Nutrient Intake Profile and Patterns with Respect to Gender and Income(NHANES)

Venkat Aashish Mathsyaraja
Computer and Information Sciences
University of Delaware
Newark, DE
aashishm@udel.edu

Abhay Hemanth
Computer and Information Sciences
University of Delaware
Newark, DE
abhayh@udel.edu

Michela Taufer
Computer and Information Sciences
University of Delaware
Newark, DE
taufer@udel.edu

***Abstract*—The NHANES dietary dataset captures data from a large diverse demographic. We use this useful data to capture nutrient intake profiles with respect to key demographic attributes such as gender and income to poverty ratio. Through a series of steps involving preprocessing, clustering and visualizing by the use of plots, we successfully show the excess consumption of harmful macronutrients and the deficiency of useful macronutrients.**

***Keywords*—NHANES, clustering, preprocessing, visualizing**

I. INTRODUCTION

Nutrient intake is a key aspect in the health of a human being. A balanced diet can go a long way in terms of staying healthy and thereby clear of diseases. The NHANES dietary dataset is a large, diverse dataset which captures dietary intake and other general demographic information. In this work, we examine the nutrient intake profiles of the surveyed population and determine the harmful macronutrients which are being consumed in excess and the useful macronutrients which are being consumed in deficient amounts. We capture this information with respect to features such as gender, income to poverty ratio and compare and contrast the key differences.

The tasks accomplished are:

- Preprocessing and handling the missing data.
- Clustering with DBSCAN and KMEANS for patterns
- Visualizing and observing results.

II. WORKFLOW

A. Preprocessing and handling missing data

First, the raw nutrient data and demographic data are cleaned by splitting header information and acquiring the necessary features. Next, missing data is removed from both datasets as we have copious amounts of data which is not missing and it can add complexity for the analysis. Next, we split the cleaned demographic data with respect to whether an instance is above or below poverty line and also with respect to a gender. Next this data is joined with the cleaned nutrient data to give us the nutrition information of people above poverty line and below poverty line separately. This is used for clustering and then to draw out some statistics.

B. Clustering and Visualizing the Joined datasets

The previously cleaned and joined data is clustered first using DBSCAN, where all the macronutrients and the sequence number are clustered. Next, the carbohydrates and fats are clustered using K-means. The clusters are observed and finally the nutrient intake profile are plotted using bar-plots and pie-charts.

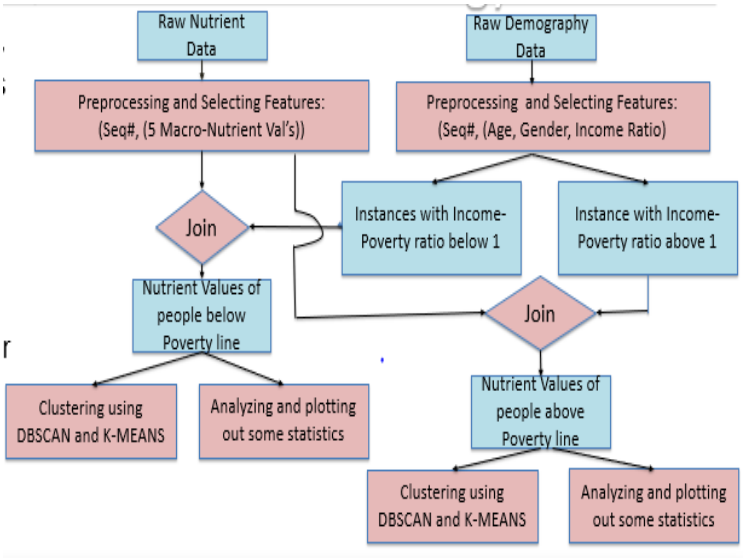


Fig 1: Workflow diagram

III. CLUSTERING USING DBSCAN AND K-MEANS

The 6-dimensional (Sequence number and 5 macro nutrients) cleaned data is clustered with respect to income to poverty ratio using DBSCAN. The DBSCAN parameters epsilon is taken as 10 and minimum points as 4. These optimal parameters were chosen by making a trade off between the number of clusters and amount

of points that are clustered.

With the parameters mentioned, for above poverty 40 clusters were obtained with 53.39% of the data points clustered and 25 clusters for below poverty with 52.3% of the data points clustered

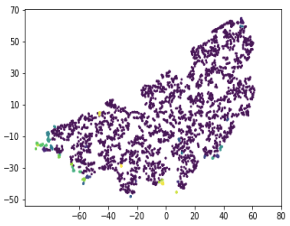


Fig 2: DBSCAN intake - above poverty line.

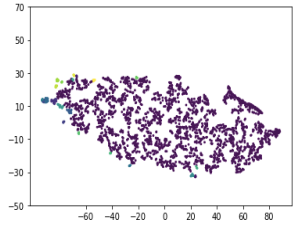


Fig 3: DBSCAN intake - below poverty line

. Next, 2-dimensional cleaned data (carbohydrate and fat) are clustered with respect to gender using K-means to get a clear graphical understanding of two major macro-nutrients and the observations are noted. Below are captured clusters.

x-axis- grams of Carbs
y-axis-grams of Fat

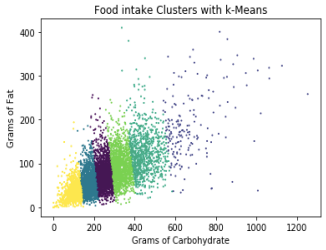


Fig 4: K-means carb-fat intake- male

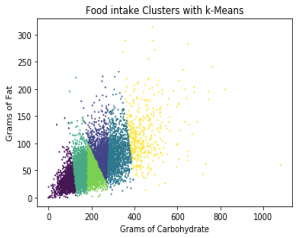


Fig 5: K-means carb-fat intake- female

IV. VISUALIZATION AND RESULTS

In our work, we considered anything above 110 grams of sugar and 100 grams of fat a day as dangerous. Anything less than 56 grams of protein and 20 grams of fiber is considered deficient. We have found out that about 40% of the population are eating excessive amounts of sugar on a daily basis, and about 20% are eating excessive amounts of fat irrespective of their income level and gender.

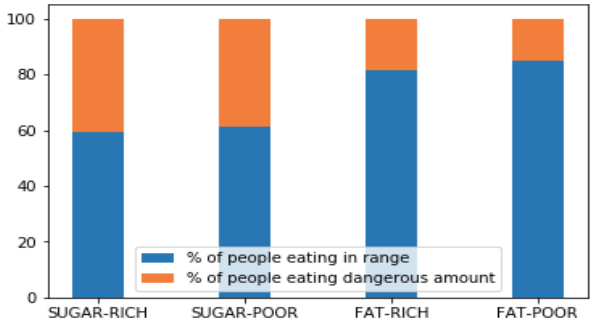
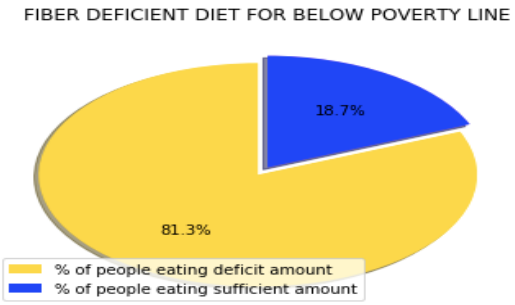
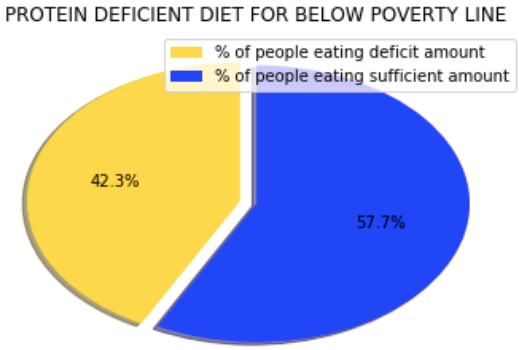


Fig 6: sugar and Fat patterns of people based on their income

There is a massive deficit of fiber in the diet of people below poverty line, about 80% of people.



And about 70% of the population above poverty line are consuming less than the required amount of fiber per day. About 38% of the population below poverty line are consuming deficit amounts of protein.



V. CONCLUSION

About 30% of the U.S population are obese and about 10% of the population are type 2 diabetic, this number increases to 25.2% above the age 60. The over intake of the macro-nutrients such as fat and sugar in a long run might be the major cause of these diseases. By clustering and drawing out these statistics from the diverse NHANES dataset, we found out some unusual food patterns which concludes that majority of the population are not aware of their nutrient intake.

REFERENCES

[1] Gregory L Austin, "Carbohydrate, fat, and protein intakes and association with energy intake in normal-weight, overweight, and obese individuals" The American Journal of Clinical Nutrition, Volume 93, Issue 4, 1 April 2011.

[2] Mikael Fogelholm , "Dietary macronutrients and food consumption as determinants of long-term weight change in adult population.