### **Introduction**

This project focuses on predicting Body Mass Index (BMI) using a dataset related to eating habits, physical activity, and demographic factors. BMI serves as an essential indicator of health, and the goal is to determine the factors that most significantly influence it. The task involves building and comparing regression models, including Linear Regression, Ridge Regression, and Random Forest, to evaluate their performance in predicting BMI. Among these, the Random Forest model demonstrated the best predictive capability, highlighting its effectiveness in capturing complex relationships within the data.

### **Data Description**

The dataset, titled **"Estimation of Obesity Levels Based on Eating Habits and Physical Condition"**, was sourced from the UC Irvine Machine Learning Repository. It includes 17 variables and 2,111 instances, with features spanning binary, categorical, and continuous data. Key attributes include age, physical activity frequency, vegetable consumption, water intake, smoking habits, alcohol consumption, and family history with overweight. The target variable, BMI, was derived from height and weight values included in the dataset.

### **Models and Methods**

The analysis began with exploratory data analysis (EDA) to identify distributions, relationships, and potential outliers in the dataset. Models were trained on an 80/20 train-test split. A Linear Regression model established a baseline for performance, followed by Ridge Regression to address multicollinearity. Finally, a Random Forest model was implemented to capture non-linear relationships and feature interactions. All models were built using pipelines to streamline preprocessing, including scaling numerical features and one-hot encoding categorical variables. Hyperparameter tuning was applied to the Ridge and Random Forest models using GridSearchCV to optimize their performance.

### **Results and Interpretation**

The baseline model, which predicted the mean BMI for all cases, showed poor performance with an **R² near 0** and a high **MSE of 66.06**. The Linear Regression model improved on this with an **R² of 0.3666** and **MSE of 41.84**, but it struggled with multicollinearity and non-linear relationships. Ridge Regression addressed these issues, achieving slightly better results, with optimized regularization improving model stability. However, the Random Forest model outperformed all others, with an **R² of 0.8507** and a significantly lower **MSE of 9.86**, demonstrating its strength in capturing complex patterns in the data.

### **Conclusion and Next Steps**

This analysis highlights the Random Forest model as the most effective tool for predicting BMI in this dataset. Key factors influencing BMI include physical activity frequency, age, and family history with overweight, which consistently emerged as the most important predictors across models. Future work could involve incorporating more granular data, such as detailed dietary information, wearable device metrics, or environmental factors, to refine predictions further. Additionally, exploring ensemble methods or hybrid models could yield even greater predictive accuracy and insights into BMI determinants.