Jackie Zhou

Professor Jacob Koehler

ECON-UB 232 Data Bootcamp

December 16th, 2024

# Final Report

## 1. Introduction

Obesity is a chronic and complex condition characterized by an accumulation of excessive body fat that can harm health. It raises the risk of developing type 2 diabetes, heart disease, and certain cancers, while also impacting bone health and reproductive functions. Obesity influences the quality of living, such as sleeping or moving. According to data from WHO (World Health Organization), in 2022, 2.5 billion adults (18 years and older), which is about 43% of all adults, were overweight; of these, 890 million, which is about 16%, were living with obesity. Therefore, obesity is now a serious problem that needs to be addressed for many of us.

For my project, I plan to use a dataset to build a predictive model capable of assessing obesity levels based on various factors such as dietary habits and physical activity. This model may be useful to both individuals who are worried about their physical condition and policymakers to get awareness of the obesity level in a specific region. These predictions can be further used to develop recommender systems that monitor obesity levels. Upon successfully implemented, the model can contribute to public health initiatives by providing data-driven insights to design better preventive measures and resource allocation. Additionally, it has the potential to positively impact individuals' quality of life and overall societal health outcomes, addressing a critical global health challenge.

## 2. Dataset Description

Data for this project comes from UCI Machine Learning Repository in csv format. The dataset includes data for the estimation of obesity levels in individuals from the countries of Mexico, Peru, and Colombia, based on their eating habits and physical condition. It contains 17 attributes and 2111 records. The attributes associated with eating habits include: Frequent intake of foods high in calories (FAVC), Frequent intake of vegetables (FCVC), Number of main meals (NCP), In-between-meal food consumption (CAEC), and Ingestion of drinking alcohol (CALC) and drinking water every day (CH20). Physical activity frequency (FAF), time spent using technology devices (TUE), transportation used (MTRANS), and calorie consumption monitoring (SCC) are the attributes associated with physical condition. Other variables that were obtained included gender, age, height, and weight. Following the labeling of all the data, the class variable NObesity was created with the following values: Insufficient Weight, Normal Weight, and Overweight Level I, Overweight Level II, Obesity Type I, Obesity Type II, and Obesity Type III. The dataset contains no missing values.

Due to the specific nature of the questions presented some receive categorical answers and some receive numerical answers. Prior to analyzing the data it should be cleaned and organized. Then, exploratory data analysis (EDA) will be performed to better understand the relationships between the variables and hopefully come up with an obesity level prediction with a certain degree of accuracy.

From the correlation vector (figure 1 in appendix), we could observe that CAEC (Consumption of food between meals) and Family history with overweight have strong positive correlation with the obesity level (especially when taken into account that weight, which is supposed to be perfectly correlated with obesity level, has only 0.38 correlation). Additionally, Age and CH2O (Consumption of water daily) also have moderately positive correlation with high levels of obesity. This means that if too much food is consumed in between meals and if there's a family history of being overweight, the person is very likely to be overweight as well. Also, the older someone gets and the more water someone drinks, the more likely he/she gets fat. The latter is somehow counterintuitive, as it's always believed that drinking water is good for our health. On the other hand, only CALC (Ingestion of drinking alcohol) and FAF (Physical activity frequency) seem to be moderately negatively correlated with obesity level. While the latter makes intuitive sense, the former is somehow counterintuitive as drinking alcohol is always deemed as bad for health. Therefore, a more detailed check of relations is needed for these attributes.

To begin, I plotted a boxplot and violin plot for CAEC vs. Obesity level (figure 2 and 3). From these plots we could observe that apparently having more in between meal food lead to higher chances of high obesity level. Similarly, from the violin plot for family history vs obesity level (figure 4), we can also see that with a family history of being overweight, the individual is more likely to end up with some overweight problems, whereas the distribution of high obesity is roughly the same for the two groups. From the box plot for obesity vs. age (figure 5), we could see a slight trend: higher obesity levels tend to include a broader range of ages, with older individuals being more prevalent at higher obesity levels, while lower obesity levels are clustered among younger individuals. From the scatter plot of water intake vs obesity (figure 6), however, we see no clear linear relationship, as obesity levels are fairly evenly distributed across all CH2O values, indicating that water intake alone is not a strong predictor of obesity. Moving on to the attributes that have negative correlations with obesity level, the box plot of alcohol vs. obesity (figure 7) shows that individuals who do not consume alcohol tend to have lower obesity levels On the other hand, individuals with higher alcohol consumption display a broader spread of obesity levels, though the relationship is not strictly linear. Finally, the scatter plot of physical activities vs. obesity (figure 8) indicates that individuals with higher physical activity frequency tend to have lower obesity levels, as there is a noticeable concentration of lower obesity levels at higher FAF values. However, some variability exists, as higher obesity levels are also observed across all FAF values.

3. Models and Methods & Results

To accurately predict the obesity level, I chose to build four different models and compare to see which one has the best performance. For each of these models, I decided to utilize an 80-20 train-test split, training my model on 80% of the data and then testing it on the remaining 20%. To evaluate the success of predictive models, I established a baseline model by using the mean value of the target variable as the prediction for all data points. This approach provides a simple reference point for comparison, ensuring that the performance of more sophisticated models can be assessed meaningfully.

The first model I built is the logistic regression because it's relatively simple and straightforward. It also works well with small datasets like this one. Moreover, it provides insights into which attributes contribute most to predicting obesity levels, corresponding to what we did in the EDA. Overall, my model performs better than the baseline, as both the train and test mean square error are lower than the baseline error. The permutation importance analysis reveals that weight is the most influential predictor, followed by, unexpectedly, gender, height, and FAVC (Frequent consumptions of high caloric food). The next important features are family history with overweight and CAEC, just like what we supposed in the EDA.

Then, I decided to build a KNN model because the model works well when data naturally forms clusters, which aligns with the trends seen in the analysis of key features. By evaluating the "neighbors" of a given instance, KNN can accurately classify individuals into appropriate obesity categories, making it a strong choice for this dataset. As it turns out, my KNN model also performs better than the baseline with a much lower mean square error in both train and test case. Likewise, in the permutation importance analysis, weight is considered to be the most important attribute, followed by gender, CALC, FAVC, family history with overweight, and MTRAN (Transportation used). However, this time height is not considered to be an important attribute at all. It is probably because the regression model assigns coefficients based on the global linear relationship between each feature and the target variable, even if the relationship is weak; whereas the KNN model uses distance-based calculations to identify patterns locally, at which height does not vary significantly.

Moving on, I decided to use the decision tree model because of its ability to handle non-linear relationships, like the KNN models. It's also able to capture the interactions between several features, making it a great tool to investigate our dataset. Likewise, my decision tree model outperforms the baseline with an even lower mean square error in both test and train data. The important features recognized by the decision tree, however, are very different from the first two models: it identifies Weight and Height as the most significant predictors of obesity level; other features like Gender and Age play smaller roles, while lifestyle-related features like CALC and NCP contribute minimally.

Finally, I chose to use the random forest as an extension of the decision tree. Unlike a single Decision Tree, which can easily overfit to the training data, Random Forest reduces overfitting by introducing randomness during training. Therefore, I believe that this will be a great fit for my dataset. The result is not disappointing: it returns the lowest mean squared error! The important features recognized by random forest includes: weight, height, gender, followed by age, CAEC, and TUE (time spent using technology devices).

4. Interpretation

A key thing to notice in my models is that, the error returned by my models is monotonically decreasing: namely, the Random Forest has the best accuracy, followed by Decision Tree, and then KNN, and finally, logistic regression has the biggest error (but still better than the baseline). The explanation for this phenomenon is the following. The baseline model simply predicts the mean of the target variable, regardless of the input features, whereas the logistic regression model learns patterns from the data and fits a linear model to make predictions. So, it outperforms the baseline model. However, the big disadvantage of logistic regression is that it assumes a linear relationship between the input and output variables, which is why the important attributes recognized by the logistic regression model is similar to what we identified in the correlation vector; whereas the KNN model makes no assumptions about the underlying data distribution. Therefore, it returns a smaller error than the logistic regression for capturing the nonlinear relationship. Still, it may not be the best model for our dataset. As it turns out, the decision tree model outperforms the KNN model. This is because, albeit both models capture the nonlinear relationship, the decision tree splits and chooses different features at different thresholds so that it identifies dominant features more effectively and could give more weights to them; KNN, however, treats all features equally when calculating distances. For example, the KNN model gives an importance value of 0.149 to the attribute "weight" – only 0.04 more than the second attribute "gender" – whereas for decision trees, it gives "weight" an importance value of 0.694. Finally, the random forest model returns the smallest error. Specifically, it outperforms the decision tree model due to its ensemble approach: it combines multiple trees to reduce overfitting of one tree. As an aggregation of multiple predictions from various trees, it returns a more reliable prediction: the average of all predictions. Typically, in most cases, random forest returns a more accurate prediction than the decision tree, so we are happy to see that our result aligns with the general case.

5. Conclusion and Next Steps

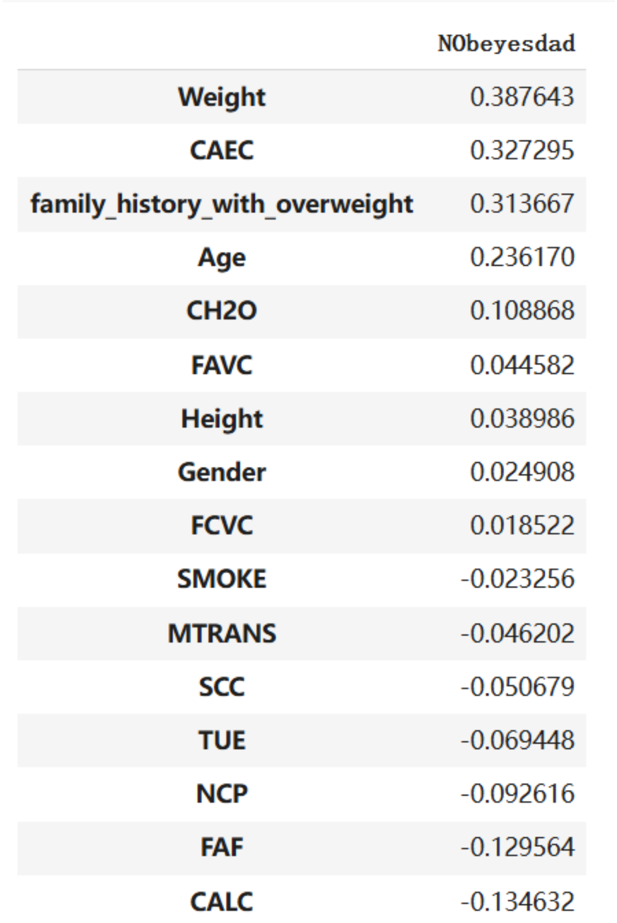
In conclusion, all of the models we built have better performance than the baseline model. They are ranked in accuracy as the following: Random Forest, Decision Tree, KNN, and logistic regression. This performance aligns with our expectations due to the ensemble nature of random forest. The decision tree outperforms the KNN for taking into consideration the dominant features in my dataset such as "weight". But they are all better than the logistic regression for capturing the nonlinear relationships. This also gives rise to another significant finding for this dataset: several features that have high correlation with the obesity level are not actually that important in the predictive models; instead, although attributes like "height" have a relatively low correlation, it appears many times in several models and are listed as important features. This shows how height is intrinsically related with several attributes and again demonstrates the nonlinearity nature of this dataset, which is why logistic regression may not be a good fit as a predictive model for this dataset. Overall, the superior performance of the random forest model highlights its ability to capture complex relationships within the data while reducing overfitting, offering valuable insights for future analyses and predictive modeling for datasets like this.

To build on the current findings and further improve the models' predictive capabilities, several next steps can be undertaken:

1. I believe that the dataset would be more meaningful and representative if the data collected were from a variety of regions instead of just Mexico, Peru and Colombia.
2. I think additional features like sleep quality, stress level,etc could also be included in the questionnaire to improve the model’s predictive power.
3. Other predictive models should also be applied to see if they could outperform random forest.
4. We should not stop at recognizing the problem of obesity. Moreover, these predictive models should be utilized to develop recommender systems that monitor obesity levels, etc. Also, important features recognized by the models should be used to inform public health campaigns and policy decisions.

By implementing these steps, a more accurate predictive model could be built in estimating obesity level, and the current analysis can evolve into a more useful and actionable framework for addressing obesity at both individual and societal levels.

## Appendix



| | NObeyesdad |
|---|---|
| **Weight** | 0.387643 |
| **CAEC** | 0.327295 |
| **family_history_with_overweight** | 0.313667 |
| **Age** | 0.236170 |
| **CH2O** | 0.108868 |
| **FAVC** | 0.044582 |
| **Height** | 0.038986 |
| **Gender** | 0.024908 |
| **FCVC** | 0.018522 |
| **SMOKE** | -0.023256 |
| **MTRANS** | -0.046202 |
| **SCC** | -0.050679 |
| **TUE** | -0.069448 |
| **NCP** | -0.092616 |
| **FAF** | -0.129564 |
| **CALC** | -0.134632 |

fig.1

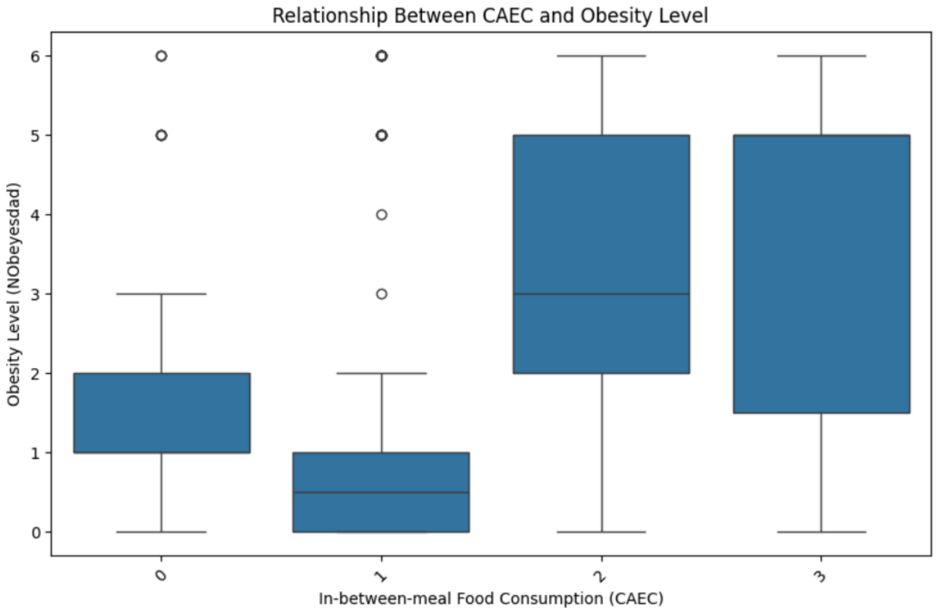


fig. 2

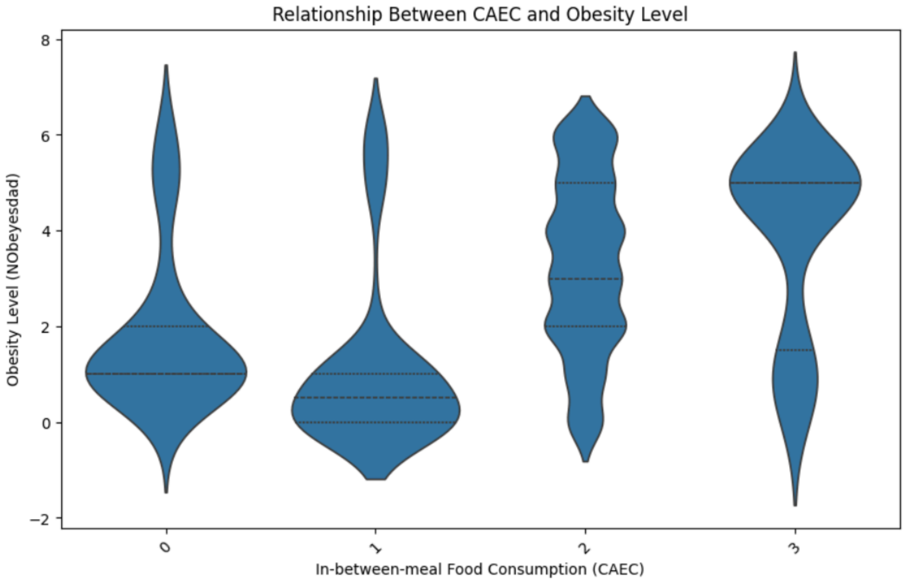

Relationship Between CAEC and Obesity Level
Obesity Level (NObeyesdad)
8
6
4
2
0
−2
0
1
2
3
In-between-meal Food Consumption (CAEC)

fig. 3

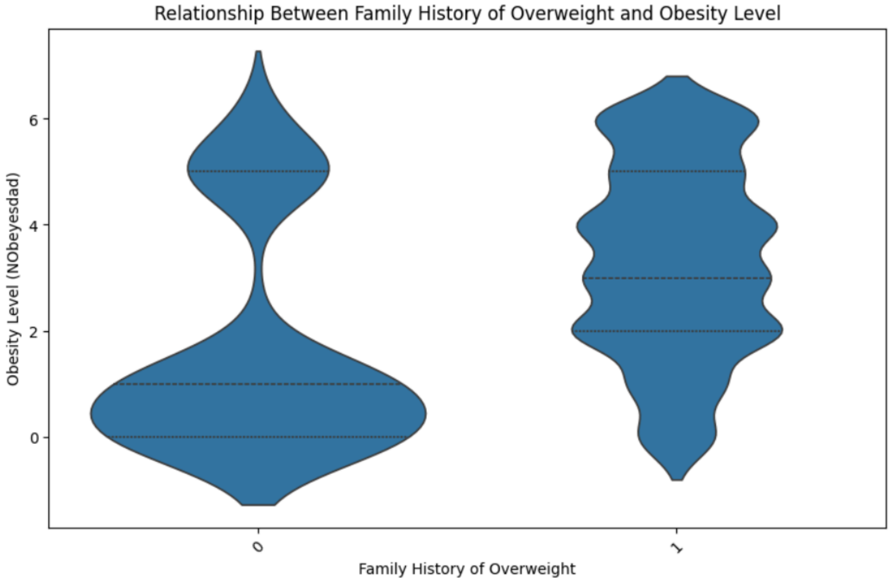

Relationship Between Family History of Overweight and Obesity Level
Obesity Level (NObeyesdad)
6
4
2
0
0
1
Family History of Overweight

fig. 4

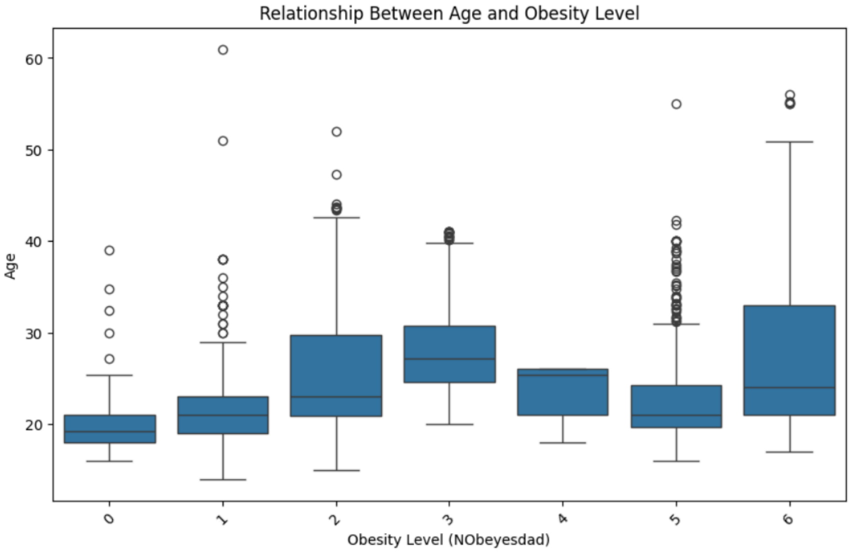

Relationship Between Age and Obesity Level
Age
60
50
40
30
20
0
1
2
3
4
5
6
Obesity Level (NObeyesdad)

fig. 5

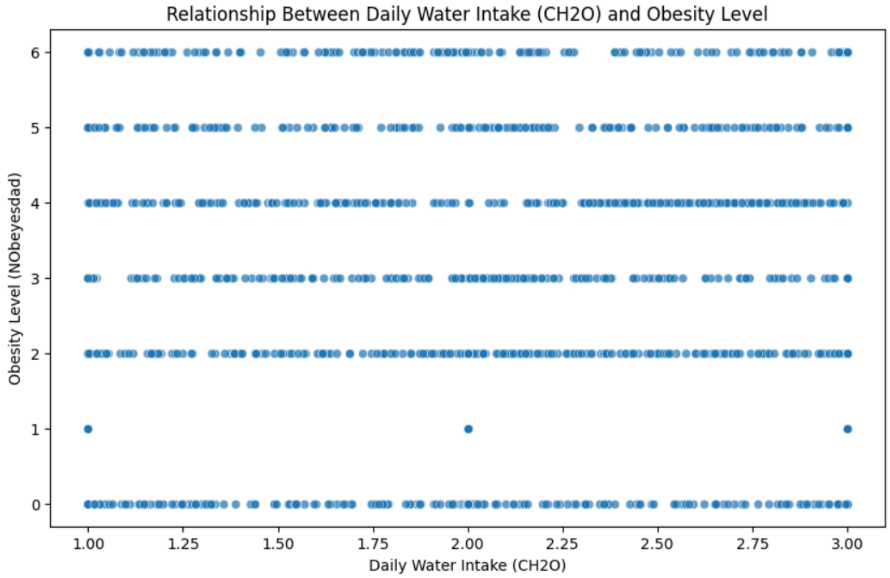

Relationship Between Daily Water Intake (CH2O) and Obesity Level
Obesity Level (NObeyesdad)
6
5
4
3
2
1
0
1.00
1.25
1.50
1.75
2.00
2.25
2.50
2.75
3.00
Daily Water Intake (CH2O)

fig. 6

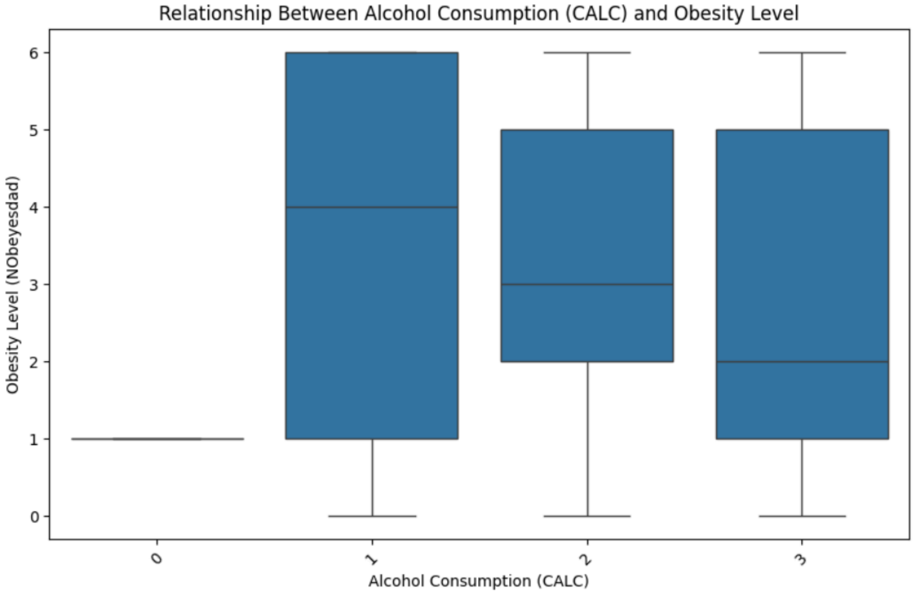

Relationship Between Alcohol Consumption (CALC) and Obesity Level
Obesity Level (NObeyesdad)
0
1
2
3
4
5
6
Alcohol Consumption (CALC)
0
1
2
3

fig. 7

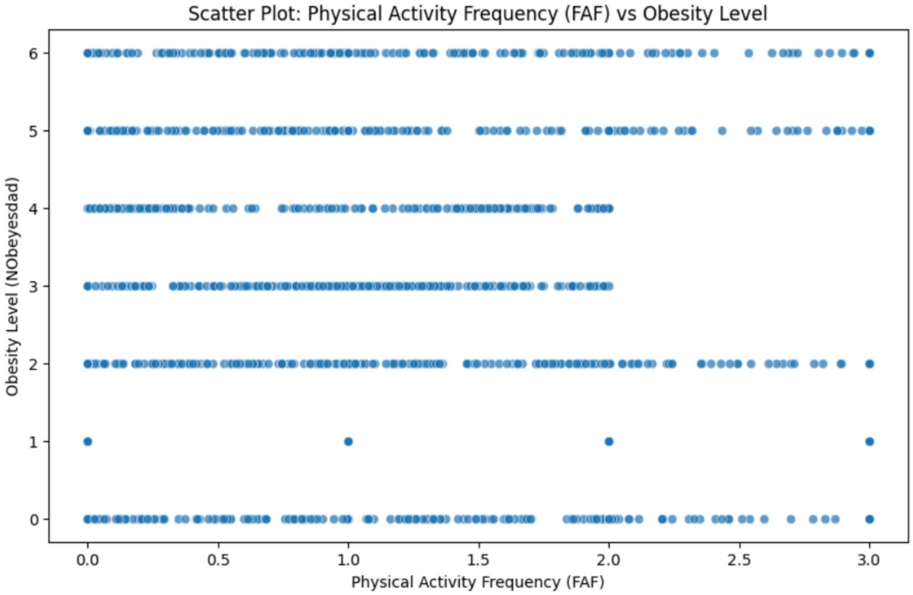

Scatter Plot: Physical Activity Frequency (FAF) vs Obesity Level
Obesity Level (NObeyesdad)
0
1
2
3
4
5
6
Physical Activity Frequency (FAF)
0.0
0.5
1.0
1.5
2.0
2.5
3.0

fig. 8