Estimation of Obesity Levels Based on Eating Habits and Physical Condition | HubbleMind

Project Description:

In this project, interns will analyze a dataset containing health and dietary information from individuals in Mexico, Peru, and Colombia. The goal is to estimate obesity levels based on physical condition and eating habits using Python for Data Science. Interns will learn to apply data cleaning, exploratory data analysis (EDA), advanced visualizations, and machine learning techniques to predict obesity levels.

Dataset : Download | Data Source : CC BY 4.0 | UCI Archive

Dataset Description:

This dataset includes data to estimate obesity levels, with 17 attributes and 2111 records. The records are labeled with a class variable `NObeyesdad` (Obesity Level), which categorizes individuals into several obesity levels:

- **Insufficient Weight**
- **Normal Weight**
- **Overweight Level I**
- **Overweight Level II**
- **Obesity Type I**
- **Obesity Type II**
- **Obesity Type III**

Key Features:

1. **Gender** (Categorical)
2. **Age** (Continuous): Age of the individual
3. **Height** (Continuous): Height in meters
4. **Weight** (Continuous): Weight in kilograms
5. **family_history_with_overweight** (Binary): Whether the individual has a family member suffering from overweight
6. **FAVC** (Binary): Whether the individual eats high-calorie food frequently
7. **FCVC** (Continuous): Frequency of vegetable consumption in meals
8. **NCP** (Continuous): Number of main meals consumed daily
9. **CAEC** (Categorical): Food consumption between meals
10. **SMOKE** (Binary): Whether the individual smokes

11. **CH2O** (Continuous): Amount of water consumed daily
12. **SCC** (Binary): Whether the individual monitors calorie intake
13. **FAF** (Continuous): Frequency of physical activity
14. **TUE** (Continuous): Time spent using technological devices
15. **CALC** (Categorical): Alcohol consumption frequency
16. **MTRANS** (Categorical): Mode of transportation used

Target Feature:

- **NObeyesdad** (Categorical): Obesity level categorized into 7 classes.

Week 1: Data Importing and Cleaning

1. **Task 1**: Import the dataset and inspect its structure.
    - Understand the data types and look for any missing values.
2. **Task 2**: Data Type Conversion and Encoding
    - Label encode binary variables like `Gender`, `SMOKE`, and one-hot encode multi-class variables like `MTRANS`, `NObeyesdad`.
3. **Task 3**: Outlier Detection and Handling
    - Detect outliers using boxplots for continuous variables like `Weight` and `Height`, and handle them by capping or transformation.
4. **Task 4**: Normalization/Standardization
    - Normalize continuous variables such as `Age`, `Weight`, `Height` using MinMax scaling.

Week 2: Exploratory Data Analysis (EDA)

1. **Task 1**: Summary Statistics
    - Generate summary statistics for continuous variables (mean, median, mode, etc.).
2. **Task 2**: Distribution Analysis
    - Plot histograms and KDE plots for key variables like `Age`, `Weight`, and `Height` to understand the data distribution.
3. **Task 3**: Relationship Exploration
    - Use boxplots to explore relationships between features (like `Weight`, `FAF`) and obesity levels.
4. **Task 4**: Correlation Analysis
    - Create a correlation heatmap to explore relationships between continuous features like `Height`, `Weight`, and `Age`.

Week 3: Advanced Visualizations and Machine Learning

1. **Task 1**: Advanced Visualizations

    - Create pair plots, feature importance plots (for Random Forest), and a heatmap of the confusion matrix.
2. **Task 2**: Feature Engineering and Scaling
    - Perform any necessary feature scaling and ensure all features are encoded properly for machine learning.
3. **Task 3**: Train-Test Split
    - Split the data into training (80%) and testing (20%) sets.
4. **Task 4**: Machine Learning Model Implementation
    - Implement **Logistic Regression** and **Random Forest** to predict obesity levels.
5. **Task 5**: Model Evaluation
    - Evaluate the models using metrics like accuracy, precision, recall, and F1-score. Compare the performance of both models.

Week 4: Model Evaluation and Reporting

1. **Task 1**: Model Evaluation Report
    - Summarize the performance of the models, focusing on key metrics such as accuracy, precision, recall, and F1-score for each class.
2. **Task 2**: Documentation
    - Write a report summarizing the entire project, including the dataset description, data preprocessing, EDA, model building, and evaluation. Discuss the insights gained from the analysis and visualizations.
3. **Task 3**: Final Project Submission
    - Prepare all the code, datasets, visualizations, and documentation in an organized manner for final submission.
4. **Submit your work:** Submit the document via the provided **Google Form**

Deliverables:

- **Code**: Clean, well-commented code for data preprocessing, EDA, and machine learning models.
- **Visualizations**: All relevant visualizations such as boxplots, heatmaps, and feature importance.
- **Report**: A comprehensive project report detailing the process and insights.