**Assignment Cover Page**

|  |  |
| --- | --- |
| **Subject Code** | COSC2999 |
| **Subject Name** | Practical Data Science with Python |
| **Location** | SGS |
| **Title of Assignment** | Assignment 2: Data Modelling |
| **Student Name** | Nguyễn Đình Đăng Nguyên |
| **Student Number** | S3759957 |
| **Lecturer Name** | DR. Thuy Nguyen |
| **Assignment due date** | December 2nd, 2024 |
| **Assignment submission date** | December 2nd, 2024 |
| **Numbers of pages including this one** | 15 |
| **Word count** |  |

*I declared that in submitting all work for this assessment I have read, understood and agree to the content and expectations of the Assessment Declaration*

# Abstract

Obesity is a medical condition characterized by an excessive accumulation of body fat, which can negatively impact an individual’s health. This study aimed to classify obesity levels using two machine learning approaches (Classification and Regression), with a focus on physical activity, nutritional habits, and genetic factors. The study utilized an observational design, collecting data from the UCI repository [ref] through a web-based survey to assess participants' eating habits and physical activity levels. The dataset included variables such as gender, age, height, weight, family history of obesity, dietary patterns, and physical activity frequency. For the modeling process, three classification algorithms were employed to predict obesity levels, including Random Forest (for both classification and regression tasks), Extreme Gradient Boosting (for classification), and k-nearest neighbors with grid search optimization for hyperparameter tuning. Model performance was evaluated using various metrics: accuracy, recall, precision, F1-score, area under the curve (AUC), and precision-recall curve for classification tasks, and Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R²) for regression tasks. The Random Forest model exhibited the most robust performance, and feature selection was shown to improve model efficiency. These findings emphasize the critical role of physical activity and nutritional habits in addressing the growing obesity epidemic.

# Introduction

The rise in obesity has become a significant public health concern worldwide. According to WHO, over the past few decades, obesity rates have steadily increased, driven by factors such as genetics, environment, diet, physical activity levels, and behaviors [ref]. Poor diet (high in calories), lack of physical activity, genetics, and certain medical conditions or medications can all contribute to obesity.

Addressing obesity is not only essential for individual well-being but also for reducing the overall burden on healthcare systems and improving societal health outcomes. Immediate and sustained efforts to prevent obesity are necessary to ensure healthier futures for individuals and communities around the world. Preventing obesity is crucial to reversing these trends and improving public health. Prevention strategies should focus on promoting healthier eating habits, increasing physical activity, and fostering environments that support healthier choices.

Obesity levels are commonly calculated using **Body Mass Index (BMI)**, a simple and widely used measure that helps categorize individuals based on their weight relative to their height. BMI is calculated using the following formula [ref]:

Where:

* Weight is in kilograms (kg)
* Height is in meters (m)

After all calculation was made to obtain the mass body index for each individual, the results were classified as [ref]:

* Underweight Less than 18.5
* Normal 18.5 to 24.9
* Overweight 25.0 to 29.9
* Obesity I 30.0 to 34.9
* Obesity II 35.0 to 39.9
* Obesity III Higher than 40

This report will outline the efficiency of using Machine Learning methodologies in detecting and preventing obesity based variables such as gender, age, height, weight, family history of obesity, dietary patterns, and physical activity frequency provided by UCI’s “Estimation of Obesity Levels Based On Eating Habits and Physical Condition” dataset [ref]

# Methodology

## Data Preparation and exploratory data analysis (EDA)

## Machine learning algorithms

**Classification**

Random Forest

XGBoost

**Regression**

KNN

Random Forest

## Grid search optimization and hyper-parameter tuning

**Classification**

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Hyper-parameter** | **Values** | **Best group** |
| Random Forest | 'max\_depth' | 3, 5, 7, 10, None | None |
| 'min\_samples\_split' | 2, 5, 10 | 2 |
| 'min\_samples\_leaf' | 1, 2, 5 | 1 |
| 'max\_features' | None, 'sqrt', 'log2' | 'sqrt' |
| XGBoost | 'n\_estimators' | 50, 100, 200 | 200 |
| 'max\_depth' | 3, 5, 7, 10 | 10 |
| 'learning\_rate' | 0.01, 0.1, 0.05 | 0.05 |
| 'subsample' | 0.5, 0.7, 0.8, 1.0 | 0.8 |

**Regression**

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Hyper-parameter** | **Values** | **Best group** |
| KNN | 'n\_neighbors' | 3, 5, 7, 9, 11 | 5 |
| 'weights' | 'uniform', 'distance' | 'distance' |
| 'metric' | 'euclidean', 'manhattan' | 'manhattan' |
| Random Forest | 'n\_estimators' | 100, 200, 300 | 200 |
| 'max\_depth' | 10, 20, 30, None | None |
| 'min\_samples\_split' | 2, 5, 10 | 2 |
| 'min\_samples\_leaf' | 1, 2, 4 | 1 |
| 'max\_features' | None, 'sqrt', 'log2' | 'sqrt' |

# Results