Estimation of Obesity Levels Based on Eating Habits and Physical Condition (Project Group 13 : BARNICHA, BENTAYFOR, DAOUDI, EL OUATILI, KAHRAMANE)

By Fabio Mendoza Palechor, Alexis de la Hoz Manotas
Universidad de la Costa, CUC, Colombia

Abstract

This document describes a dataset for estimating obesity levels among individuals from Mexico, Peru, and Colombia based on eating habits and physical conditions. The dataset includes 2,111 records with 17 attributes, offering valuable insights into obesity factors.

1 Introduction

Obesity is a growing health concern globally, driven by lifestyle factors. This dataset provides anonymized data collected via online surveys to analyze the correlation between lifestyle habits and obesity in Latin American populations. Principal Component Analysis (PCA) can be used here to reduce data dimensionality, thus enhancing the focus on significant attributes and minimizing the bias of majority classes.

2 Data Collection and Preprocessing

Data was collected online from individuals aged 14 to 61 across Mexico, Peru, and Colombia. The survey, conducted over 30 days, included questions about dietary habits, physical activity, and technology use. After data collection, PCA can be applied to retain the most informative components, thus improving the dataset's suitability for classification.

3 Attributes and Classification

The dataset encompasses 17 attributes divided as follows:

Eating Habits

- High-caloric food consumption (FAVC)
- Vegetable consumption frequency (FCVC)
- Main meals per day (NCP)
- Food between meals (CAEC)
- Daily water intake (CH20)
- Alcohol consumption (CALC)
- Smoking habits (SMOKE)

Physical Condition

- Caloric monitoring (SCC)
- Physical activity frequency (FAF)
- Technology use (TUE)
- Transportation mode (MTRANS)

Demographics and Additional Factors

- Gender, Age, Height, Weight
- Family history of overweight

4 Labeling and Obesity Classification

Each individual is classified into one of seven obesity levels, using BMI thresholds as per WHO and Mexican standards:

1. Insufficient Weight: BMI < 18.5
2. Normal Weight: $18.5 \leq$ BMI < 24.9
3. Overweight Level I: $25.0 \leq$ BMI < 29.9
4. Overweight Level II: $30.0 \leq$ BMI < 34.9
5. Obesity Type I: $35.0 \leq$ BMI < 39.9
6. Obesity Type II: $40.0 \leq$ BMI < 44.9
7. Obesity Type III: BMI > 45

5 Dimensionality Reduction with PCA

To enhance interpretability and address class imbalance, PCA can be applied to the dataset. By reducing dimensionality, PCA allows for the selection of principal components that capture the most variance in the data, focusing the analysis on the most significant attributes. This process also reduces the likelihood of bias toward majority classes, making the data more robust for training models.

6 Data Application

The dataset's structure supports various data mining techniques, including classification, prediction, segmentation, and clustering. It enables the development of tools to estimate obesity risks and inform public health strategies.

7 Conclusion

This dataset offers a rich foundation for studying the impact of lifestyle on obesity. Its format allows researchers to apply data analysis techniques effectively, fostering advancements in obesity detection and intervention tools.