**DSC 510 Project Proposal**

**Project Team:** William Ranick, Nibedita Bal, Hima Spandana Barla

**Data:** National Health and Nutrition Examination Survey (NHANES)

**Subject Area or Field of Interest**: Health and Nutrition

**Source of Data (provide a link to data):**

<https://www.kaggle.com/cdc/national-health-and-nutrition-examination-survey>

**Specific dataset(s):** There are 6 datasets and they are Demographic, Diet, Examination, Labs, Medication and Questionnaire. Each of the 6 datasets uses the same patients (their ID numbers are the same between different datasets).

**Description of its scope (# metric variables, #categorical variables, #samples, multiple**

**related tables?) :**

Each of the 6 datasets has between 13 and 424 attributes and around 10,000 observations. Also, every dataset has a mixed number of categorical and numerical variables depending on the dataset. For example, the “Labs” dataset contains 424 variables, the vast majority of which are numerical.

**Explain your research questions? :**

*The below questions were formulated based on initial analysis of the available data. As the data is incredibly diverse, there are multiple topics to explore. After further exploration, more concise questions based on a single topic will be generated.*

1. Using clustering on the examination dataset how many wellness groups(healthy, unhealthy) can be formed?

2. Using Questionnaire dataset finding out the factors responsible for the different wellness buckets.

3. How well can digestive cancer diagnosis (esophageal, stomach, colorectal, etc.) be predicted from the diet and lab findings of patients?

4. How well can diabetes diagnosis be predicted from the diet and lab findings of patients?

**Technology group plans to use for Project (i.e. Python, R, SPSS, Tableau, etc.) o How do you plan to use the technology?**

We will useR, Python, SAS, and Tableau for this project. We will cover exploratory data analysis, visualization, and predictive analytics for forecasting future events.

**Journal articles :**

1. Logistic Regression based Feature Selection and Classification of Diabetes Disease using Machine Learning Paradigm

<http://www.ru.ac.bd/stat/wp-content/uploads/sites/25/2020/03/ICDSSDG_COR-2019_paper_11F.pdf>

# 

# Trends in Healthy Eating Index Scores Among Women of Varying Ethnic Backgrounds with Gynecological and Breast Cancers

<https://academic.oup.com/cdn/article/4/Supplement_2/346/5845822>

1. Plan and Operation of the Third National Health and Nutrition Examination Survey, 1988-94

<https://books.googleusercontent.com/books/content?req=AKW5Qafe611oEs2hwHblaiEfskA5K541PRyYVcCfKC-wrwZ_QH3RPFD1Y_Pm3r6I4Na2twLcg08OkB9pnWdSMIDeyKyOuKPj2ytnUxAUWF7wbh08H_3ZPi6Q44l_NPW34FOe_CJMQnXzYi5ofa9KZQlY1TTDHS3TPilmJp66kligl5yOQNkeBB33yMDvB6TXZ40G_zFPIos7jkVVlRF-Ads_QUkF3vRix0HXj3QumXyOar0JHWeizmzi6SQrGRRQ3uFso_oF_P0lBxD2h7NunBu7soMbo5qxfjblsi69jTKBdacUmau4nXI>

1. Dietary Intake Source Data, United States, 1971-74

<https://books.googleusercontent.com/books/content?req=AKW5QafxvPQa8oLe8qShy59XX9CfoWNzgU2Ci6cIdFm2L91B7J9Vefft29f4cDEfE05TGdOASt4kKg7MDBg5mZp66HAi5nZL5wGrksBVe1pAnRs4uy-nR-eo7UOq7kJuFbP1zMqNrafCdOO6btA60x6J5kj04dMDAHYJmjCNm3wUaA_2QDPxLbyTvzn2vUsi9g0SKl5Xqe3kp7kQGT3J0UV8n8ib459B0h-IEYGwI7H5MhW89UMKpJtTmxBUKfu4VJjQqAbe5fcjKx2gb1rZZ-nUb0kXoV4wnKzQI-RQQZQrd-mlpmwn5Ow>

1. Activity-based nutrition management model for healthcare using similar group analysis.

<https://doi-org.ezproxy.depaul.edu/10.3233/THC-191731>

6).Assessment of Functioning of Village Health Nutrition and Sanitation Committees in a District in Maharashtra.

<https://doi-org.ezproxy.depaul.edu/10.4103/ijcm.IJCM_171_17>