**Project Title: Estimation of Obesity Levels Using Machine Learning**

**Data Analysis and Visualization  
Semester Project Proposal**

**Session *2022-2026***



|  |  |  |
| --- | --- | --- |
| 1 | *Aqeel Abbas Khan* | *NUM-BSCS-2022-38* |
| 2 | *Muhammad Asim Khan* | *NUM-BSCS-2022-14* |

Department of Computer Science

Namal University Mianwali

07 May 2025

**Introduction**

Obesity is a critical global health challenge, with the World Health Organization reporting that over 1 billion people were living with obesity in 2022 ([WHO Obesity Factsheet](https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight)). It is linked to chronic diseases like diabetes and heart disease, making it essential to understand its contributing factors and predict risk for early intervention. This project proposes using a publicly available dataset to analyze obesity levels through machine learning, aiming to identify key predictors and develop accurate predictive models. By exploring demographic, dietary, and lifestyle factors, the study seeks to provide insights that could inform public health strategies and personalized interventions.

**Dataset Source and Description**

The dataset, titled "Estimation of Obesity Levels Based On Eating Habits and Physical Condition," is sourced from the UCI Machine Learning Repository ([UCI Dataset](https://archive.ics.uci.edu/dataset/544/estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition)). It comprises 2111 records and 17 attributes, collected from individuals in Mexico, Peru, and Colombia. The attributes include:

| **Attribute** | **Type** | **Description** |
| --- | --- | --- |
| Gender | Categorical | Male or Female |
| Age | Continuous | Age in years |
| Height | Continuous | Height in meters |
| Weight | Continuous | Weight in kilograms |
| family\_history\_with\_overweight | Binary | Family history of overweight (yes/no) |
| FAVC | Binary | Frequent consumption of high caloric food (yes/no) |
| FCVC | Integer | Frequency of vegetable consumption in meals |
| NCP | Continuous | Number of main meals daily |
| CAEC | Categorical | Consumption of food between meals |
| SMOKE | Binary | Smoking status (yes/no) |
| CH2O | Continuous | Daily water intake in liters |
| SCC | Binary | Monitoring of calorie consumption (yes/no) |
| FAF | Continuous | Frequency of physical activity |
| TUE | Integer | Time using technological devices in hours |
| CALC | Categorical | Alcohol consumption frequency |
| MTRANS | Categorical | Mode of transportation (e.g., walking, public transport) |
| NObeyesdad | Categorical | Obesity level (target variable: Insufficient Weight, Normal Weight, Overweight Level I/II, Obesity Type I/II/III) |

The dataset includes no missing values, with 77% of the data synthetically generated using the Weka tool and SMOTE filter, and 23% collected directly from users via a web platform ([Dataset Paper](https://www.sciencedirect.com/science/article/pii/S2352340919306985)). This composition supports robust analysis but requires caution in generalizing findings.

**Research Questions**

The project aims to address the following questions:

1. **Key Influencing Factors:** What are the primary factors contributing to different obesity levels in the studied population? This involves analyzing the impact of demographic, dietary, and lifestyle variables.
2. **Predictive Accuracy:** How effectively can machine learning models predict obesity levels based on the provided features? This focuses on developing and evaluating classification models.
3. **Demographic Variations:** Are there significant differences in obesity levels across demographic groups, such as gender or age? This explores potential disparities for targeted interventions.

These questions align with the dataset’s comprehensive feature set and support the goal of deriving actionable health insights.

**Proposed Methodology**

The project will follow a structured approach to analyze the dataset and answer the research questions:

* **Data Preprocessing:** Categorical variables will be encoded, and numerical features (e.g., Age, Height) will be standardized to ensure model compatibility.
* **Exploratory Data Analysis (EDA):** Initial analysis will examine feature distributions, correlations, and relationships with obesity levels to identify patterns and potential issues like multicollinearity.
* **Model Development:** Multiple machine learning models will be implemented for multi-class classification to choose the best fit model.

**Potential Challenges and Mitigation Strategies**

Several challenges may arise during the project, each with a planned mitigation strategy:

| **Challenge** | **Description** | **Mitigation Strategy** |
| --- | --- | --- |
| Synthetic Data | 77% of the data is synthetic, potentially limiting real-world applicability. | Focus on dataset-specific patterns and acknowledge generalizability limitations in findings. |
| Class Imbalance | Uneven distribution of obesity levels may skew model performance. | Apply class weighting to balance classes during model training. |
| Feature Correlation | High correlation between features may cause multicollinearity. | Conduct correlation analysis and use feature selection to reduce redundant features. |

Additional considerations include the dataset’s regional focus (Mexico, Peru, Colombia), which may limit applicability to other populations, and ethical concerns around obesity-related stigmatization, which will be addressed by framing results sensitively.

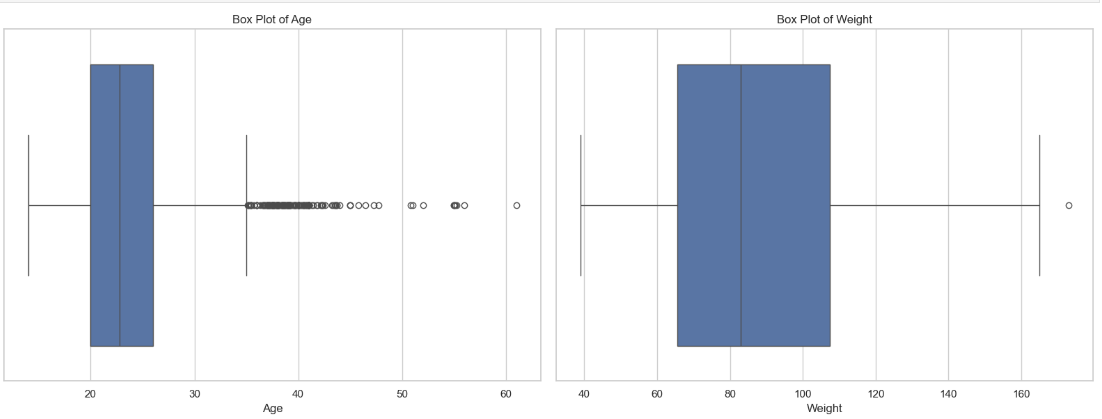
**Expected Outcomes**

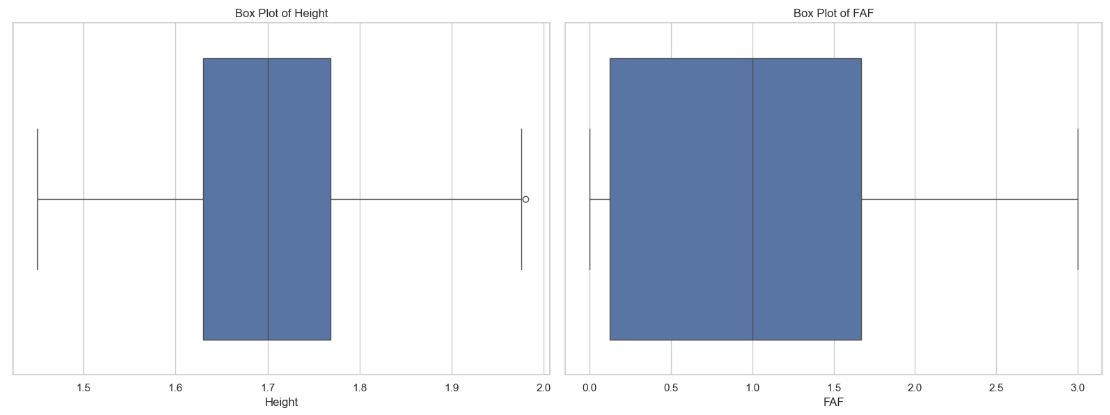
The project anticipates the following outcomes:

* **Identification of Key Factors:** Determining which demographic, dietary, or lifestyle factors most strongly predict obesity levels, potentially highlighting actionable areas like physical activity or dietary habits.
* **Accurate Predictive Models:** Developing machine learning models with high accuracy, comparable to studies achieving good accuracy.
* **Demographic Insights:** Uncovering variations in obesity levels by gender or age, informing targeted public health interventions.
* **Public Health Implications:** Providing insights that could guide personalized health recommendations and broader obesity prevention strategies.

These outcomes aim to contribute to the growing body of research on obesity prediction and support evidence-based health policies.

## **FINDING OUTLIERS:**

Let’s plot the box plot for finding outliers:  




* Box plot of a shows that there lies too much outlier in age column. So can not remove from this feature because this process can loss to much data from our data set.
* Weight feature contain just single outlier. So we have removed row related to that.
* Height also contain just single outlier. So we have also removed that one row.
* Physical activity have no outlier.

## EDA:

Let try to analyze the correlation of different other feature:

## Highly Correlated Feature Pairs:

1. **Height and Weight** (Correlation = **+0.46**)

Taller individuals tend to have higher body weight.

1. **Height and FAF (Physical Activity Frequency)** (Correlation = **+0.29**)

Taller individuals tend to engage more in physical activities.

1. **Height and NCP (Number of Main Meals Per Day)** (Correlation = **+0.24**)

Taller individuals may have slightly more main meals daily.

1. **Weight and FCVC (Vegetable Consumption Frequency)** (Correlation = **+0.21**)

Heavier individuals tend to eat vegetables slightly more frequently.

1. **CH2O (Water Intake) and FAF (Physical Activity Frequency)** (Correlation = **+0.16**)

Individuals who drink more water are somewhat more physically active.

## Most Negative (Inverse) Correlated Feature:

1. **Age and TUE (Technology Usage Time)** (Correlation = **-0.29**)

Older individuals spend less time using technological devices compared to younger individuals.

## Weakest (Almost No) Correlations:

1. **Age and Height** (Correlation = **-0.02**)

No meaningful relationship between a person's age and their height.

1. **FCVC (Vegetable Consumption) and FAF (Physical Activity)** (Correlation = **+0.02**)

No strong relation between vegetable eating habits and physical activity frequency.

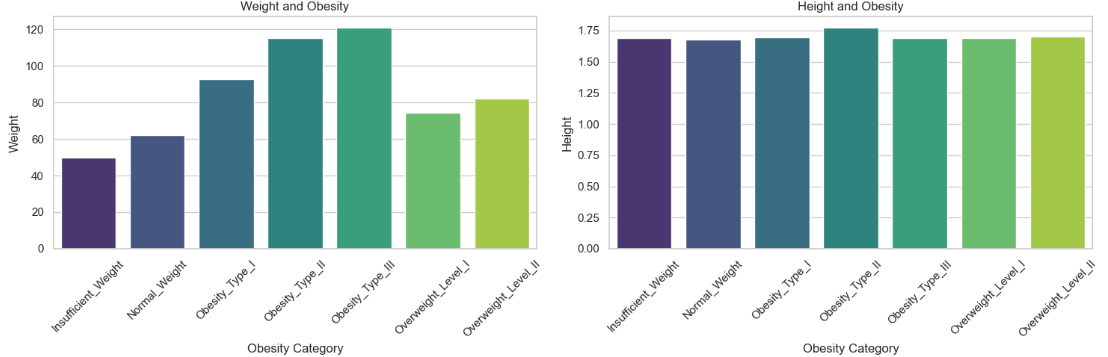
1. **NCP (Number of Meals) and TUE (Technology Usage)** (Correlation = **+0.04**)

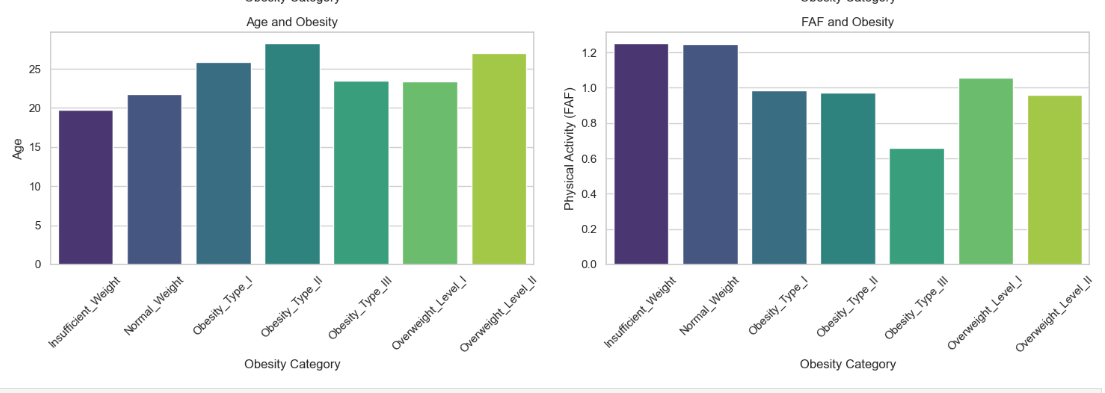
No strong relationship between how often someone eats and their technology usage time.

**Relationship Between Different Features And Obesity:**

In this section we will try to analyze the relation and effect of other features on target variable.

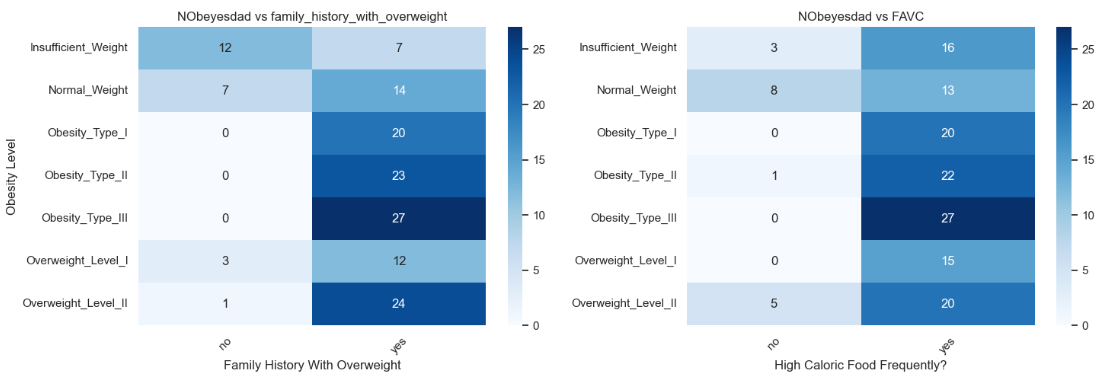
For that first we will plot bar graph on different features to analyze the impact of those features on obesity.

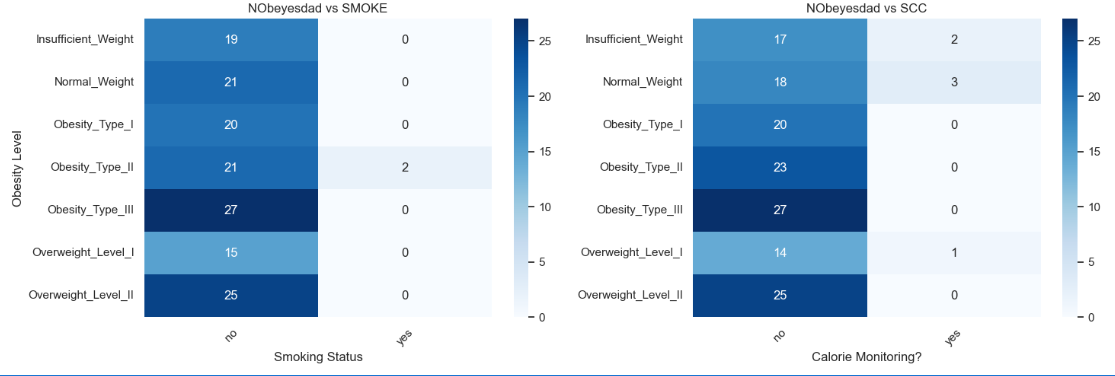




1. **Weight and Obesity**  
   As obesity category increases from **Insufficient Weight** to **Obesity\_Type\_III**, weight also increases significantly. Weight then decreases again slightly in **Overweight\_Level\_I** and **II**, but remains higher than normal weight categories.
2. **Height and Obesity**  
   Height remains relatively constant across all obesity categories with only minor fluctuations. There is no strong relationship between height and obesity in this dataset.
3. **Age and Obesity**  
   Age tends to increase with higher obesity categories up to **Obesity\_Type\_II**, but decreases slightly in **Obesity\_Type\_III** and **Overweight\_Level\_I** before rising again in **Overweight\_Level\_II**. Overall, higher obesity levels are associated with middle to older age groups.
4. **FAF (Physical Activity) and Obesity**  
   Higher physical activity levels (FAF) are associated with **Insufficient Weight** and **Normal Weight**. As obesity level increases, physical activity generally decreases, showing lower activity in **Obesity\_Type\_II** and **III**, but a slight increase again in **Overweight\_Level\_I**.

Now we will analyze the binary features to check their relation with obesity.





1. **NObeyesdad vs family\_history\_with\_overweight**  
   Individuals with a family history of overweight ("yes") have higher counts in obesity categories, especially **Obesity\_Type\_III** and **Overweight\_Level\_II**. Lack of family history shows higher counts in lower obesity levels or insufficient weight.
2. **NObeyesdad vs FAVC (High Caloric Food Frequently?)**  
   People who frequently consume high-calorie food ("yes") are concentrated more in higher obesity types (Obesity\_Type\_II, III). Those who do not consume high-calorie food regularly ("no") are more represented in Insufficient\_Weight and Obesity\_Type\_I.
3. **NObeyesdad vs SMOKE (Smoking Status)**  
   Both smokers and non-smokers show similar distribution across obesity levels, but non-smokers ("no") slightly dominate higher obesity categories (Obesity\_Type\_III, Overweight\_Level\_II). Smoking doesn't appear to drastically skew obesity levels.
4. **NObeyesdad vs SCC (Calorie Monitoring?)**  
   Individuals who monitor their calories ("yes") show slightly higher counts in lower obesity levels and insufficient weight. Those who do not monitor calories ("no") have greater numbers in higher obesity types, especially Obesity\_Type\_III and Overweight\_Level\_II.