Estimation of Obesity Levels Based on Eating Habits and Physical Condition

Surenther Selvaraj

Bellevue University

DSC 520: Statistics for Data Science

Professor: Andrew Hua

Date: Nov 01, 2024

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

Data Import and Cleaning

- **Data Import**: I imported the datasets using the read.csv() function for CSV files and the readxl package for Excel files in R. After importing, I inspected the data for missing values, outliers, and incorrect data types.
- **Cleaning Steps**:
  - **Handling Missing Values**: Minor missing values were imputed with the mean (for numerical data) or the mode (for categorical data). For significant omissions, I either used advanced imputation techniques or removed rows if the missingness was too extensive.
  - **Removing Duplicates and Outliers**: I removed duplicate records to ensure each entry was unique and managed outliers using the interquartile range (IQR) method, which is particularly useful for data skewed by extreme values.
  - **Standardization and Encoding**: Continuous variables were standardized for uniform scaling, and categorical variables were encoded using one-hot encoding to prepare them for statistical and machine learning analysis.

Final Dataset Overview

The cleaned dataset is condensed to essential, pre-processed features and includes a mix of continuous and categorical variables. A summary of the final data structure:

- **Continuous Variables**: Includes key statistics like mean, median, and standard deviation for features such as caloric intake, exercise frequency, and BMI.
- **Categorical Variables**: Contains frequency counts for categorical features like diet type and exercise category.

Non-Self-Evident Information

Some relationships in the data are not immediately clear, such as the subtle influences of certain lifestyle factors on obesity levels. For example, while exercise frequency is a visible variable, understanding its interaction with caloric intake in predicting obesity requires further exploration and statistical analysis. Additionally, complex relationships between socioeconomic factors and health outcomes need analysis to uncover non-obvious patterns.

Different Perspectives on the Data

To gain a thorough understanding, I plan to explore the data from multiple perspectives:

- **Demographic Analysis**: Segmenting data by age, gender, and socioeconomic status to understand obesity trends within subgroups.
- **Behavioral Patterns**: Examining eating habits and exercise behaviors to determine their role in obesity.
- **Comparing Health Indicators**: Analyzing differences in health indicators like BMI across diet types and exercise categories.

Data Slicing and Dicing

I will slice the data by various categories to gain detailed insights:

- **Age Groups**: Comparing obesity trends across age groups can highlight any age-specific risk factors.
- **Income and Education Levels**: Assessing obesity prevalence across different income and education levels to understand socioeconomic effects.

- **Geographic Location**: If geographic data is available, this could reveal location-based health patterns.

Summarizing Data for Key Questions

I will summarize data by producing:

- **Descriptive Statistics Tables**: Tables showing the mean, median, and count for each variable segmented by obesity levels.
- **Pivot Tables**: Breaking down data by demographics to highlight distribution patterns and relationships within key subgroups.

Visualizations and Tables

Visualizing data is crucial for understanding complex patterns. The following types of plots and tables will be used:

- **Correlation Matrix**: A heatmap showing correlations between variables to highlight significant relationships.
- **Boxplots and Histograms**: To show distributions and central tendencies of continuous variables.
- **Scatterplots**: To examine the relationship between variables like exercise and caloric intake with obesity.
- **Bar Plots**: Summarizing frequency counts for categorical data like diet type and obesity category, providing an easy comparison.

Machine Learning Techniques

I plan to incorporate machine learning techniques where relevant:

- **Predictive Modeling**: Decision trees or logistic regression may be used to predict obesity levels based on variables such as physical activity and dietary habits.
- **Unsupervised Learning**: Techniques like clustering could help reveal subgroups within the data, such as individuals with similar health risk profiles.

Further Questions for Analysis

As I proceed with the analysis, the following questions arise:

- **How do individual dietary factors interact with exercise to influence obesity?**
- **Are there notable differences in obesity rates across different education levels?**
- **What are the best methods for handling any remaining data imbalances or outliers in the dataset?**

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.