Dataset Title

Estimation of obesity levels based on eating habits and physical condition (source: UC Irvine Machine Learning Repository)

"Obesity being a global health concern, affects almost every country in the world. It's a major crisis and a real threat to global health progress Machine learning (ML) offers promising tools to predict obesity risk by analyzing lifestyle, genetic, and socioeconomic factors. ML gives more realistic and predictive outcomes than the traditional methods like surveys etc. which relies on linear aassumptions."

"In 2022, over 1 billion people worldwide were obese (including 650 million adults, 340 million adolescents, and 39 million children). Obesity rates have nearly tripled since 1975."
Source**:** World Health Organization (WHO)

"Annual U.S. medical costs for obesity-related conditions (e.g., diabetes, heart disease) exceed $173 billion."
Source: CDC Obesity Consequences (2022)

Research Questions and Justification

Q1. How do dietary habits and physical activity influence obesity levels?

- Dietary habit and physical activity closely linked with obesity levels.
- High calorie food and vegetable intake reflect eating pattern.
- Physical activity is a factor in obesity.
- Understanding how these factors interact can guide targeted intervention.

Techniques & Tools

- Multinomial Logistic Regression
    - Reason: The outcome (NObeyesdad) is categorical (e.g., Normal_Weight, Obesity_Type_I)
- Application: Model obesity levels as a function of:
    - Diet (FAVC, FCVC, NCP, CAEC
    - Physical activity (FAF)
- ANOVA/Post-hoc Tests

Tools

- Python: statsmodels, scikit-learn
- Visualization: matplotlib,

Formula for Multinomial Logistic Regression:

$$P(y = k \mid \mathbf{x}) = \frac{e^{\beta_k^T \mathbf{x}}}{\sum_{j=0}^{K} e^{\beta_j^T \mathbf{x}}} \quad \text{for each class } k = 0, 1, \ldots, K$$

Where:

- y: obesity level (encoded 0–N)
- X: vector of input features like (FAVC, FCVC, NCP, CAEC_frequently, FAF)
- βk: coefficients for each class k

Q2. Does family history of overweight influence the relationship between lifestyle choices (diet, exercise) and obesity?

Techniques & Tools

- Moderated Regression Analysis

    - Tests if family history of obesity changes the effect of lifestyle on obesity.
- Subgroup Analysis
    - Stratify by family history (yes/no) and compare regression coefficients.
- Visualization of interaction with diet and physical exercise

Tools

- Python: statsmodels (for moderation),

Formula

Multinomial logistic regression models the **log-odds of each class** $y = k$ relative to a **reference class** (typically $y = 0$):

$$\log\left(\frac{P(y=k)}{P(y=0)}\right) = \beta_{0k} + \beta_{1k} \cdot \text{FAVC} + \beta_{2k} \cdot \text{FAF} + \beta_{3k} \cdot \text{FH} + \beta_{4k} \cdot (\text{FAVC} \times \text{FH}) + \beta_{5k} \cdot (\text{FAF} \times \text{FH}) \quad \text{for } k = 1, 2, ..., K$$

Where:

- P(y=k): probability of being in obesity class k
- P(y=0): probability of the reference class (e.g., "Normal_Weight")
- FAVC: indicator for frequent high-caloric food (0 or 1)
- FAF: FAF: physical activity frequency
- FH: family history (0 or 1)
- FAVC×FH: interaction term
- FAF×FH: interaction term
- βjk: coefficient for predictor j and class k

Q3: Can individuals be classified into obesity categories using machine learning models based on their daily habits and physical characteristics?

Techniques & Tools

• Classification Models

- Train/Test Split (Stratified
  - Ensure balanced class distribution.
- Random Forest
  - Train a robust, ensemble-based classifier

• Classification Report

- Evaluate precision, recall, f1-score

• Confusion Matrix

- Visualize correct vs incorrect predictions

Tools

• Python:

- scikit-learn, statmodels.

Formula

Radom Forest

$$\text{Class} = \arg\max_k \left(\sum_{i=1}^{n_{\text{trees}}} 1[\text{Tree}_i(x) = k] \right)$$

- 1[·] is an indicator function (1 if condition is true, else 0).
- k is the class label.
- The final prediction is based on majority voting across all trees.

Precision, Recall, F1-score

Precision

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Recall

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

F1-score

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Ref.

Mozaffarian, D. et al. (2011)

- "*Changes in Diet and Lifestyle and Long-Term Weight Gain in Women and Men*"
- New England Journal of Medicine

- Key Finding: Identified specific dietary components (e.g., sugary beverages, processed foods) as significant predictors of obesity (OR: 1.23–1.93).
- Logistic Regression Use: Modeled weight gain as a binary outcome (obese/non-obese).
- DOI: 10.1056/NEJMoa1014296

Donnelly, J. E. et al. (2009)

- "*Physical Activity and Weight Management*"
- Medicine & Science in Sports & Exercise
- Key Finding: Sedentary behavior predicted obesity (OR: 2.1) independent of diet.
- Logistic Regression Use: Controlled for covariates like age and baseline BMI.
- DOI: 10.1249/MSS.0b013e3181949333

Obesity Level Estimation from Lifestyle (Kaggle notebook)

- Techniques: Data cleaning, feature engineering, Random Forest, SVM, etc.
- Link: https://www.kaggle.com/code/rafaelmarino/obesity-level-estimation