Obesity Prediction Project

The goal of this project is to develop a machine learning model that predicts obesity levels based on lifestyle habits, family history, and physical conditions. This system will help in understanding the key factors contributing to obesity and enable early intervention strategies.

Dataset Overview

The dataset contains 2111 divided to 1900 train data and 211 for test data. It is collected from individuals in Mexico, Peru, and Colombia. It includes 16 lifestyle and health-related features, with obesity levels classified into different categories ranging from underweight to various obesity types.

Features:

- **Gender**: Male or Female.
- **Age**: The person's age in years.
- **Height**: Height in meters.
- **Weight**: Weight in kilograms.
- **family_history_with_overweight**: Whether the person has a family history of being overweight (yes/no).
- **FAVC**: If the person frequently consumes high-calorie foods (yes/no).
- **FCVC**: Frequency of vegetable consumption (scale from 1 to 3).
- **NCP**: Number of main meals per day.

- **CAEC**: Frequency of consuming food between meals (Never, Sometimes, **Frequently**, Always).
- **SMOKE**: Whether the person smokes (yes/no).
- **CH2O**: Daily water intake (scale from 1 to 3).
- **SCC**: If the person monitors their calorie intake (yes/no).
- **FAF**: Physical activity frequency (scale from 0 to 3).
- **TUE**: Time spent using technology (scale from 0 to 3).
- **CALC**: Frequency of alcohol consumption (Never, Sometimes, Frequently, Always).
- **MTRANS**: Main mode of transportation (Automobile, Bike, Motorbike, Public Transportation, Walking).

Target Variable:

- **NObeyesdad:** Categorized as Insufficient Weight, Normal Weight, Overweight Level I, Overweight Level II, Obesity Type I, Obesity Type II, Obesity Type III

Requirements:

Data Preprocessing:

- Handle missing values, encode categorical variables, and normalize numerical features.
- Implement feature selection techniques to improve model performance.
- *Bonus:* Use advanced preprocessing techniques and feature engineering not covered in labs.

Exploratory Data Analysis (EDA):

- Visualize data distribution and correlations among features.
- Identify key factors contributing to obesity levels.
- Use statistical and graphical analysis to interpret the features impact on obesity.

Model Development:

- Train and compare **at least 4** machine learning models, such as:
    - Logistic Regression, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Gradient Boosting (XGBoost), Neural Networks… etc.
- Evaluate models using:
    - Accuracy, Precision, Recall, F1-score, and Confusion Matrix.

Reporting (PDF):

- Document all preprocessing steps, including missing value handling, encoding, scaling, and feature selection.
- Present and interpret relevant data visualizations.
- Explain relationships between lifestyle factors and obesity levels.
- Detail model selection, hyperparameters, and tuning approaches. Show trials for at least 2 hyperparameters change's and their effect on model performance.
- Evaluate and compare models using metrics such as accuracy, precision, recall, F1-score, and confusion matrix. Discuss the strengths and weaknesses of each model.
- Summarize your work and write a conclusion.

Bonus Tasks:

- Implement different feature extraction techniques to improve model interpretability.
- Develop at least one machine learning model from scratch and understand it very well.
- Use powerful new models not covered in the labs.
- Develop a user-friendly desktop application using Tkinter, PyQt or any other library you know. Allow users to input lifestyle attributes and predict obesity levels along with confidence scores.
- Extend the application to include personalized health recommendations based on predictions.

Mentor

TA. Maryam Alsawaf

Email: maryam.sawaf@cis.asu.edu.eg