**PROJECT PROPOSAL**

**Title**: Prediction and Analysis of Obesity Levels Based on Lifestyle and Physical Attributes

**Name:** Tejalakshmi Gangadhar, Chalumuri

**Data Mining:** CS 613, CRN: 21453

**Professor:** D. Paul Benjamin

**Introduction:**

This proposal outlines the approach for a final project aimed at predicting obesity levels using data mining techniques. The objective is to apply classification algorithms to understand how various factors like eating habits, physical activity, and personal attributes correlate with different obesity levels.

**Dataset Description:**

* The dataset to be analyzed includes data from individuals in Mexico, Peru, and Colombia, focusing on their eating habits, physical condition, and obesity levels.
* The dataset comprises 17 attributes and 2111 records, with a detailed distribution of both synthetically generated data (77% via Weka and SMOTE) and real user inputs (23%).
* The attributes cover a range of factors from dietary habits to physical activity and technological usage. The dataset is available at [Obesity Levels](https://www.kaggle.com/datasets/fatemehmehrparvar/obesity-levels)

**Problem Statement:**

* The primary goal is to predict the obesity level (classified into categories such as Insufficient Weight, Normal Weight, Overweight, and different types of Obesity) based on predictive modeling.
* This analysis can help in understanding the impact of lifestyle choices on obesity, which is critical for designing targeted health interventions.

**Data Mining Techniques:**

* **Classification Learning**: We will employ three different algorithms to build robust predictive models:
  + **J48**: A decision tree algorithm which is effective for categorical target variables.
  + **Random Forest**: An ensemble method known for its high accuracy and control over overfitting.
  + **Logistic Regression**: A statistical model that estimates the probabilities of binary or multinomial outcomes.
* These methods were selected to compare their efficacy in handling different types of data distributions and their ability to model nonlinear relationships.

**Data Transformation and Preparation:**

* **Normalization**: All continuous variables such as age, height, weight, and daily water intake will be normalized to ensure that the scale of these variables does not bias the models.

**Model Validation and Testing:**

* The dataset will be split into a training set (80%) and a test set (20%).
* The models will be evaluated using cross-validation within the training set to fine-tune parameters and select the best-performing model.
* Performance metrics such as accuracy, precision, recall, and F1-score will be used to assess each model.

**Expected Outcomes:**

We expect to identify which factors are most predictive of obesity levels and to understand the relative performance of different data mining algorithms in this context. This will provide insights into how lifestyle choices impact obesity, potentially informing public health strategies.

**Conclusion:**

This project aims to utilize advanced data mining techniques to provide deeper insights into the factors influencing obesity. By applying multiple classification algorithms, we expect to develop a predictive model that can serve as a tool for understanding and combating obesity.