**Estimation of Obesity Levels Based on Eating Habits and Physical Condition**

## *Introduction*

Obesity has emerged as a significant global public health issue, affecting millions of people across diverse demographics. The rise in obesity levels over recent decades can be attributed to various factors, including changes in lifestyle, dietary habits, and physical activity levels. Obesity not only affects the individual's quality of life but also poses a major economic burden on healthcare systems due to its association with chronic diseases like diabetes, hypertension, cardiovascular diseases, and certain cancers.

Given the widespread nature of this problem, understanding the factors that contribute to obesity is crucial for developing targeted interventions. The goal of my research project is to estimate obesity levels based on eating habits and physical conditions. Data science provides a powerful set of tools to analyze and model the complex relationship between eating behaviors, physical activity, and obesity.

This research project will explore how certain variables, such as eating patterns, physical exercise, and demographic characteristics, are associated with obesity levels. This project will focus on creating a recommendation for a model to predict obesity levels based on these variables.

## *Research Questions*

To guide the research, I will be addressing the following research questions:

1. What are the key eating habits that contribute to obesity?
2. How do physical conditions such as exercise frequency and duration affect obesity levels?
3. Can demographic factors such as age, gender, and socioeconomic status be strong predictors of obesity levels?
4. What types of foods (processed vs. natural) are most strongly correlated with obesity?
5. How can I create a model to accurately predict an individual's obesity level based on a combination of physical condition and eating habits?
6. Is there any evidence of interaction between physical activity and eating habits in relation to obesity?

### *Proposed Approach*

To address these research questions, I plan to perform a comprehensive exploratory data analysis (EDA) using R programming. The focus will be on identifying the relationships between the variables related to physical activity, eating habits, and obesity levels. I will clean and transform the data, conduct statistical analyses, and produce visualizations to uncover trends and patterns.

Key aspects of my approach include:

- **Data Cleaning**: Handle missing values, remove outliers, and standardize the dataset for consistent analysis.
- **Exploratory Data Analysis**: Use summary statistics and visualizations to explore the distribution and relationships between variables.
- **Feature Engineering**: Create new variables where necessary to capture meaningful relationships (e.g., creating composite variables for total caloric intake or intensity of physical exercise).
- **Modeling**: Based on the results of the EDA, I will recommend a machine learning model (e.g., linear regression, decision trees) or statistical approach (e.g., logistic regression) to estimate obesity levels.
- **Validation**: Use cross-validation techniques to ensure the robustness of the model.

### *Addressing the Problem Statement*

The proposed approach will address the problem statement by:

- **Exploring the Data**: Thorough analysis of the dataset will reveal patterns related to eating habits, physical condition, and obesity levels.
- **Predictive Modeling**: By recommending a predictive model, I will provide insights into how various factors contribute to obesity. This could be valuable in designing interventions for individuals or communities at risk.

### *Datasets*

I have identified three datasets that will be used to address the research questions:

1. **Obesity Levels Based on Eating Habits and Physical Conditions** (available on UCI Machine Learning Repository):

https://archive.ics.uci.edu/dataset/544/estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition

2. **National Health and Nutrition Examination Survey (NHANES)**:

   https://wwwn.cdc.gov/Nchs/Nhanes/Search/DataPage.aspx?Component=Demographics&CycleBeginYear=2021

3. Kaggle : https://www.kaggle.com/datasets/abdelrahman16/obesity-dataset

### *Packages to Use*

To perform the analysis, the following R packages will be essential:

- **ggplot2**: For creating visualizations that represent trends and relationships.
- **dplyr**: For data transformation and summarization.
- **MASS**: For performing statistical analysis, including ANOVA.
- **readxl** : The *readxl package* makes it easy to get data out of Excel
- **GGally***: GGally* extends 'ggplot2' by adding several functions

### *Visualizations and Plots*

Some of the key types of plots and tables I will create include:

1. **Correlation Matrix**: To visualize relationships between different predictors and the target variable (obesity levels).
2. **Boxplots and Histograms**: To show the distribution of eating habits, physical conditions, and their relationship with obesity levels.
3. **Scatterplots**: To visualize relationships between continuous variables such as calorie intake and BMI.
4. **Bar Plots**: To explore categorical variables such as the frequency of exercise and obesity categories.
5. **Summary Tables**: For providing descriptive statistics like means, medians, and standard deviations across different groups.

### *What I Need to Learn*

Currently, I need to enhance my understanding in the following areas:

1. **Handling Complex Datasets**: NHANES and USI datasets involve complex sampling methodologies, so I will need to understand how to handle them in R.
2. **Cross-Validation and Model Selection**: I need to learn how to systematically select models and apply cross-validation techniques to avoid overfitting and ensure generalization.
3. **Model Interpretation**: Interpreting complex models in the context of obesity predictions will be important to make actionable recommendations.

### *Conclusion*

In conclusion, my research project will focus on understanding and estimating obesity levels based on eating habits and physical conditions. This will be accomplished through rigorous data analysis and visualization using R, followed by model recommendations.