**Background**

The aim of our project was to explore the relationship between nutrition and disease to better understand how dietary factors influence health outcomes. By examining whether nutritional data can be used to predict disease status, we wanted to identify potential applications for improving public health and developing preventative strategies. To achieve this, we compared the performance of various classification models in predicting health conditions based on dietary intake data.

More specifically, we used data to answer these two questions:

1. Can we reliably predict diagnosis of the outcome conditions from mean daily nutrient intake?
2. Which classification models predict with the best performance?

We used a Naïve Bayes classifier, a Tree and Random Tree model, and then finally, a Neural Network model to answer these questions.

**Dataset**

The data used for this project was the National Health and Nutrition Examination Survey (NHANES), which is a long-running program conducted by the CDC. NHANES is designed to assess the health and nutritional status of adults and children in the United States. This survey has been conducted in various forms since the early 1960s and is now an ongoing program with continuous data collection, but we decided to just take the dataset from 2021-2023.

Participants are selected using a complex probability sampling method to ensure that the results are representative of the U.S. population. NHANES has five different sections in total: demographic data, dietary data, examination data, laboratory data, and questionnairee data. For our project, we only used the dietary data and the questionnaire data.

We chose the eight most frequently occurring conditions as our response variables: asthma, hay fever, arthritis, congestive heart failure, coronary heart disease, heart attacks, thyroid problems, and cancer. For our predictor variables, we kept all the variables that summarized the total dietary intake of the individuals within one day, which encompasses broad categories like total fats, carbohydrates, and fiber, as well as more specific nutrients such as vitamin B-12, caffeine, and magnesium

To clean the data, we selected all the relevant columns from the dietary data and the questionnaire data, removed all the NA values, and then used the participant sequence number to match the people across the datasets.

Reference: <https://wwwn.cdc.gov/nchs/nhanes/>