Team 18: Obesity Level Classification and Clustering Analysis

1 Introduction

The `18.csv` **Obesity Level Estimation** dataset contains health-related information from individuals in **Mexico, Peru, and Colombia**, with ages ranging from **14 to 61 years**. The dataset consists of **2111 records** and includes **17 attributes** related to eating habits, physical activity, and demographic factors. Data collection was performed via an online survey, allowing anonymous users to provide responses regarding their lifestyle and health conditions.

This dataset enables **both classification and clustering tasks**, allowing for the development of models to predict obesity levels and segment individuals based on their health characteristics. The results can be used for health policy planning, personalized recommendations, and obesity prevention strategies.

2 Dataset Description

The dataset consists of the following key attributes:

2.1 Demographic and Anthropometric Features

- **Gender** – Male or Female.
- **Age** – Age of the individual.
- **Height** – Height in meters.
- **Weight** – Weight in kilograms.

2.2 Eating Habit Attributes

- **Frequent Consumption of High-Caloric Food (FAVC)** – Yes or No.
- **Frequency of Vegetable Consumption (FCVC)** – Scale-based frequency.
- **Number of Main Meals (NCP)** – Number of daily main meals.
- **Consumption of Food Between Meals (CAEC)** – Level of snacking behavior.
- **Consumption of Water Daily (CH20)** – Amount of daily water intake.
- **Consumption of Alcohol (CALC)** – Frequency of alcohol consumption.

2.3 Physical Activity and Lifestyle Attributes

- **Calories Consumption Monitoring (SCC)** – Whether the individual tracks calorie intake.
- **Physical Activity Frequency (FAF)** – Frequency of physical exercise.
- **Time Using Technology Devices (TUE)** – Daily screen time.
- **Transportation Mode (MTRANS)** – Primary mode of transportation.

2.4 Target Variable: Obesity Levels (Classification Task)

- **Underweight** – BMI Less than 18.5.
- **Normal** – BMI between 18.5 and 24.9.
- **Overweight** – BMI between 25.0 and 29.9.
- **Obesity I** – BMI between 30.0 and 34.9.
- **Obesity II** – BMI between 35.0 and 39.9.
- **Obesity III** – BMI higher than 40.

3 Tasks and Requirements

This dataset enables two primary machine learning tasks: classification and clustering.

3.1 Obesity Level Classification (Supervised Learning)

- Train classification models such as **Logistic Regression, Decision Trees, Random Forest, Support Vector Machines (SVM), and Neural Networks** to predict obesity levels.
- Evaluate model performance using **accuracy, precision, recall, and F1-score**.
- Identify the most significant factors contributing to obesity levels.

3.2 Obesity Segmentation (Unsupervised Learning - Clustering)

- Apply clustering algorithms to segment individuals into different health-related groups.
- Use **Elbow Method** and **Silhouette Score** to determine the optimal number of clusters.
- Analyze the characteristics of each cluster to understand obesity trends in different populations.

3.3 Visualization and Reporting

- Generate bar charts and scatter plots to explore feature distributions.
- Create correlation heatmaps to examine relationships between eating habits, physical activity, and obesity.
- Develop clustering visualizations to showcase health-based group segmentation.

4 Submission Requirements

- A well-structured report detailing the methodology, results, and analysis in a given report format.
- Python code is used for implementation.
- A presentation summarizing key findings and recommendations in a given presentation format.