Estimation of Obesity Levels Based on Habits

Introduction

The goal of this project is to estimate obesity levels based on physical conditions and eating habits using a dataset from individuals in Mexico, Peru, and Colombia. Through this project, various data preprocessing steps, exploratory data analysis (EDA), and machine learning techniques were employed to analyze health and dietary data, with the final objective of building a predictive model for obesity levels.

Key Objectives

- Clean and preprocess the dataset.
- Conduct EDA and visualize relationships between features.
- Build machine learning models to estimate obesity levels.
- Evaluate the performance of the models and interpret the results.

Dataset Description

The dataset comprises 17 features and 2,111 records. The target variable, NObeyesdad (Obesity Level), categorizes individuals into seven obesity levels:

- Insufficient Weight
- Normal Weight
- Overweight Level I
- Overweight Level II
- Obesity Type I
- Obesity Type II
- Obesity Type III

Key Features:

- Demographic and Physical Attributes: Gender, Age, Height, Weight
- Dietary Habits: Frequency of high-calorie food consumption (FAVC), frequency of vegetable consumption (FCVC), number of main meals (NCP), alcohol consumption (CALC), and food consumption between meals (CAEC).
- Lifestyle Factors: Smoking (SMOKE), physical activity frequency (FAF), water consumption (CH2O), and mode of transportation (MTRANS).
- Family and Medical History: Family history with overweight (Binary) and calorie monitoring (SCC).

Data Preprocessing

Data Importing and Inspection

The dataset was imported and inspected for missing values, incorrect data types, and outliers. There were no missing values; however, categorical variables needed to be encoded, and continuous variables required normalization.

Data Type Conversion and Encoding

- Binary variables such as *Gender*, *SMOKE*, and *family_history_with_overweight* were label-encoded.
- Multi-class variables such as *MTRANS* (mode of transportation) and the target variable *NObeyesdad* were one-hot encoded.

Outlier Detection and Handling

Outliers in continuous features like *Weight* and *Height* were detected using box plots. These outliers were capped at an appropriate range.

Normalization

Continuous variables such as *Age*, *Weight*, *Height*, *FAF*, *CH2O* and *TUE* were normalized using MinMax scaling to ensure they were on the same scale for machine learning models.

Exploratory Data Analysis (EDA)

Summary Statistics

Summary statistics were computed for continuous features such as *Age*, *Weight*, and *Height*. These statistics provided insights into the central tendencies and spread of the data.

Below is a reference table that appreciates the insights of the normalized data

Table 1

Normalized Values	Age	Height	Weight
0.0	14.0	1.450	39.0
0.1	18.7	1.503	52.4
0.2	23.4	1.556	65.8
0.3	28.1	1.609	79.2
0.4	32.8	1.662	92.6
0.5	37.5	1.715	106.0
0.6	42.2	1.768	119.4

0.7	46.9	1.821	132.8
0.8	51.6	1.874	146.2
0.9	56.3	1.927	159.6
1.0	61.0	1.980	173.0

With reference to the able above, the image below displays the statistical summary of the stated variables.

	Age	Weight	Height
count	2111.000000	2111.000000	2111.000000
mean	0.219417	0.411229	0.492595
std	0.135021	0.240426	0.226706
min	0.000000	0.000000	0.000000
25%	0.126536	0.216677	0.317473
50%	0.186764	0.378461	0.489903
75%	0.255319	0.603975	0.656135
max	1.000000	1.000000	1.000000

Figure 1. Statistical Summary

Distribution Analysis

Histograms and Kernel Density Estimate (KDE) plots were used to understand the distribution of key variables. The image below shows the distribution graph:

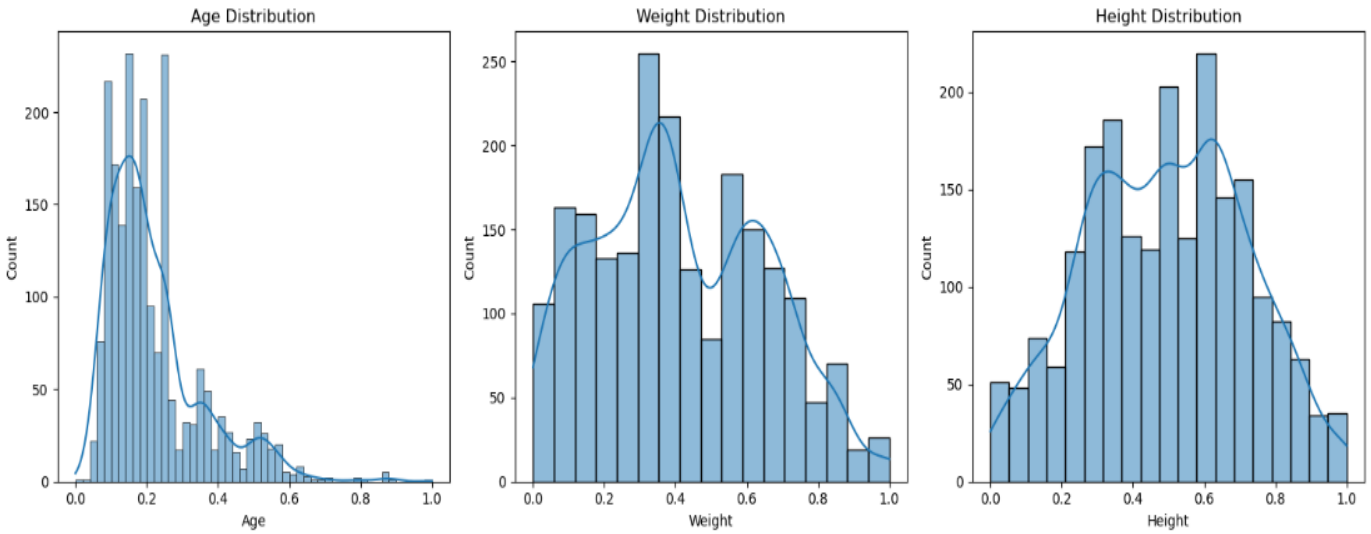

Figure 2. Histogram & KDE

Based on the histograms and KDE plots:

Age Distribution:

- The distribution appears to be slightly skewed to the right.
- Most individuals are in the age range between 15 and 30.
- There is a noticeable peak around the age of 20-25

Weight Distribution:

- The distribution is fairly symmetric with a peak between 0.4 - 0.6 range (after normalization)
- It indicates a majority of individuals are in the average weight range.

Height Distribution:

- The distribution is also fairly symmetric with a peak between 0.4 - 0.6 range (after normalization)
- It indicates most individuals have an average height.

In summary, the distributions of age, weight, and height are all relatively normal, with a tendency towards a slight right skew for age.

Relationship Exploration

Box plots were used to explore relationships between continuous variables (*Weight*, *FAF*) and the target variable *NObeyesdad*. It was observed that individuals with higher physical activity tended to have lower obesity levels.

For more clarity of the insight of the box plot, the table below shows the corresponding obesity levels to the one-hot encoding of the NObeyesdad to multiclass variables.

Table 2

NObeyesdad_0	Insufficient weight
NObeyesdad_1	Normal weight
NObeyesdad_2	Obesity Type I
NObeyesdad_3	Obesity Type II
NObeyesdad_4	Obesity Type III
NObeyesdad_5	Overweight level I
NObeyesdad_6	Overweight level II

The box plots explore the relationship between *Weight* and *FAF* (Frequency of Physical Activity) with respect to different categories of *NObeyesdad* (Obesity Level).

These box plots show the distribution of *Weight* and *FAF* for the classified individuals. By observing the median, quartiles, and potential outliers for both *Weight* and *FAF* within this group, we conclude that, if there's a significant difference in the median or the spread of the boxes between different *NObeyesdad* categories, it suggests that there's a relationship between obesity levels and *Weight* or *FAF*. The visualizations below show the relationships between obesity levels and *Weight* or *FAF*.

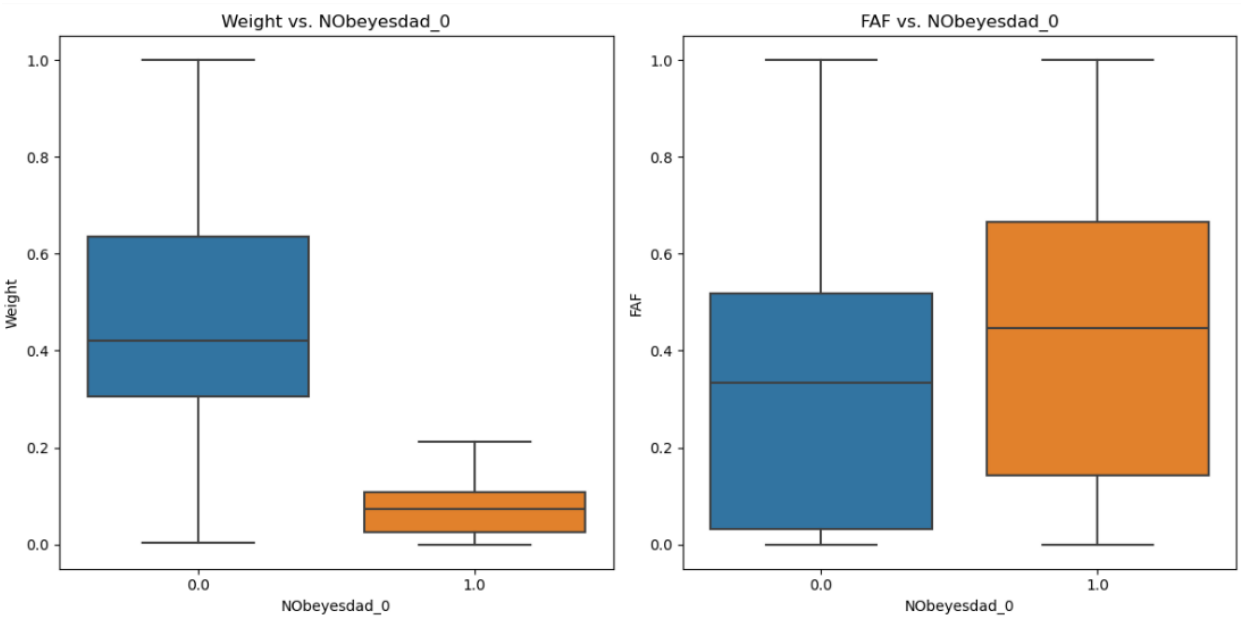

Figure 3. Boxplot of Insufficient weight

Figure 3 indicates that there is a relationship between the weight, and insufficient weight obesity level and also a relationship between the frequency of physical activities (FAF) and the insufficient weight obesity levels of individuals.

- Individuals with insufficient weight obesity level category are associated with lower weight.
- Individuals with insufficient weight obesity level category are associated higher frequency of physical activities

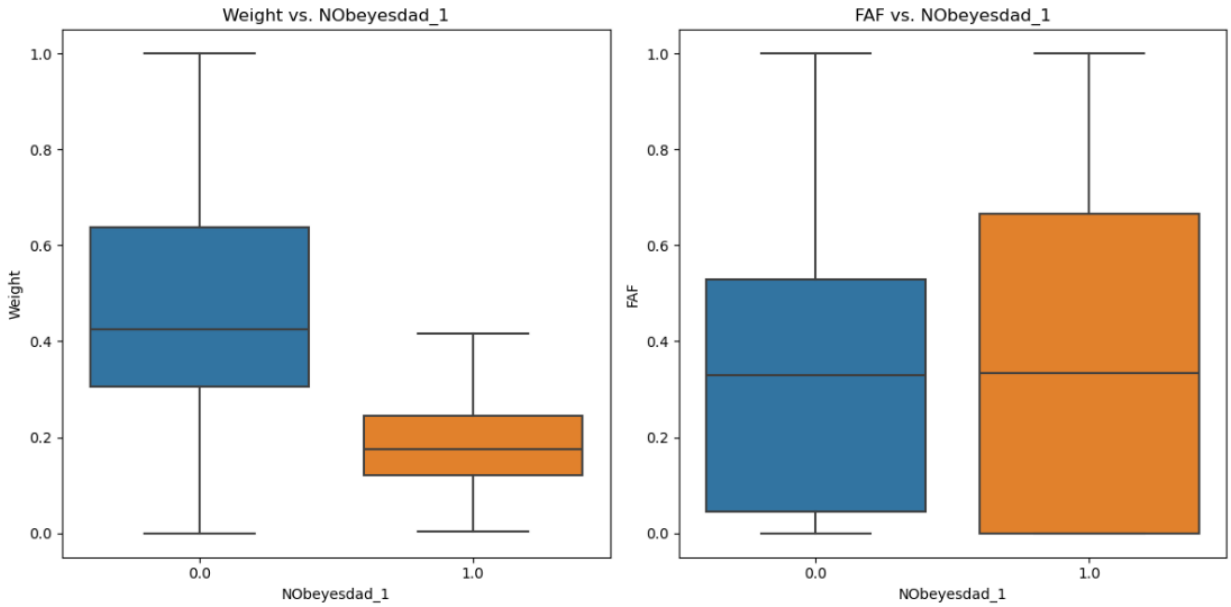

Figure 4. Normal weight box plot

Figure 4 indicates that there is a relationship between the weight and normal weight obesity level and no significant relationship between the frequency of physical activities (FAF) and the normal weight obesity levels of individuals

- Individuals with normal weight obesity level category are generally associated with lower weight as compared to the other obesity levels.

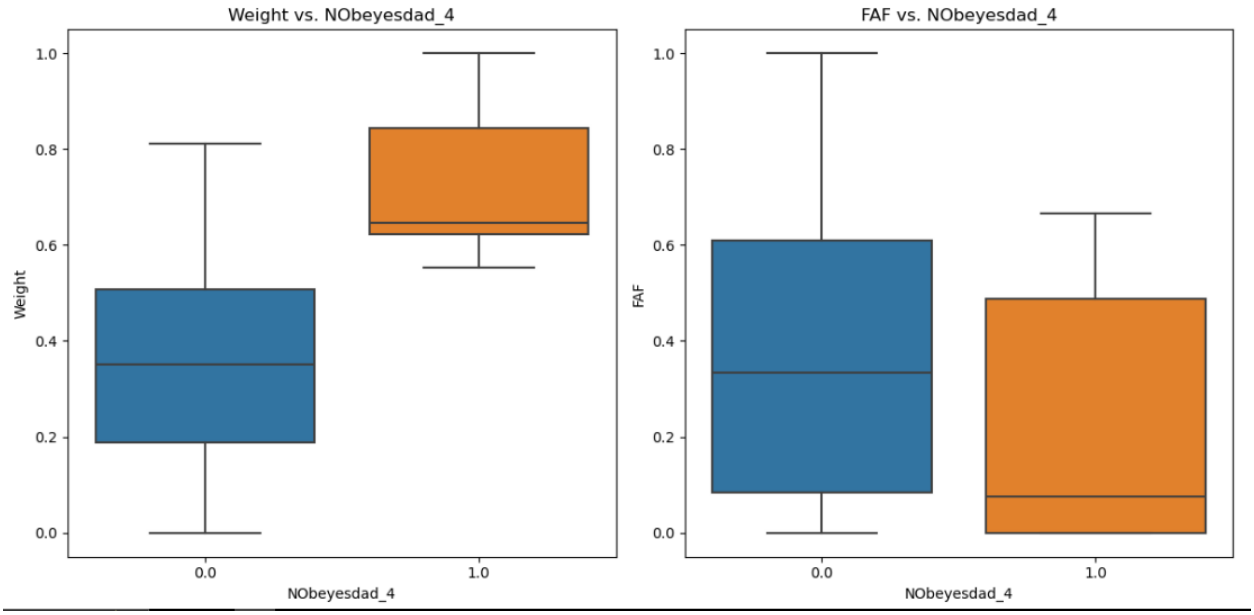

Figure 5. Obesity type III

Figure 5 indicates that there is a relationship between the weight and obesity Type III level and also a relationship between the frequency of physical activities (FAF) and the obesity Type III levels of individuals.

- Individuals with obesity Type III level category are associated with much higher weights as compared to other categories.
- Individuals with obesity Type III level category are associated very low frequency of physical activities.

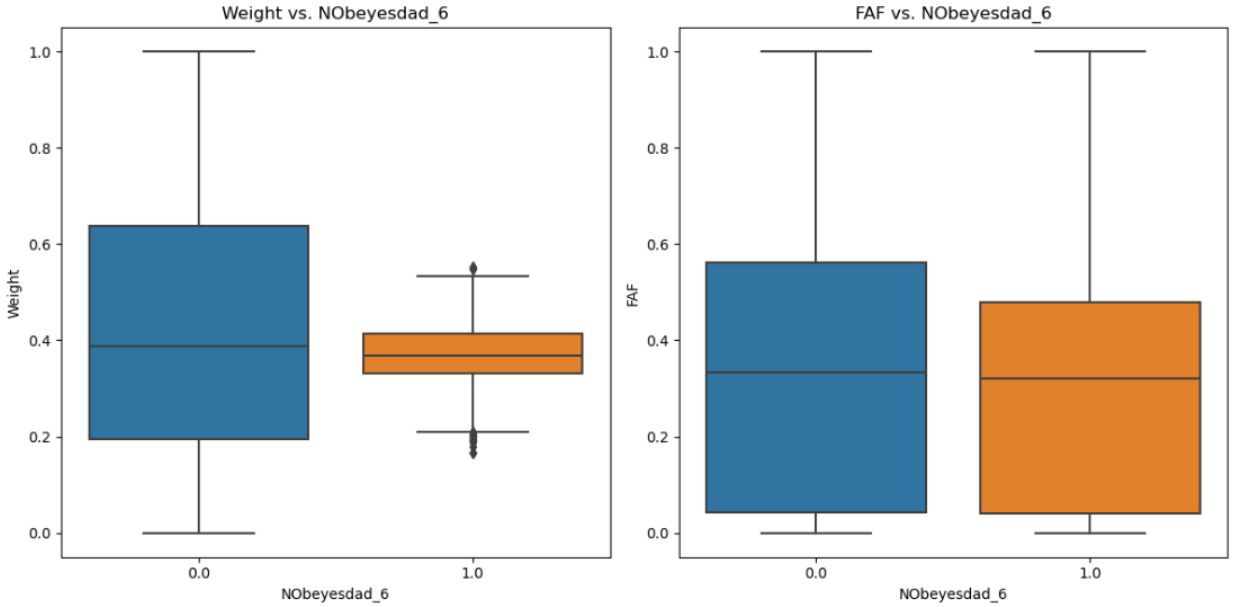

Figure 6. Overweight II

Figure 6 indicates that there is no recognizable relationship between the weight and overweight II level and also no relationship between the frequency of physical activities (FAF) and the overweight II levels of individuals.

Correlation Analysis

A correlation heatmap was generated to explore relationships between continuous variables such as *Age*, *Height*, and *Weight*. The figure below displays the heatmap of the correlations

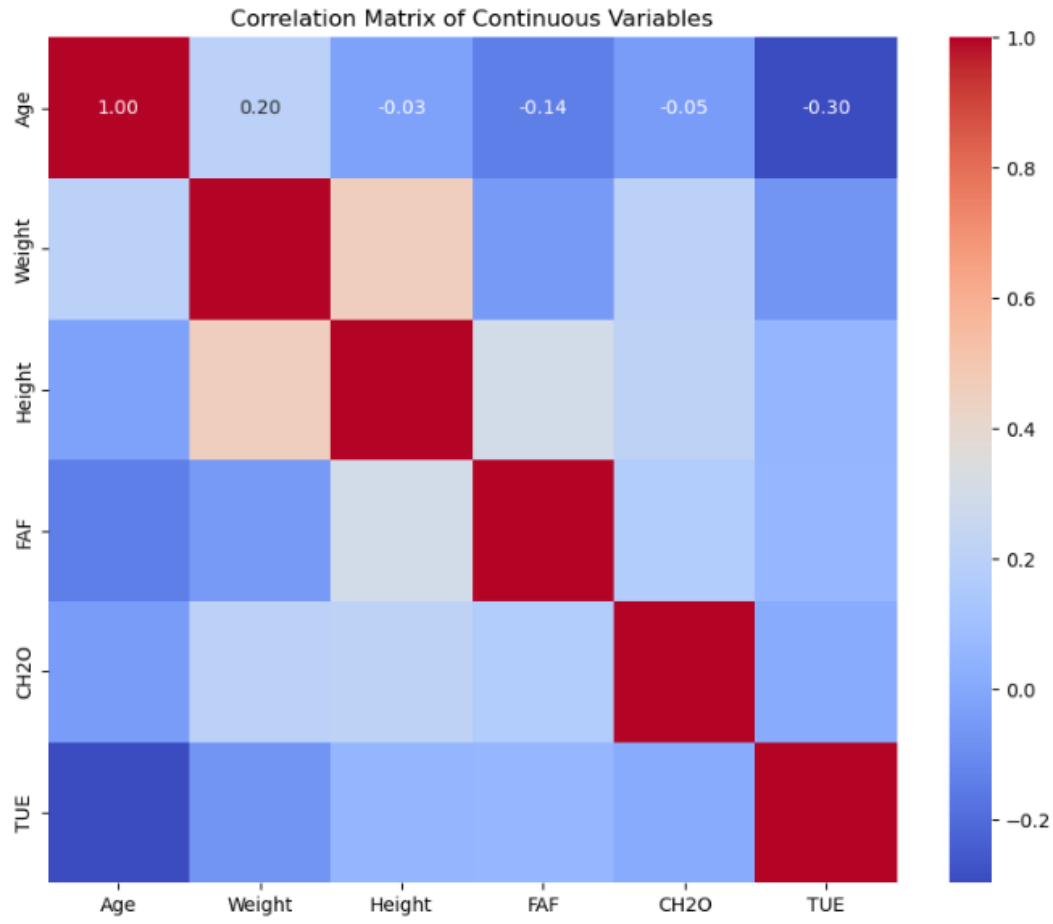

Figure 7. Correlation

According to the correlation heatmap, the following observations were made:

- Positive correlation between Age and Weight of 0.20
- Negative correlation between Age and Height of -0.03
- Positive correlation between Height and Weight of 0.46 which indicates a strong correlation.

There are other observable correlations in the figure above that was not mentioned.

Advanced Visualizations and Machine Learning

Advanced Visualizations

Pair plots and feature importance plots were created to gain insights into feature relationships and their significance in predicting obesity levels. A confusion matrix heatmap helped to visualize model performance.

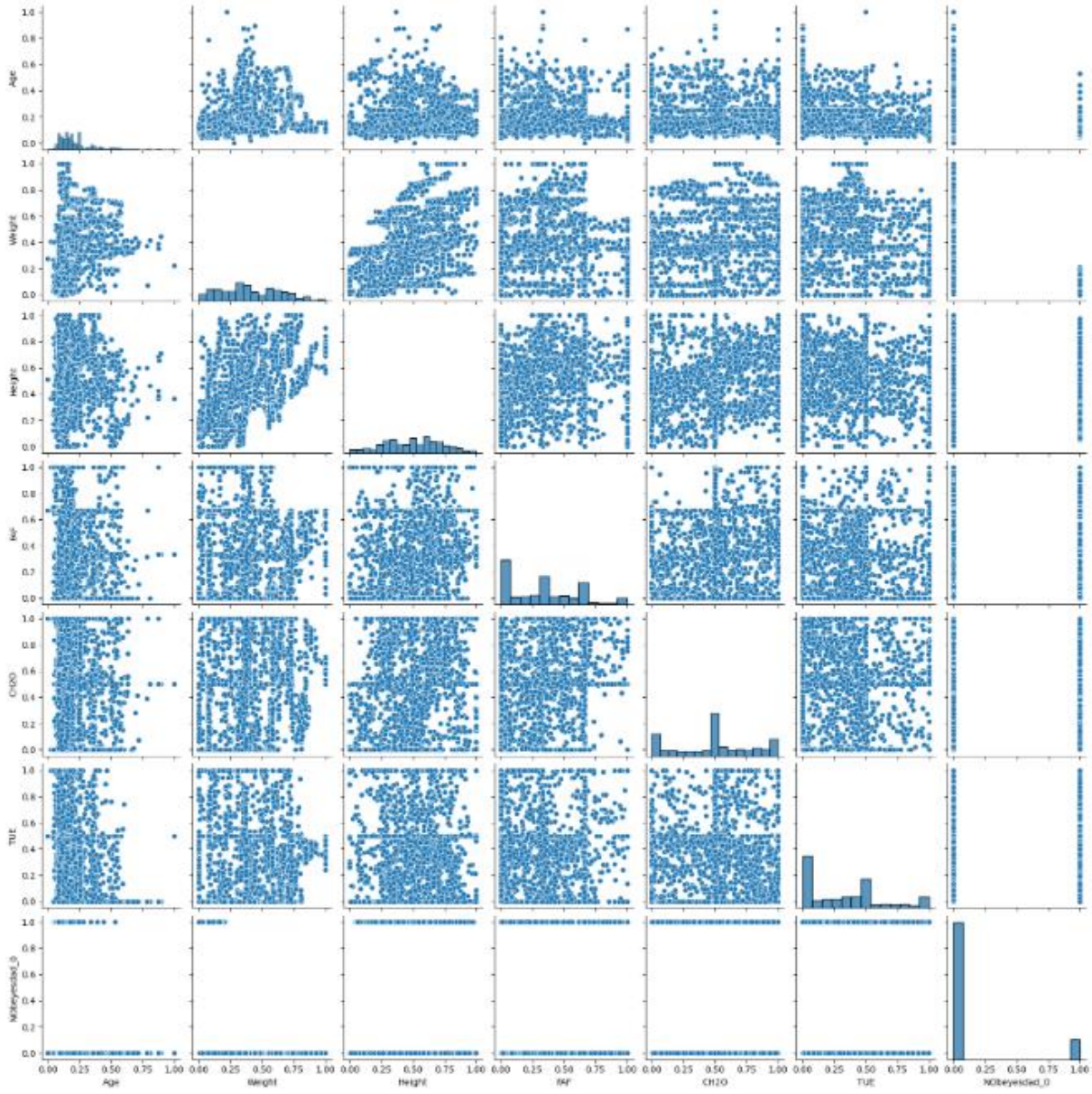

Figure 8. Confusion Matrix

The pair plots provide a visual representation of the relationships between different pairs of features in your dataset. Each small plot in the grid shows the scatter plot of two features against each other. The diagonal plots display the distribution of individual features (histograms or KDEs).

Pair Plots
Scatterplots: The scatter plots show the relationships between pairs of features. For instance, the scatter plot for "Age" vs. "Weight" suggests a positive relationship (as Age increases, Weight tends to increase).

Histograms: The diagonal plots show the distribution of each individual feature.

Linear Relationships: Some features, like "Age" and "Weight", appear to have linear relationships.

Potential Correlations: The relationships between features like "Age" and "Height" or "Height" and "TUE" suggest there might be some correlation, though it's not perfectly linear.

Outliers: There are potential outliers, especially for features like "TUE" where there are a few data points with significantly higher values).

Feature Importance Plot
Bars: Each bar represents a feature, and the height of the bar indicates its importance in the Random Forest model's predictions.

Most Important Features: The features with the highest bars are the most influential in the model's predictions.

Less Important Features: Features with shorter bars have a lesser impact.

Model Interpretation: This plot helps you understand which factors are most strongly associated with the target variable ("NObeyesdad_0").

Confusion Matrix Heatmap
Grid: The confusion matrix shows how many times the model correctly classified each category and how many times it made mistakes.

Diagonal: The diagonal entries represent the correct predictions.

Off-Diagonal: The off-diagonal entries represent misclassifications.

Model Performance: By examining the confusion matrix, you can evaluate the model's accuracy, precision, recall, and other performance metrics.

Error Patterns: If you see a significant number of misclassifications in certain categories, it might indicate that the model needs improvement.

Key Features: "Weight", "Age", and "Height" seem to be highly correlated with the target variable.

Feature Engineering and Scaling
Additional feature scaling and transformations were applied to ensure that features were prepared for machine learning models. All categorical variables were one-hot encoded and continuous variables were scaled.

Train-Test Split
The dataset was split into training (80%) and testing (20%) sets to evaluate model performance on unseen data.

Model Building

Two machine learning models were implemented:

- Logistic Regression: This model was chosen for its simplicity and interpretability when predicting multi-class classification problems.
- Random Forest Classifier: This ensemble model was chosen for its ability to handle a large number of features and its robustness against overfitting.

Observations
Logistic Regression typically performs well for certain obesity levels but struggles with minority classes (those labeled 1.0). It consistently has lower accuracy and F1-scores when there is an imbalance between the 0.0 and 1.0 classes (for example, in NObeyesdad_1, NObeyesdad_2, and NObeyesdad_5).

Random Forest outperforms Logistic Regression in almost all cases, achieving higher accuracy and F1-scores, especially for imbalanced classes. It handles the minority class (1.0) better, as seen in NObeyesdad_2, NObeyesdad_5, and NObeyesdad_6.

There is variation in performance as the obesity levels increase (NObeyesdad_1 to NObeyesdad_6), the accuracy tends to decrease for Logistic Regression, while Random Forest maintains higher accuracy.

For instance, NObeyesdad_1 shows a significant drop in recall for the positive class (1.0) in Logistic Regression, suggesting difficulty in predicting that class

Model Evaluation

The evaluation was a comparison of Logistic Regression and Random Forest models in predicting obesity levels, indicated by various classes NObeyesdad_0 to NObeyesdad_6. The metrics evaluated are accuracy, precision, recall, and F1-score, which help measure different aspects of the models' performances.

Performance Metrics:

- Accuracy: Measures the proportion of correctly predicted instances out of the total instances.
- Precision: Indicates the proportion of true positive predictions among all positive predictions. High precision means fewer false positives.
- Recall: Measures the proportion of true positive predictions among all actual positives. High recall indicates fewer false negatives.
- F1-score: The harmonic means of precision and recall, providing a balance between the two metrics

Results and Insights

Logistic Regression

- Generally, performs well for NObeyesdad_0 and NObeyesdad_4, but struggles with higher levels.
- Performance deteriorates in more challenging levels like NObeyesdad_1, NObeyesdad_2, and NObeyesdad_5, where accuracy ranges from 81% to 87%. The precision and F1-scores also indicate room for improvement, especially for classifying imbalanced or harder-to-predict categories.
- The logistic regression model seems to have difficulty generalizing to more complex patterns for these categories.

Random Forest

- Consistently outperforms Logistic Regression across all obesity levels. It achieves near-perfect or perfect scores in several categories (NObeyesdad_0, NObeyesdad_3, NObeyesdad_4) with accuracies close to 100%, suggesting better handling of non-linear patterns and feature interactions.
- Achieves perfect accuracy, precision, recall, and F1-score for NObeyesdad_4 and even for the more challenging categories (NObeyesdad_1, NObeyesdad_5, NObeyesdad_6) where Random Forest significantly outperforms Logistic Regression, achieving higher accuracy and better F1-scores, showcasing its robustness.
- Maintains high performance even in the presence of class imbalance, which is evident in levels with fewer positive cases.

Model Insights

- Random Forest outperforms Logistic Regression across most obesity categories, demonstrating better model performance, particularly in more complex classification

tasks. This suggests that Random Forest is better at capturing non-linear relationships and interactions between features.

- Logistic Regression shows a significant drop in performance for higher obesity levels, particularly in recall. This suggests that the model may struggle to identify positive cases (obesity levels) when they are less frequent in the dataset.
- The results indicate that while both models can predict obesity levels, Random Forest's ability to handle variability in the data makes it the preferred choice for this task.

Feature Importance

Given Random Forest's stronger performance, its feature importance capabilities can help identify which features are critical in predicting obesity levels. The model's inherent ability to rank feature importance can guide which factors have the highest predictive power in determining obesity levels. By extracting feature importance from the Random Forest model, we gain insights into which features are most significant for predicting obesity classification.

Table 3. Feature Importance

Feature	**Importance**
Weight	0.355141
Height	0.227643
Age	0.175697
FAF	0.089541
TUE	0.077501
CH2O	0.074477

Conclusion

This project successfully demonstrated the process of using machine learning to estimate obesity levels based on physical and dietary factors. Through effective data preprocessing, EDA, and model building, we gained valuable insights into the key factors influencing obesity. The Random Forest model proved to be the most effective, providing a reliable method for estimating obesity levels in individuals based on their eating habits and physical activity.