Predicting Obesity from Health and Lifestyle Data

📌 Project Overview

This project aims to build predictive models that estimate the likelihood of a person being obese based on their health metrics, physical activity, and demographic characteristics. The analysis is based on data from the NHANES dataset and includes both **interpretable models** (like logistic regression) and **black-box models** (like random forests).

🔍 Research Question

Can we predict whether an individual is obese using a combination of health, demographic, and lifestyle data?

📊 Variables Used

- **Target Variable**: Obesity (defined as BMI > 35)
- **Predictor Variables**:
    - Demographics: Age, Gender, Marital Status
    - Clinical Measures: Weight, Height, BMI, Blood Pressure, HDL, Total Cholesterol
    - Lifestyle: Sedentary minutes, work activity, recreational activity, and walking or biking

⚙️ Methodology Summary

Two predictive models are developed:

1. **Logistic Regression**: Useful for interpretability, this model provides odds ratios to understand how each predictor affects the likelihood of obesity.
2. **Random Forest**: A machine learning approach that captures non-linear patterns and ranks variable importance, offering potentially better predictive performance.

The data is split into training and testing sets, and each model is evaluated using:

- **Accuracy**
- **AUC (Area Under the Curve)**
- **Confusion Matrix**
- **Variable Importance (for random forest)**

🧠 Interpretation

Logistic regression helps us explain the relationship between obesity and predictors (e.g., more sedentary time = higher odds of obesity), while random forest helps uncover interactions and non-linear relationships. Comparing the two offers insight into the trade-off between accuracy and interpretability.