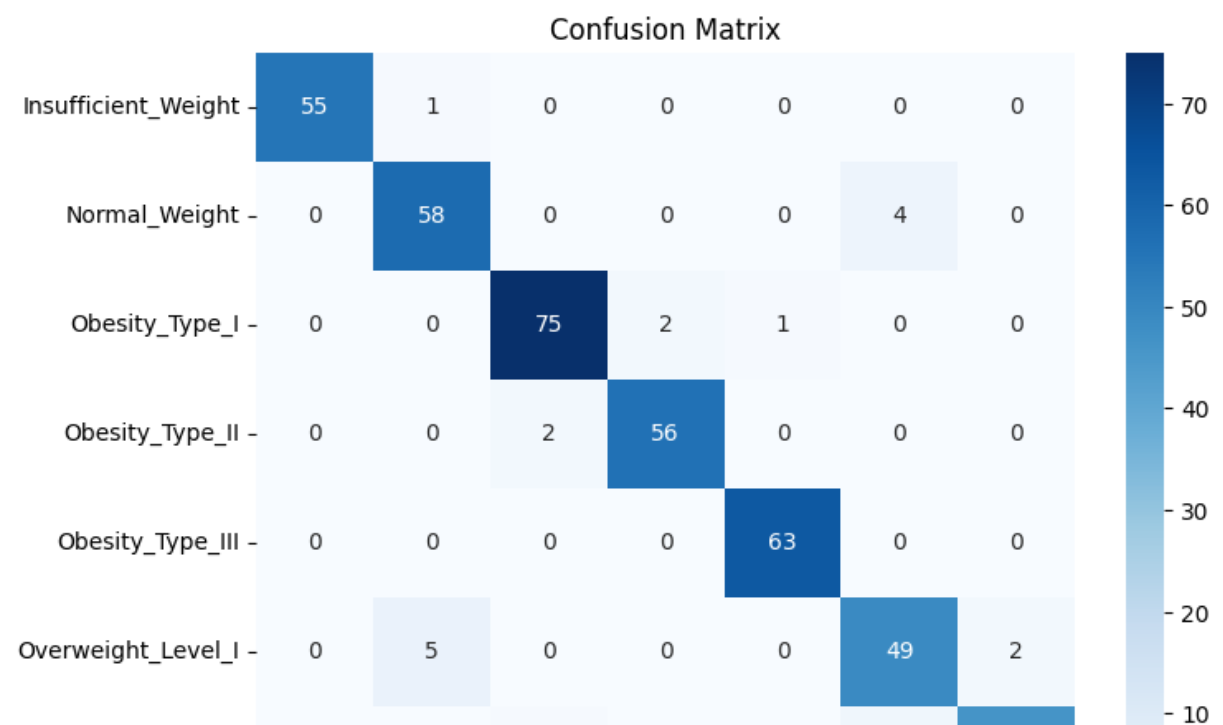
FAF vs Height: Weak positive relationship

**Overall summary**

The clearest relationship is between Height and Weight (positive correlation)

Most relationships show considerable scatter, indicating complex relationships

There appear to be some outliers, particularly in the Weight measurements



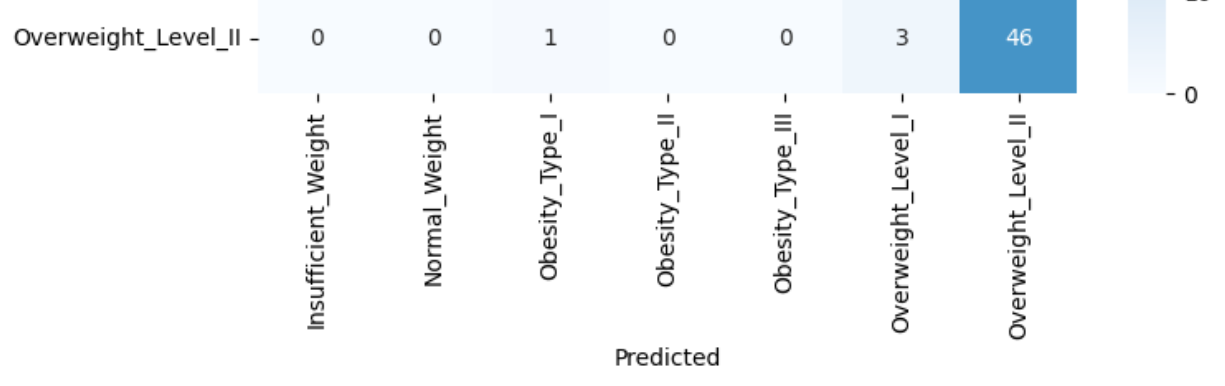


Figure 21Confusion matrix

The confusion matrix above shows the performance of a classification model for different weight categories:

True Positives (Diagonal):

Insufficient\_Weight: 55 correct predictions

Normal\_Weight: 58 correct predictions

Obesity\_Type\_I: 75 correct predictions (highest accuracy)

Obesity\_Type\_II: 56 correct predictions

Obesity\_Type\_III: 63 correct predictions

Overweight\_Level\_I: 49 correct predictions

Overweight\_Level\_II: 46 correct predictions

Misclassifications:

Between Normal\_Weight and Overweight\_Level\_I:

4 Normal\_Weight cases were predicted as Overweight\_Level\_I

5 Overweight\_Level\_I cases were predicted as Normal\_Weight

Between Obesity Types:

2 cases of Obesity\_Type\_I were predicted as Type\_II

1 case of Obesity\_Type\_I was predicted as Type\_III

Between Overweight Levels:

3 Overweight\_Level\_II cases were predicted as Level\_I

2 Overweight\_Level\_II cases were predicted as Level\_I

**Overall summary**

This appears to be a well-performing model with most predictions falling on the diagonal, indicating good accuracy across all weight categories. Obesity\_Type\_I has the highest number of correct predictions (75)

.

## Task 2: Feature Engineering and Scaling

The analysis of feature importance was conducted to determine which factors significantly

contribute to the model's predictions of obesity levels

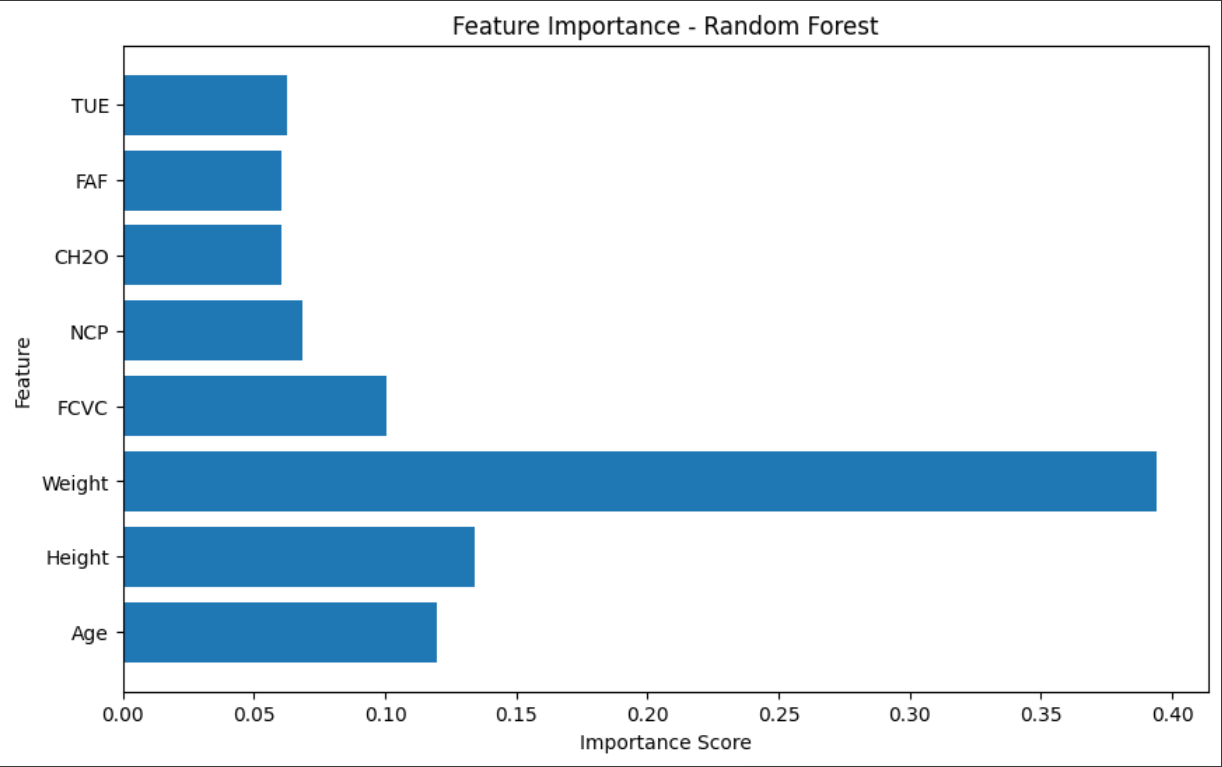


Figure 22feature importance plot

The feature importance plot above shows that weight is the most influential feature, dominating the chart with its significantly higher importance score compared to others.

Other features have lower importance scores, highlighting lesser impact on obesity levels

## Task 3: Train-Test Split

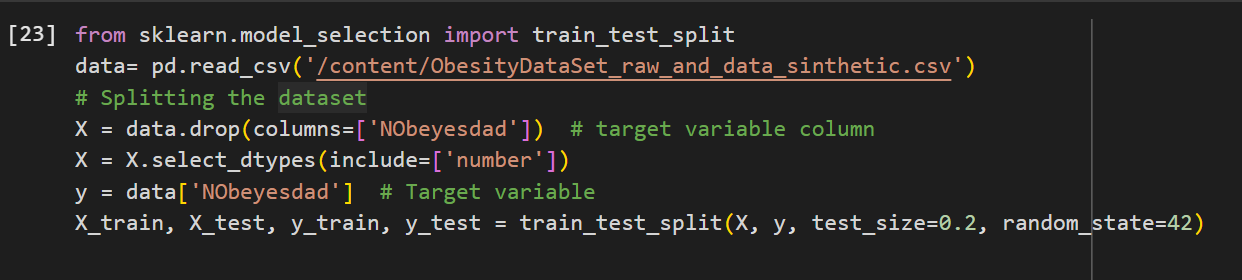


Figure 23Train-Test split

## Task4: Machine Learning Model Implementation

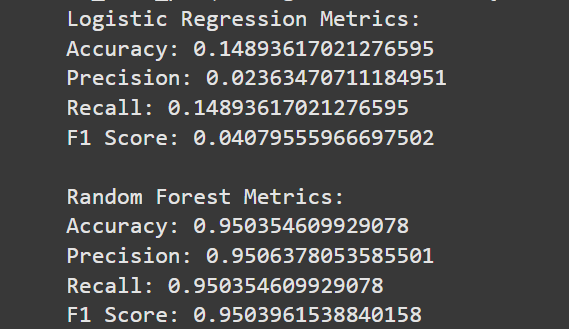


Figure 24Logistic regression and random forest metrics

## Task 5: Model Evaluation

**Logistic Regression Metrics:**

Accuracy: 0.1489

This is the percentage of correctly predicted labels out of all predictions. A value this low indicates the model is performing poorly, barely better than random guessing.

Precision: 0.0236

Precision measures how many of the positive predictions are actually correct. This low value indicates that the model is making many false-positive predictions (i.e., incorrectly predicting a label as positive when it is not).

Recall: 0.1489

Recall measures how many actual positive cases the model correctly identifies. The model only identifies ~14.89% of the true positive cases, showing poor sensitivity to the target class.

F1 Score: 0.0408

The F1 score is the harmonic mean of precision and recall. A low F1 score reflects the poor balance between precision and recall.

Conclusion: Logistic Regression is performing very poorly. This could be due to several reasons, such as insufficient feature scaling, high multicollinearity, an unsuitable model for the data, or imbalanced classes.

**Random Forest Metrics:**

Accuracy: 0.9504

The model is correctly predicting 95.04% of the cases, indicating strong overall performance.

Precision: 0.9506

A high precision value means that most positive predictions are correct, with very few false positives.

Recall: 0.9504

A high recall value means the model is correctly identifying the vast majority of true positives.

F1 Score: 0.9504

The F1 score is very high, indicating a strong balance between precision and recall.

Conclusion: Random Forest is performing exceptionally well, with high accuracy, precision, recall, and F1 score. This suggests that it is well-suited to this dataset and has effectively learned patterns in the data.

**Overall summary**

Logistic Regression:

Performs very poorly, likely due to its linear nature, sensitivity to feature scaling, or inability to handle complex relationships in the dataset.

Accuracy and F1 Score are too low for practical use.

Random Forest:

A non-linear model that works well with complex relationships and feature interactions.

Achieves high performance across all metrics, making it the better choice for this task.

# Conclusion

Obesity levels can be effectively predicted using eating habits, physical activity, and other health-related factors.

Random Forest was identified as the best model for this task due to its high performance across all evaluation metrics.

# Recommendations

Adopt Random Forest for Prediction:

Deploy the Random Forest model for practical applications in predicting obesity levels.

Ensure hyperparameter optimization to maintain its high performance.

Awareness Programs:

Leverage insights about obesity-related factors to design targeted health awareness campaigns.

Encourage healthier eating habits and increased physical activity, especially among younger demographics.

Future Improvements:

Explore ensemble methods or deep learning models to further enhance prediction accuracy.

Conduct longitudinal studies to examine how obesity levels change over time with lifestyle modifications.

# Github repository

https://github.com/vjarenga/Estimation-of-Obesity-Levels-Based-on-Eating-Habits-and-Physical-Condition.git