Introduction

For this project, I have selected a study on obesity named "ObesityDataSet\_raw\_and\_data\_synthetic.csv." This dataset contains various features that could be used to predict the possibility of a person becoming obese, based on factors such as age, weight, and family history of obesity. According to a 2022 study by the World Health Organization (WHO), approximately 2.5 billion adults worldwide were estimated to be overweight, with 890 million being obese.

The goals for this project are to identify which features from this dataset contribute to an individual becoming obese and to utilize prediction and classification algorithms to determine what percentage of the study sample are most at risk of becoming obese.

Firstly, we will conduct Exploratory Data Analysis (EDA) to identify any anomalies within the dataset. Then, we will identify the most important feature in the dataset and designate it as our target variable (X). We will explore the best approach for choosing which classification algorithm to use and the reasons behind this decision. If necessary, we will implement data processing techniques such as Principal Component Analysis (PCA).

Throughout this project, we will employ various methods for characterizing the data. We will utilize cross-validation techniques and data visualization to explain the variances and argue from a statistical perspective which features are more important than others.

https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight#:~:text=In%202022%2C%201%20in%208,million%20were%20living%20with%20obesity.