PROJECT REPORT

Estimation of Obesity Levels Based on Eating Habits and Physical Condition

By Oghenerukevwe Esemitodje

Introduction

The objective of this project is to analyze a dataset containing health and dietary information from individuals in Mexico, Peru, and Colombia. The aim is to estimate obesity levels based on physical condition and eating habits. The analysis was conducted using Python, using various machine learning techniques to extract meaningful insights from the data and predict obesity levels.

The purpose of this report is to walk through the entire analysis process, highlighting the methods used and the insights gained from the data.

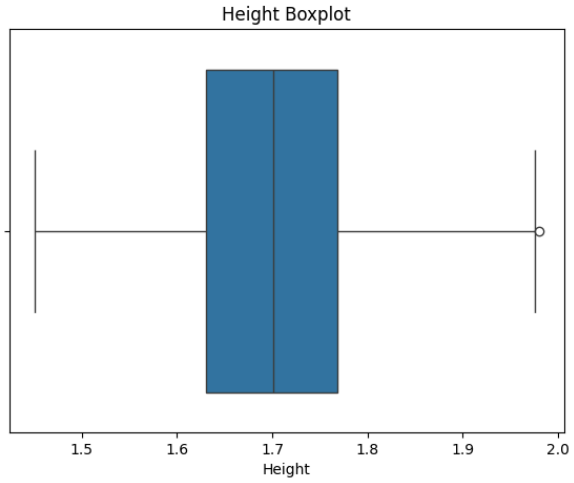
https://drive.google.com/file/d/1q21Y2_XVvxPpIVANtGzxpHwH2JLpiQLR/view?usp=sharing

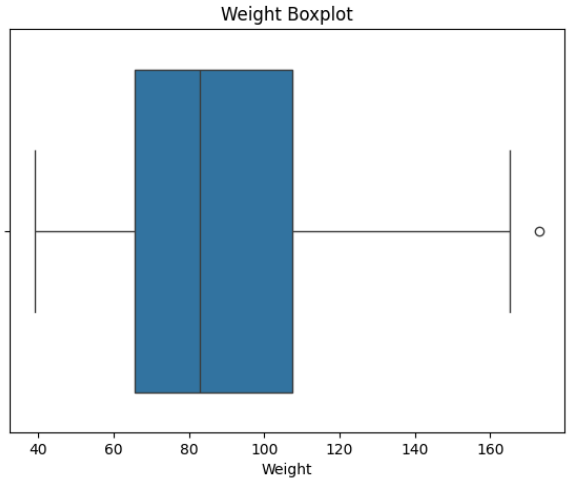
https://github.com/rukyese/obesity-prediction-project

WEEK 1: Dataset Description

Data Inspection and Cleaning

- Imported the dataset and inspected the structure.
- Removed duplicate records, reducing the data from **2111** to **2087** entries.
- Label encoding was used for binary features like **Gender**, **Smoke**, **family_history_with_overweight**, **FAVC** and **SCC**.
- One-hot encoding was applied to multi-class variables like **MTRANS**, **CALC**, **CAEC** and **NObeyesdad**.
- Outliers in continuous variables such as **Weight** and **Height** were detected using boxplots and managed by capping values using the interquartile range (IQR) method.
- MinMax scaling was applied to normalize the values of features such as **Age**, **FCVC**, **NCP**, **CH2O**, **FAF**, **TUE**, **Weight**, and Height to a scale of **0-1**.





WEEK 2: Exploratory Data Analysis (EDA)

Summary Statistics for Continuous Features

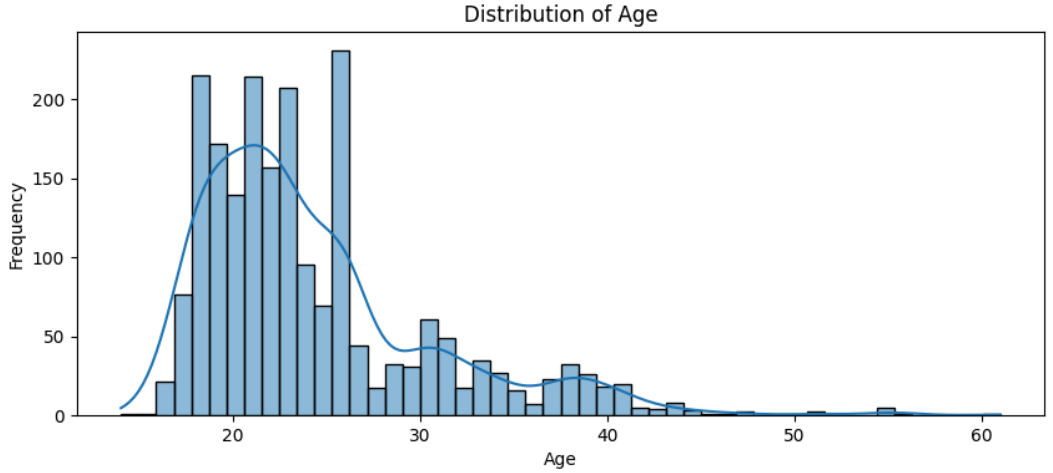
A summary of key continuous features (Age, Height, Weight, family_history_with_overweight, FCVC, NCP, CH2O, FAF, TUE) was computed, including mean, median, standard deviation, minimum, maximum, and quartiles.

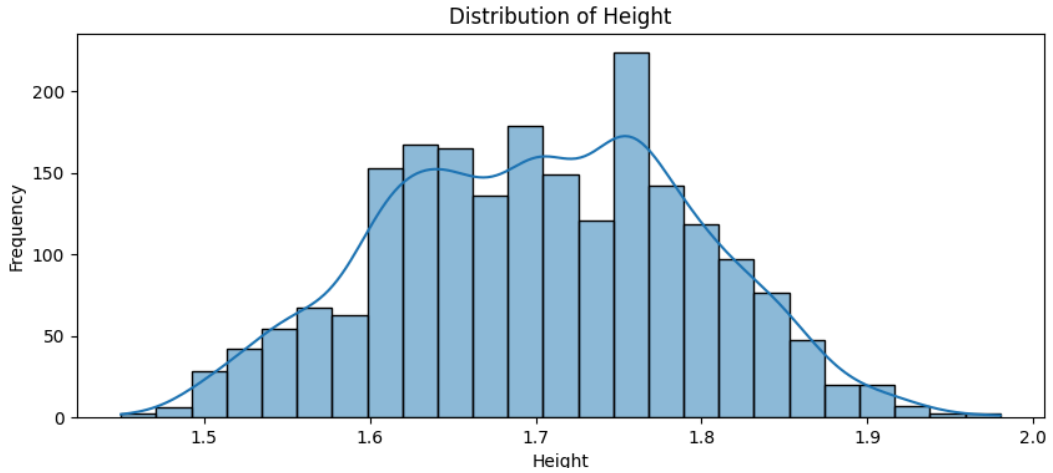
- **Gender**: Mean of **0.504**, indicating roughly equal distribution between male and female participants.
- **Age**: Mean of **24.35** years, with a range from **14** to **61** years, suggesting a focus on younger individuals.
- **Height**: Average height of **1.70** meters, ranging from **1.45** to **1.98** meters.
- **Weight**: Mean weight of **86.86** kg, with notable variance, ranging from **39** to **173** kg.
- **family_history_with_overweight**: Majority (**82.5%**) of individuals had a family history of being overweight.
- **FCVC (Frequency of Vegetable Consumption)**: Average score of **2.42** out of **3**, indicating moderate vegetable consumption.
- **NCP (Number of Main Meals)**: Most participants had approximately **3** meals per day.
- **CH2O (Water Consumption)**: Mean value of **2.00**, indicating most individuals drank around **2** liters per day.
- **FAF (Physical Activity Frequency)**: Mean of **1.01**, suggesting low physical activity frequency overall.
- **TUE (Time using Technology Devices)**: Mean of **0.66**, reflecting moderate use of technology.

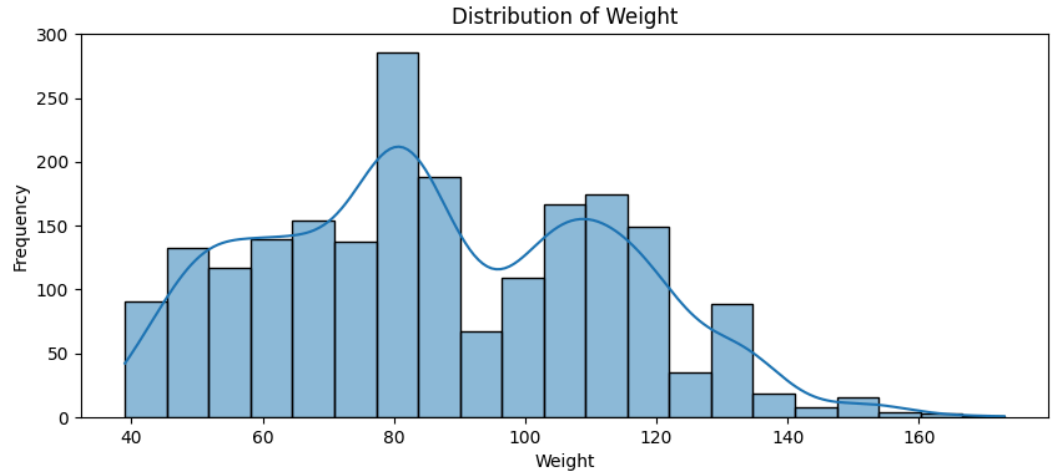
Distribution Analysis

- **Age Distribution**: Skewed towards younger individuals, with most between **18–25**. Few participants are over **30**, indicating a younger demographic bias.
- **Height Distribution**: Roughly normal, peaking at **1.7–1.8** meters. Slight variation at extremes may reflect gender or genetic diversity.

- **Weight Distribution**: Bimodal, with peaks at **70–80** kg and **90–100** kg. A right tail indicates some higher-weight outliers, likely linked to obesity.







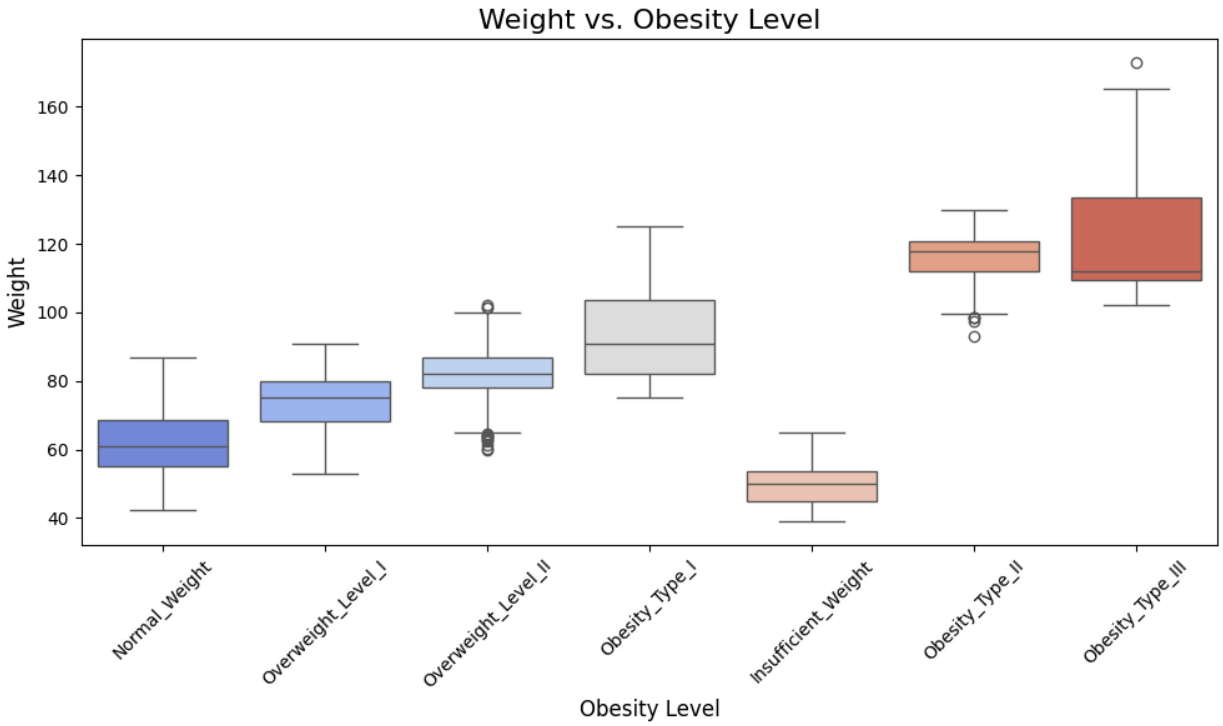
Relationship Exploration

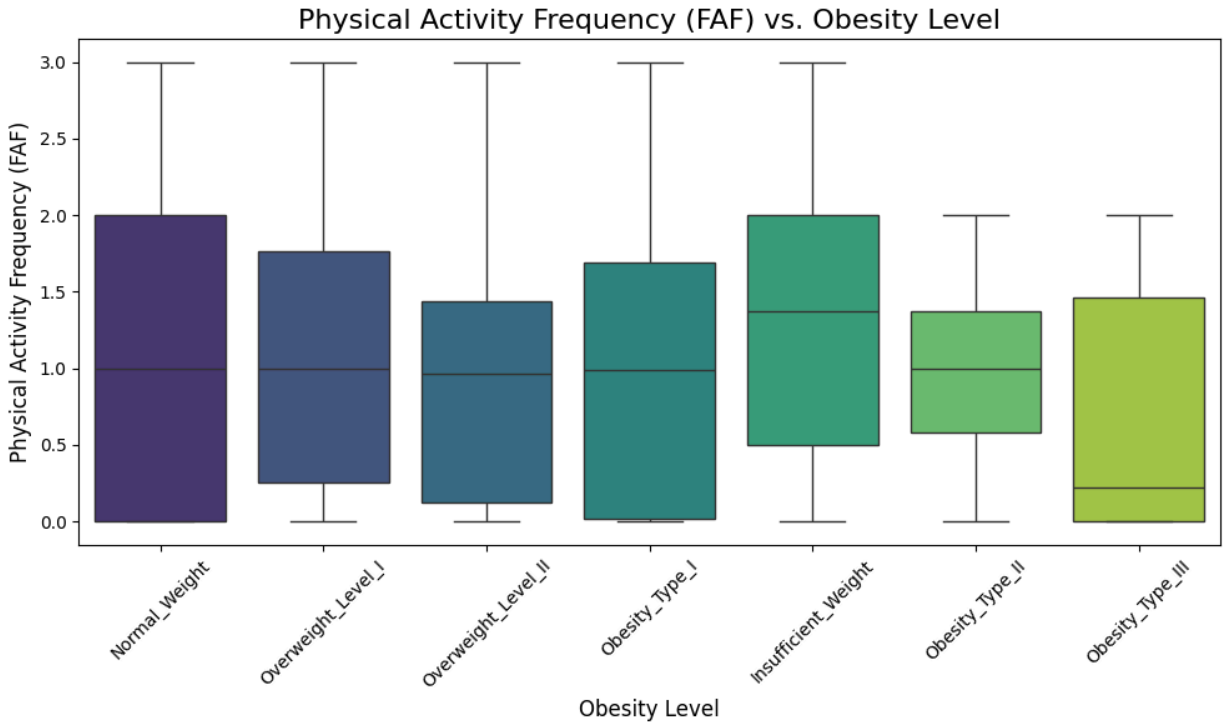
- **Weight vs. Obesity Level**
    - The boxplot shows a clear positive correlation between weight and obesity level.
    - As obesity levels progress from **Normal Weight** to **Obesity Type III**, there is a **noticeable increase in median weight**.
    - **Obesity Type III** has the **highest median weight** and the widest range, indicating significant variability among individuals in this category.
    - Individuals with **Normal Weight** have a **narrower range**, reflecting a more consistent weight pattern.
    - Outliers are present, particularly in **Overweight** Levels and **Obesity Type III**, suggesting some individuals fall significantly outside the expected weight distribution for their category.

- **Physical Activity Frequency (FAF) vs. Obesity Level**
    - The inverse relationship is apparent, as higher obesity levels correlate with **lower median physical activity frequency**.
    - **Normal Weight** individuals exhibit the **highest median FAF**, showing greater engagement in regular physical activity.
    - Categories such as **Obesity Type II** and **Obesity Type III** show **lower median FAF**, with many individuals having minimal or no physical activity.
    - The variability in physical activity decreases as obesity levels increase, suggesting **fewer individuals engage in regular physical activity** at higher obesity levels.

Conclusions from Relationships:

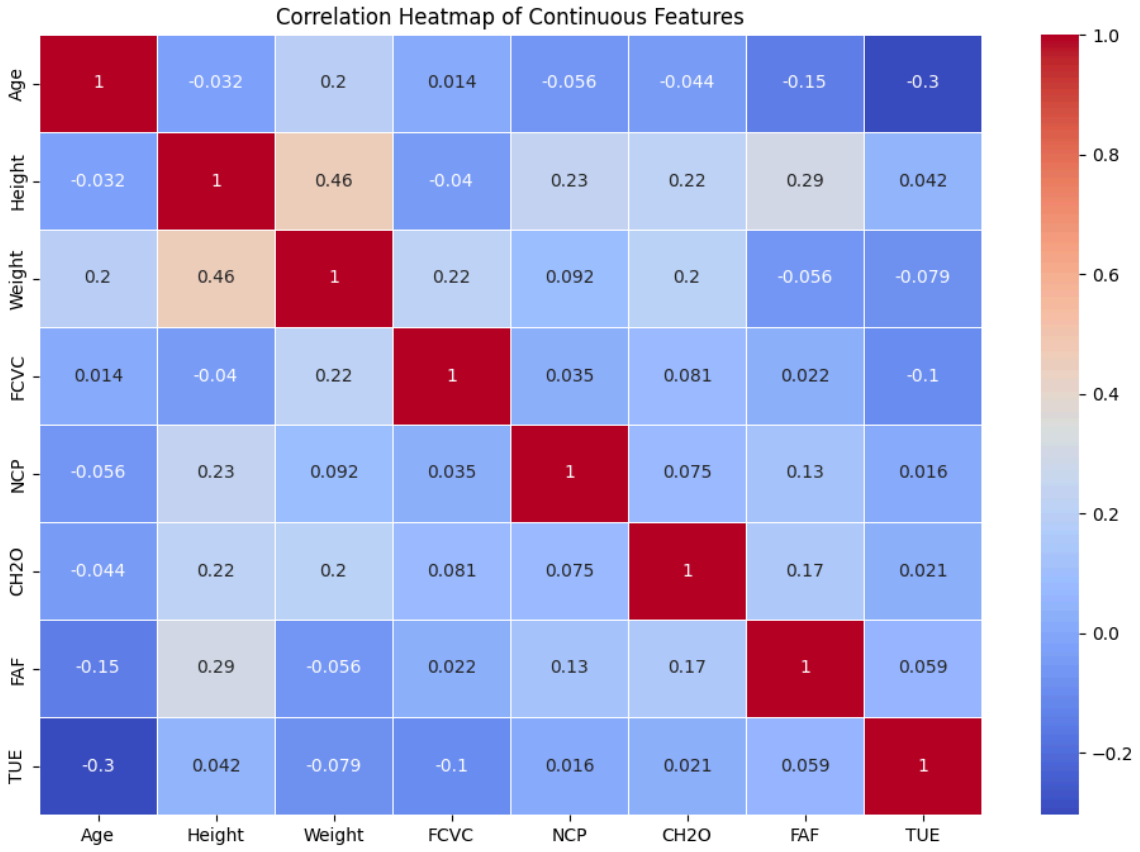
- The **progressive increase in weight** aligns with higher obesity levels, emphasizing the significance of weight management in combating obesity.
- The **inverse relationship between FAF and obesity levels** underscores the importance of regular physical activity in maintaining healthy weight and preventing obesity progression.


Weight vs. Obesity Level
Weight
160
140
120
100
80
60
40
Normal_Weight
Overweight_Level_I
Overweight_Level_II
Obesity_Type_I
Insufficient_Weight
Obesity_Type_II
Obesity_Type_III
Obesity Level


Physical Activity Frequency (FAF) vs. Obesity Level
Physical Activity Frequency (FAF)
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Normal_Weight
Overweight_Level_I
Overweight_Level_II
Obesity_Type_I
Insufficient_Weight
Obesity_Type_II
Obesity_Type_III
Obesity Level

Correlation Analysis

- A correlation heatmap was generated to explore relationships between variables.
- **Height** and **Weight** had a moderate positive correlation (**0.46**), implying that taller individuals tend to weigh more, although not always.
- **Age** showed a weak correlation with **Weight** and no significant correlation with **Height**, suggesting these variables are mostly independent.



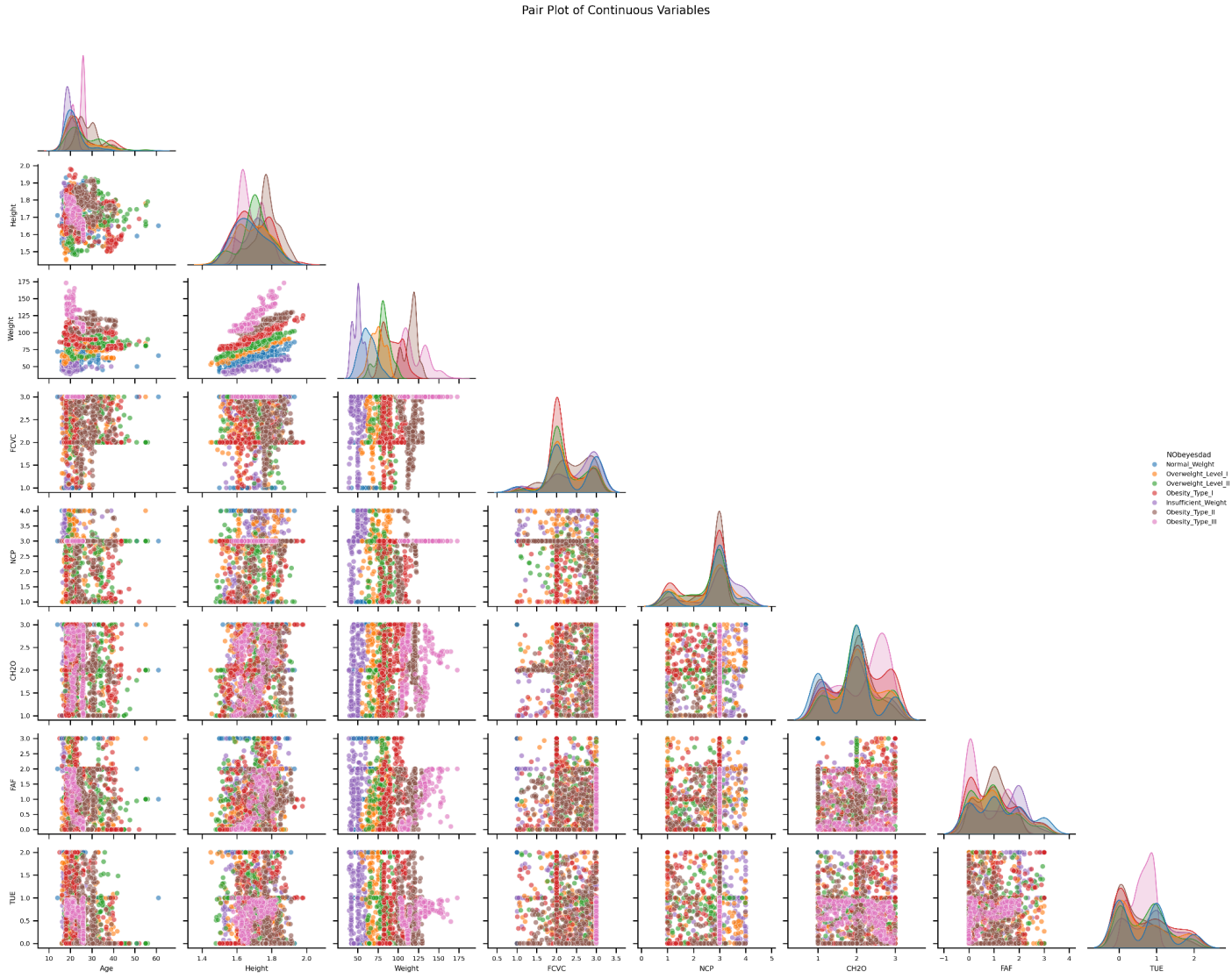
WEEK 3: Advanced Visualizations and Machine Learning

Pair Plot

The pair plot reveals several key relationships among the continuous variables:

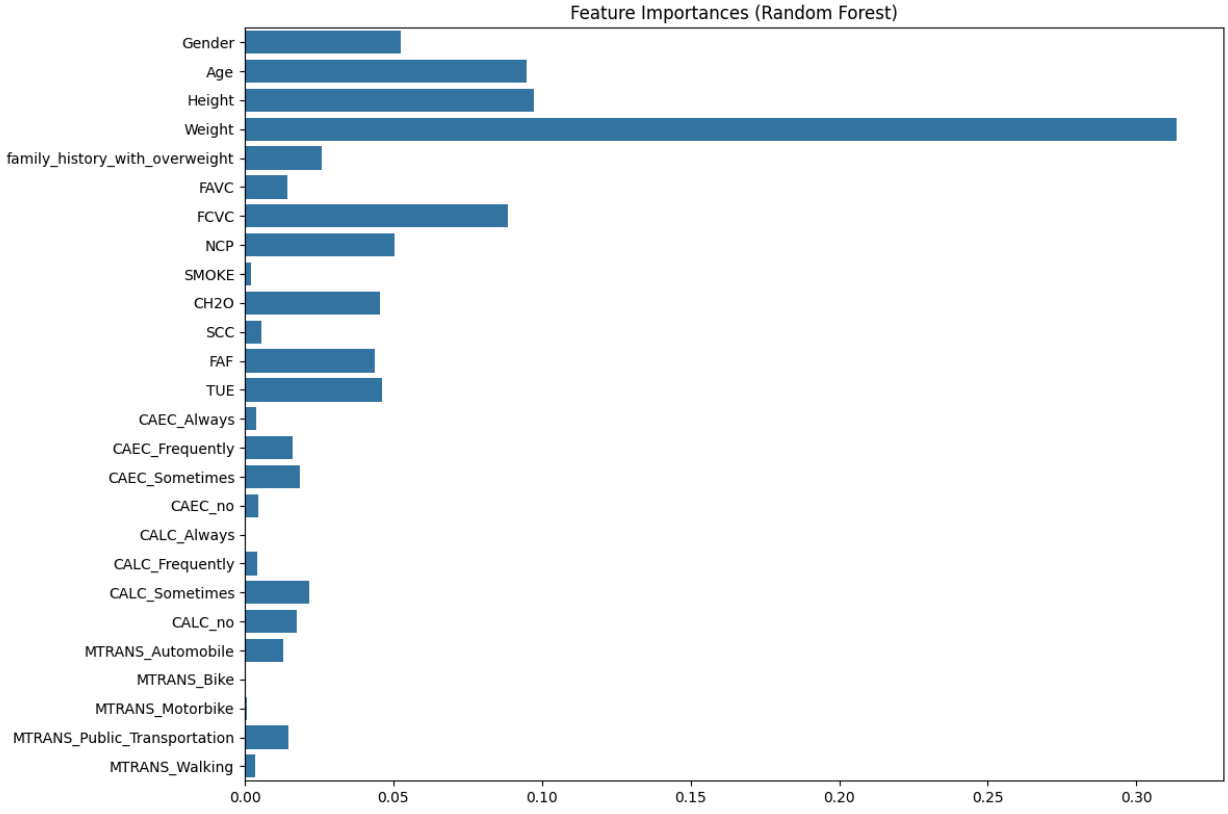
1. **Height and Weight:** A positive linear correlation is observed, indicating that taller individuals tend to weigh more, although there are some outliers.
2. **Physical Activity Frequency (FAF) and Weight:** An inverse trend is noted, suggesting that higher physical activity is associated with lower weight, aligning with its role in obesity prevention.
3. **Age:** No significant correlations are observed with weight, height, or other variables, suggesting age is largely independent of these factors in this dataset.
4. **Clusters and Outliers:** Distinct clusters in weight and height likely correspond to different obesity categories, while outliers highlight unique cases, such as individuals with atypical body compositions or behaviors.

These insights underscore the importance of height, weight, and physical activity as key factors in understanding obesity levels.


Pair Plot of Continuous Variables
NObeyesdad
Normal_Weight
Overweight_Level_I
Overweight_Level_II
Obesity_Type_I
Insufficient_Weight
Obesity_Type_II
Obesity_Type_III
Age
Height
Weight
FCVC
NCP
CH2O
FAF
TUE

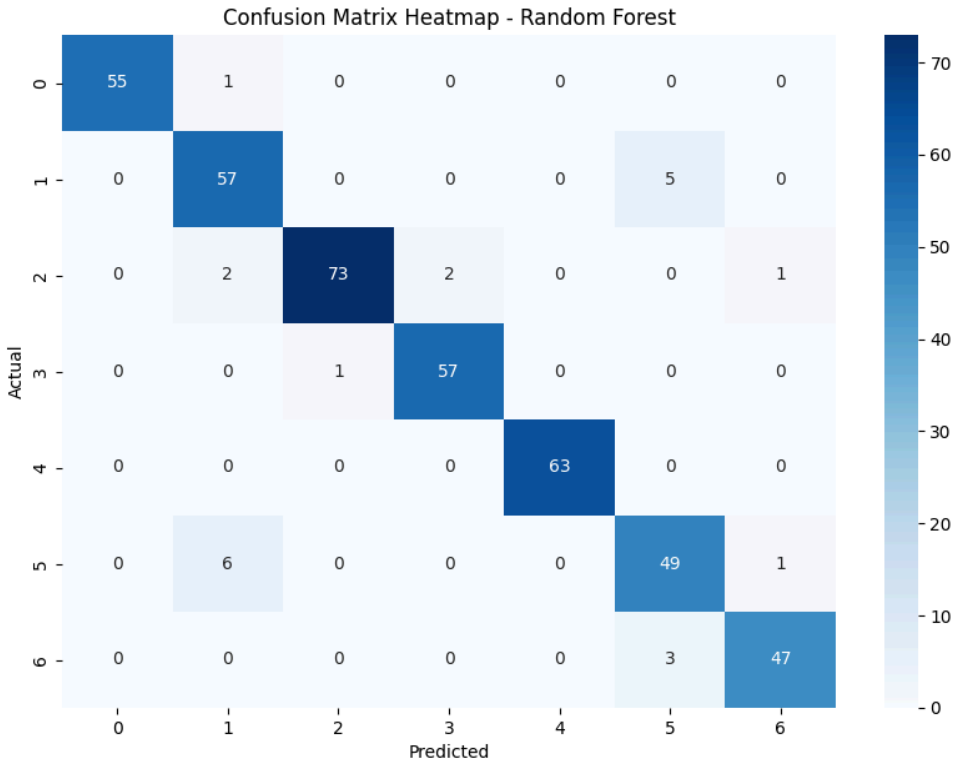
Feature Importance Plot

- **Weight** was found to be the most crucial feature, followed by **Height** and **Age**.
- Other features like **vegetable consumption (FCVC)** and **family history of overweight** were also important, but factors like **Smoking** had little impact.



Confusion Matrix Heatmap (Random Forest)

- The confusion matrix highlighted that the **Random Forest model** was highly effective, with most values concentrated along the diagonal, indicating correct predictions.



WEEK 4: Model Evaluation and Reporting

Machine Learning Model Implementation

- Implemented two machine learning models: **Logistic Regression** and **Random Forest Classifier**.
- The dataset was divided into **training and testing sets** with an 80-20 split, and **StandardScaler** was applied for feature normalization.

Logistic Regression

- **Overall Accuracy:** 86%
- **Performance Across Categories:**
    - **Class 0 (Normal Weight):** High precision (**87%**) and strong recall (**92%**) result in an F1-score of **0.89**, indicating reliable classification.
    - **Class 1 (Overweight Level I):** Moderate precision (**83%**) and lower recall (**74%**) lead to an F1-score of **0.78**, showing challenges in identifying all cases correctly.
    - **Class 4 (Obesity Type II):** Achieves perfect precision and recall (**1.00**), demonstrating excellent classification performance.
    - **Class 6 (Severe Obesity):** Precision (**66%**) and recall (**78%**) suggest room for improvement in identifying severe cases.
- **Weaknesses:**
    - Struggles with overlapping categories, such as **1 (Overweight Level I)** and **5 (Obesity Type III)**, causing reduced recall and F1-scores.
    - Slightly weaker performance in minority classes, such as **6 (Severe Obesity)**.

Random Forest

- **Overall Accuracy:** 95%
- **Performance Across Categories:**
    - **Class 0 (Normal Weight):** Excellent precision (**98%**) and good recall (**90%**) yield an F1-score of **0.94**.

    - **Class 1 (Overweight Level I):** Improved performance with precision (**82%**) and high recall (**95%**), resulting in an F1-score of **0.88**.
    - **Class 4 (Obesity Type II):** Perfect precision, recall, and F1-score (**1.00**), highlighting the model's exceptional accuracy.
    - **Class 6 (Severe Obesity):** Strong performance with precision (**92%**) and recall (**96%**), resulting in an F1-score of **0.94**.
- **Strengths:**
    - Handles overlapping categories and non-linear relationships effectively, outperforming Logistic Regression in all metrics.
    - High precision and recall across all obesity levels, especially for distinct classes like **4 (Obesity Type II)** and **6 (Severe Obesity)**.
- **Weaknesses:**
    - Slightly reduced precision and recall in categories with smaller sample sizes, such as **1 (Overweight Level I)** and **5 (Overweight Level II)**.

Key Insights

- The **Random Forest model outperformed Logistic Regression** in all metrics, achieving higher accuracy, precision, recall, and F1-scores across all obesity categories.
- **Class Disparities**: Both models performed better for distinct categories (e.g., Obesity Type II) but struggled slightly with categories that may have overlapping features (e.g., Overweight Level I and Overweight Level II).
- **Conclusion**: Random Forest is the preferred model due to its superior ability to handle non-linear relationships and high performance across all obesity levels.

PROJECT SUMMARY

1. **Dataset**: The dataset includes various health and lifestyle attributes such as age, weight, height, frequency of vegetable consumption (FCVC), number of main meals per day (NCP), water intake (CH2O), physical activity frequency (FAF), and time spent using technology (TUE).
2. **Data Preprocessing**: To ensure consistency, features were standardized using a scaler, and the target variable (obesity levels) was converted into a format suitable for machine learning models.
3. **Exploratory Data Analysis (EDA)**: Visualizations like pair plots and heatmaps were used to explore relationships between different features. These plots provided valuable insights into patterns and correlations in the data.
4. **Feature Importance**: Among all features, **weight** emerged as the most critical factor in predicting obesity levels, followed by height and physical activity.
5. **Model Building**: Two machine learning models were developed—Logistic Regression and Random Forest. Both were trained and tested on the dataset to evaluate their performance.
6. **Evaluation**: While both models performed well, the Random Forest model stood out with its ability to handle complex relationships in the data, delivering higher accuracy and better overall results.

Insights:

- **Weight** is the most significant factor influencing obesity levels, with a clear progression in weight as obesity levels increase.
- The **Random Forest model's** strong performance highlights the presence of non-linear relationships between the features.
- The high accuracy of both models confirms that the chosen features are effective predictors of obesity levels.