**Project Topic:** Predicting Diabetes based on food intake on daily basis.

For this project, I will use data from CDC (Centers for Disease Control and Prevention) site, specifically from The NHANES program (National Health and Nutrition Examination Survey). The data is related to daily nutrition intake in day-to-day life, demographic details, and individual’s laboratory results. Based on the data, I will predict the most influential factor for diabetes.

In this capstone project, the goal is to analyze the influence of various factors on diabetes using data from the National Health and Nutrition Examination Survey (NHANES) program, obtained from the Centers for Disease Control and Prevention (CDC). The dataset comprises information on daily nutrition intake, demographic details, and laboratory results of individuals. By exploring this data, I aim to identify the most significant factor contributing to the development of diabetes. This analysis will provide valuable insights into understanding the relationship between lifestyle, demographic characteristics, and the occurrence of diabetes, ultimately aiding in the prevention and management of this chronic condition.

**Why this project topic:**

The choice of this project topic is motivated by the increasing prevalence of diabetes worldwide and its significant impact on public health. Diabetes is a chronic condition that affects millions of individuals and can lead to various health complications if not managed effectively. By studying the influential factors for diabetes, we can gain valuable insights into the underlying causes and potential risk factors associated with the disease.

The NHANES program conducted by the CDC provides a comprehensive dataset that includes information on nutrition, demographics, and laboratory results. This dataset offers a unique opportunity to explore the relationships between these factors and the development of diabetes. By analyzing this data, we can identify the most influential factors that contribute to the occurrence of diabetes, which can guide preventive strategies, intervention programs, and public health policies.

The insights gained from this project can help healthcare professionals, policymakers, and individuals make informed decisions about lifestyle modifications, early screening, and targeted interventions to reduce the incidence of diabetes. Furthermore, understanding the most significant factors for diabetes can contribute to the development of personalized medicine approaches and tailored interventions, improving the management and treatment outcomes for individuals living with diabetes.

Overall, this project has significant relevance and potential to make a positive impact on public health by identifying and understanding the influential factors associated with diabetes.

This project could involve following steps:

1. Data Collection: Collect Demographic data, nutrition data, laboratory result data of individuals.
2. Data Processing: Clean and process the data for use with feature engineering techniques.
3. Model Selection: Validate the available data to choose the best model for predicting the readmission to hospital.
4. Model Training: Train the selected model with available set of data through different techniques.
5. Model Evaluation: Evaluate the model accuracy.

The other area which may not be part of the project is deployment. As part of it, the finalized model should be integrated with the current clinical system to do the prediction and provide the required recommendations.

Overall, this project would demonstrate the potential of machine learning to improve the healthcare outcomes and reduce the cost by identifying the patients who are at high risk of readmission.

**Data Sources:**

<https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?cycle=2017-2020>

Demographic Data - <https://wwwn.cdc.gov/nchs/nhanes/search/datapage.aspx?Component=Demographics&Cycle=2017-2020>

Dietary Data - <https://wwwn.cdc.gov/nchs/nhanes/search/datapage.aspx?Component=Dietary&Cycle=2017-2020>

Examination Data - <https://wwwn.cdc.gov/nchs/nhanes/search/datapage.aspx?Component=Examination&Cycle=2017-2020>  
Laboratory Data

<https://wwwn.cdc.gov/nchs/nhanes/search/datapage.aspx?Component=Laboratory&Cycle=2017-2020>

Questions:

1. Is there any relevant data source which can be used? Considering the data may have data related to compliance (PII, HIPPA) and may need some specific approval.
2. What are your suggestions related to topic selection and the most appropriate method to approach it?