Kira Glasmacher
BIS/CBB 634
Final Project Report

# Insights from NHANES Data to Improve Nutritional Literacy

## Introduction

Nutrition is an essential determinant of health and well-being, influencing everything from metabolic processes to chronic disease risk. Yet, despite its importance, access to adequate nutrition remains uneven across the United States. A large number of households in the United States experience food insecurity, with marginalized communities disproportionately affected. These inequities not only compromise immediate nutritional needs but also contribute to long-term disparities in health outcomes, including elevated rates of conditions such as diabetes, heart disease, and obesity. Addressing these disparities is a critical public health challenge that requires innovative solutions capable of analyzing and mitigating the multifaceted factors underlying nutritional inequities.

The relationship between nutrition, health, and individual lifestyle factors are very complex, and variables such as dietary intake, demographics, and environmental exposures interact in ways that are not necessarily direct causal pathways. Hence, it is difficult for many people to fully understand how nutrition and lifestyle choices affect their health. Often the education differences about nutrition among different demographics may even exarcebate disparities in food security in addition to accessibility, also widening the divide in health and disease between demographics.

On another note, behaviors such as the consumption of dietary supplements, alcohol, or even chewing gum can significantly impact biomarkers used in clinical assessments, introducing potential biases or inaccuracies in lab results. For instance, alcohol consumption may skew liver enzyme readings or alter blood sugar levels, while certain dietary supplements can influence nutrient profiles or even interfere with medications. However, lab results are often used to make clinical decisions in many different scenarios. Obscured results due to patients breaking fasting advice before a blood draw could therefore lead to potentially inaccurate clinical decisions. This challenge highlights the need for tools that not only analyze labs and measurements but may also account for factors such as nutritional intakes that reflect some behavioral structure that could predict fast breaking in a cohesive framework.

This report describes an API developed to address these challenges by providing insights into the National Health and Nutrition Examination Survey (NHANES), a dataset maintained by the National Center for Health Statistics (NCHS). The API offers functionality for analyzing nutritional values and demographic factors, exploring their relationships, and providing predictive insights. One of its key features is the ability to predict whether individuals consumed dietary supplements, gum, or alcohol prior to a blood draw. This predictive capability could help clinicians contextualize lab results and make more informed decisions, either by adjusting their interpretations or by reinforcing pre-blood draw fasting recommendations.

Better education and understanding of nutrition are crucial for creating healthier communities and addressing widespread health disparities. Empowering individuals with knowledge on this topic can significantly influence their ability to make informed decisions that enhance well-being and prevent chronic diseases. The complexities of nutrition science often make this knowledge inaccessible to the general population. Hence, this API aims to provide an easily interpretable interface with data-driven insights to educate users about nutritional habits of people in the United States. By integrating demographic, behavioral, and nutritional data, it provides clear visualizations and predictive analyses, enabling users to explore factors like supplement use, alcohol consumption, or nutrient deficiencies.

# Methods

## Dataset Description and Acquisition

This project utilizes data from NHANES, a comprehensive resource provided by the NCHS. NHANES offers publicly available datasets that capture detailed information about the nutritional and health status of individuals across the United States. The datasets used in this analysis were obtained directly from the official NCHS repository (https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?Cycle=2021-2023), in compliance with the NCHS Data User Agreement (https://www.cdc.gov/nchs/policy/data-user-agreement.html).

To ensure a robust foundation for analysis, all available datasets from August 2021 to August 2023 were initially downloaded and merged. Following a thorough inspection of missing values and relevance to the study's objectives, the analysis focused on three key datasets: demographics data (DEMO), which provides variables like age and household size; dietary information (DR1IFF), capturing detailed nutrient intake data within a 24 hour period; and the fasting questionnaire (FASTQX), which records pre-blood work fasting behaviors. This targeted selection ensures the datasets align with the project's goals of understanding nutrition-related health outcomes while maintaining data integrity.

The NHANES datasets align with the FAIR principles, ensuring transparency and reproducibility. Hosted on the NCHS website, the data are accompanied by comprehensive documentation and intuitive search tools, facilitating seamless discovery and navigation of data. These public-use files are freely downloadable without significant restrictions, enabling broad accessibility. Provided in standardized XPT format, the datasets are compatible with various analytical tools and software. Additionally, extensive metadata in HTML format are available for each dataset, detailing variables and data collection methodologies.

## Data Preprocessing and Standardization

The data underwent a thorough cleaning and preprocessing process to ensure its quality and relevance for this analysis. Each XPT file was imported into a Pandas DataFrame and merged using SEQN (Sequence Number) as the common identifier, maintaining the alignment of data across datasets. To ensure accurate interpretation, variable definitions and units were cross-referenced with the NHANES documentation.

Columns with over 6,000 missing observations, predominantly those unrelated to nutritional or demographic data, were excluded to focus on variables most relevant to the analysis. Observations missing critical nutritional or demographic information were also removed, as these gaps often indicated that the corresponding surveys were not administered to those respondents. This approach minimized bias while retaining respondents with complete and meaningful data for the study.

Outliers were not removed, since closer inspection revealed that outliers likely represent genuine variations rather than errors or anomalies in the data. Therefore, these data points were preserved to reflect the true variability within the population. Additionally, no initial standardization of values was performed, as most laboratory results and measurements in NHANES datasets are inherently standardized to specific scales. This preserved the integrity and interpretability of the raw data while ensuring compatibility with subsequent analytical methods.

## Analysis

The API includes several analytical tools to enhance the understanding of the NHANES dataset, aiming to make it accessible to all users and not just clinical researchers or public health professionals.

Summary statistics serve as the foundation for data exploration, offering key descriptive measures such as mean, range, and other statistical summaries for numerical variables. To complement these statistics, visualizations are generated tailored to the variables: histograms for numerical variables and bar charts for categorical variables. These visual tools help users intuitively grasp data distributions and frequency patterns.

For a more detailed examination of variable relationships, the application includes scatterplots. Users can plot numerical variables against one another, optionally coloring the data points by a categorical variable to add an additional layer of context. For instance, users can investigate how nutrient intake varies by demographic factors such as age or gender. Each scatterplot incorporates linear regression lines and equations, making trends and associations between variables clearer and more interpretable with a statistical foundation.

To further identify relationships among variables, the application also features a correlation heatmap, visualizing the correlation coefficients between numerical variables, using a gradient of colors to highlight the strength and direction of associations. This helps with identifying patterns of correlation and areas of interest to focus on, however there may not always be a direct causal factor to the relationship.

Beyond exploratory analysis, the API also includes a predictive tool using k-nearest neighbors (kNN) for classification. This tool allows users to classify new data points based on selected features and input values. For example, the kNN feature can predict whether a participant consumed dietary supplements, alcohol, or gum prior to a blood draw, providing clinicians or researchers with actionable insights to refine their interpretations. The classification output includes not only the predicted category but also class probabilities, ensuring transparency.

## Data Analysis Infrastructure

### Server API

The server-side API was developed using Flask to facilitate data exploration and analysis, facilitating smooth interaction between the front-end and back-end. The API provides endpoints for different analyses, including /analyze for summary statistics and scatter plots, /correlation_heatmap for generating heatmaps, and /knn_predict for k-NN predictions. It accepts POST requests with user-specified parameters in JSON format, validates inputs, processes the data, and returns results in either JSON format or as Base64-encoded images for visualization. The server API also takes care of error handling, ensuring that users receive informative feedback if inputs are invalid, improving accessibility.

### Web Front-end

The web front-end complements the server API with a dynamic interface. Users can select analysis options through dropdown menus, which are dynamically populated based on the dataset's features. Interactive forms guide users to specify relevant parameters, such as columns for analysis or feature values for k-NN predictions. The interface is responsive to user input, dynamically adjusting displayed options based on the analysis type selected.

The web interface enables users to choose from various analysis options, specify parameters through dropdown menus, and view results directly on the interface. For example, users can select variables for summary statistics, specify both X and Y columns for scatterplots, or provide input values and a parameter k for k-NN predictions. The application dynamically responds to these parameters, ensuring that only relevant options are displayed, thus reducing confusion. Visualizations, such as histograms or the correlation heatmap, are rendered directly in the interface.

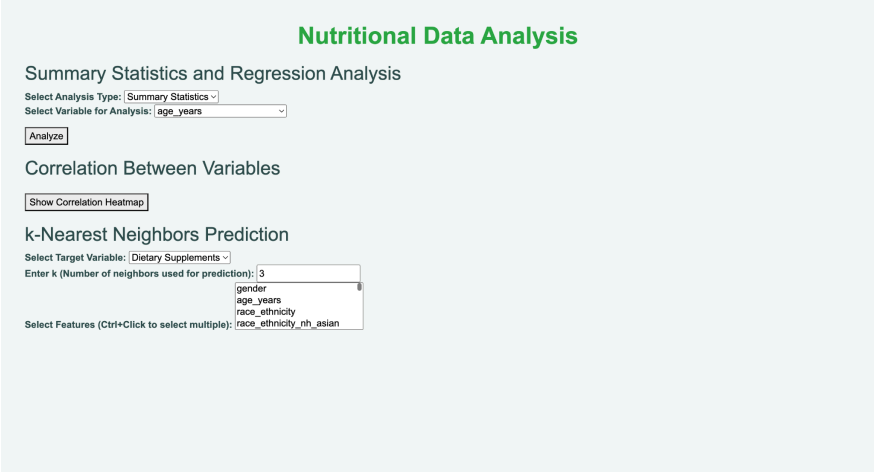
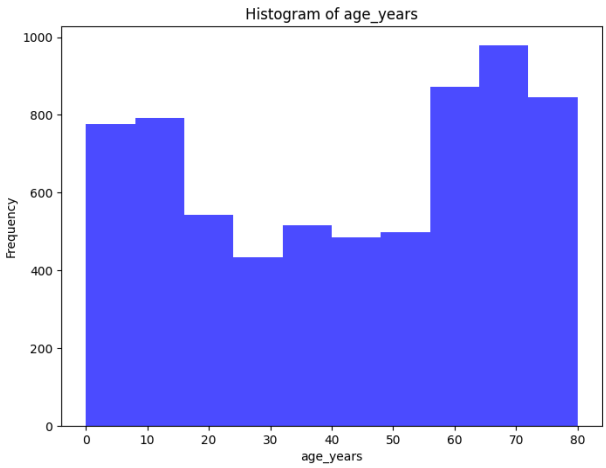



**Figure 1: Default web interface setup when launching the API.**

# Results

## Summary Statistics

The summary statistics revealed important insights into the demographic characteristics of the respondents and their nutritional patterns. Respondents ranged in age from newborns to 80 years, with a bimodal distribution showing a high frequency of children and adults above 60 years (Figure 2). In terms of gender, women accounted for a slight majority (54.4%), with no other gender identities reported. The sample was predominantly White (55.7%), followed by non-Hispanic Black and Hispanic groups, each representing about 10% of respondents. Most respondents (84.5%) were born within the 50 states or Washington, DC, with the remainder born outside the U.S.




**25%:** 17

**50%:** 44

**75%:** 65

**count:** 6744

**max:** 80

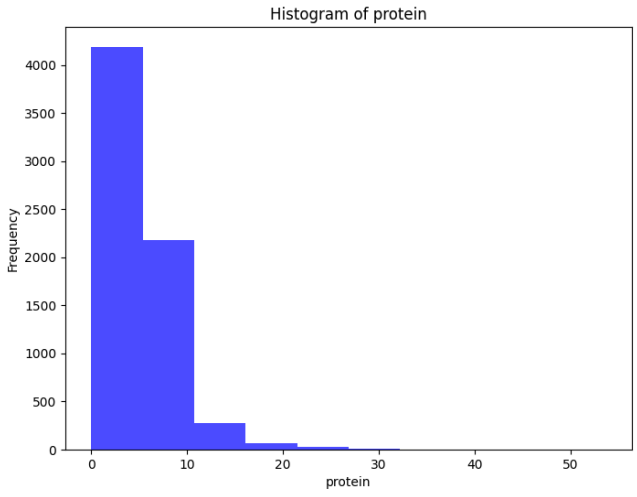
**mean:** 41.87826215895611

**min:** 5.397605346934028e-79

**std:** 25.25034392816576

**Figure 2: Age distribution in the NHANES dataset after processing and cleaning.**

Nutritional variables showed heavily right-skewed distributions, such as energy, protein, and vitamin C intake, making measures like the interquartile range (IQR) more informative than mean values. For instance, the distribution of protein intake highlights the utility of using the IQR to understand typical consumption patterns (Figure 3).




**25%:** 3.2868840579710144

**50%:** 4.588948717948718

**75%:** 6.4805

**count:** 6744

**max:** 53.69

**mean:** 5.357332790489602

**min:** 5.397605346934028e-79

**std:** 3.4274071220528675

**Figure 3: Distribution and summary statistics of protein intake (in grams).**

## Scatterplots and Linear Regression

The analyses of scatterplots and linear regressions revealed numerous linear relationships among nutritional variables. As expected, there were strong positive correlations between variables such as total fat and monounsaturated fat (Figure 4), since higher levels of monounsaturated fat inherently require more total fat.

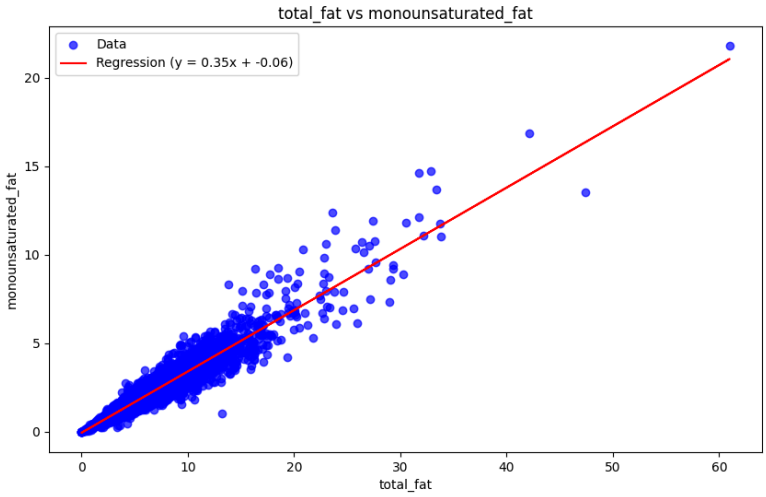



**Figure 4: Linear regression of monounsaturated fat and total fat.**

However, some unexpected relationships were also observed. For example, a strong positive correlation between choline and cholesterol was identified (Figure 5a). Although these nutrients are not typically linked, choline has been shown to boost cholesterol homeostasis, which could explain this relationship. Similarly, a positive correlation between vitamin K and lutein/zeaxanthin (Figure 5b) likely arises because these nutrients are commonly consumed together in foods like leafy greens and egg yolks.

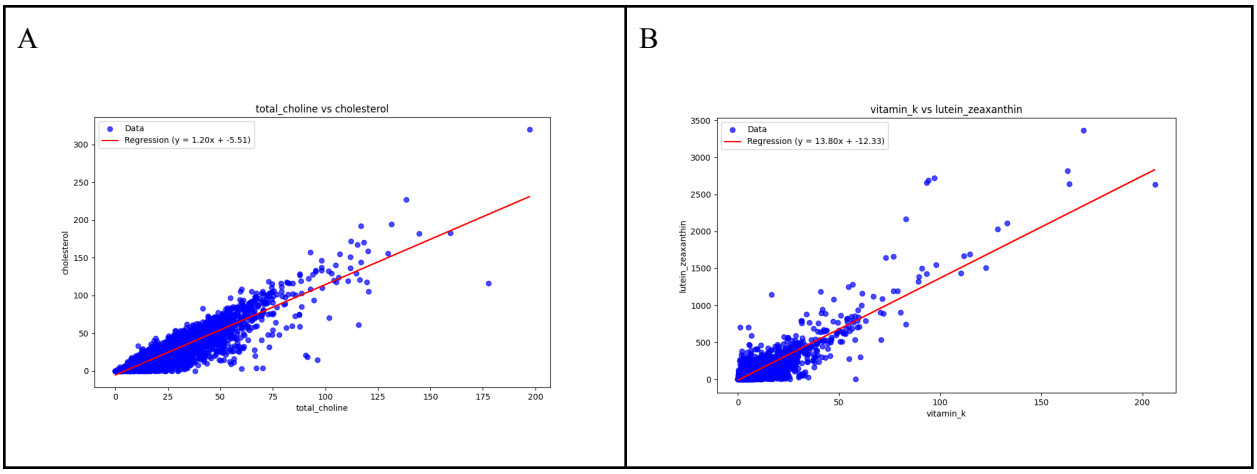



**Figure 5: Unexpected regression results. (A) Linear regression of cholesterol versus choline intake. (B) Linear regression of lutein and zeaxanthin against vitamin k intake.**

Conversely, some relationships reflected mutually exclusive consumption patterns. For instance, there was an inverse relationship between caffeine and vitamin C intake (Figure 6). This may reflect distinct dietary behaviors: individuals with high vitamin C intake often prioritize a diet rich in fruits and vegetables, which may align with a health-conscious lifestyle and less caffeine consumption. On the other hand, high caffeine consumers, potentially through energy drinks, may not prioritize a fresh or nutrient-dense diet.

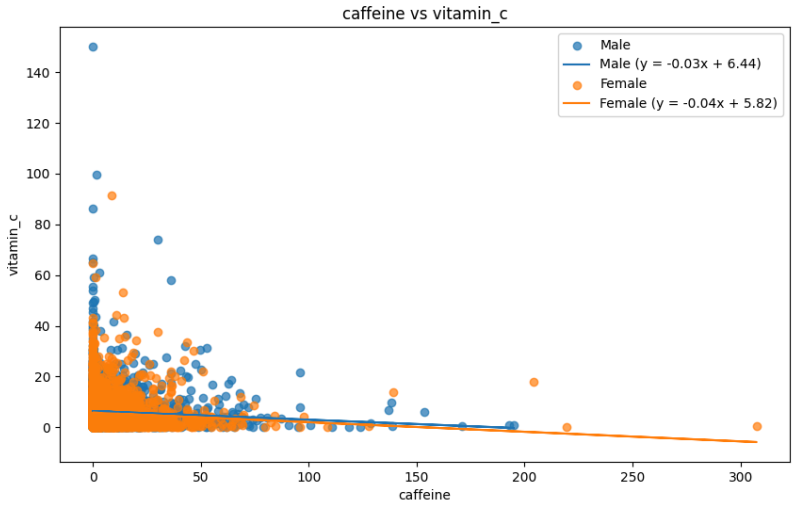



**Figure 6: Non-linear relationship between vitamin c and caffeine intake.**

Nutritional patterns were generally consistent across demographic groups, though certain trends were observed. For instance, when plotting total sugar against energy and grouping by ethnicity, regression lines were similar across most groups, except for non-Hispanic White respondents and those classified as "other," who tended to consume more sugar as energy intake increased (Figure 7).

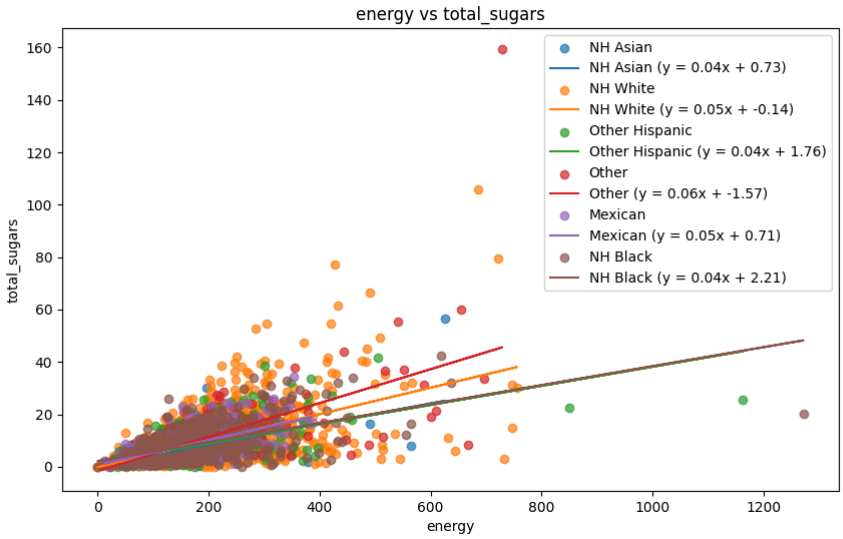



**Figure 7: Total sugar intake versus total energy plotted and analyzed by ethnicities.**

## Correlation Matrix

A correlation matrix (Figure 8) highlighted clusters of relationships among nutritional variables. Predictable clusters were observed among general macronutrients (energy, protein, fats, sugars) and minerals (calcium, magnesium, sodium). Notably, unexpected correlations emerged, such as between cholesterol and choline intake, as well as between vitamin K and lutein/zeaxanthin. These relationships further underscore the complex interplay of dietary patterns and nutrient co-occurrence in common food sources.

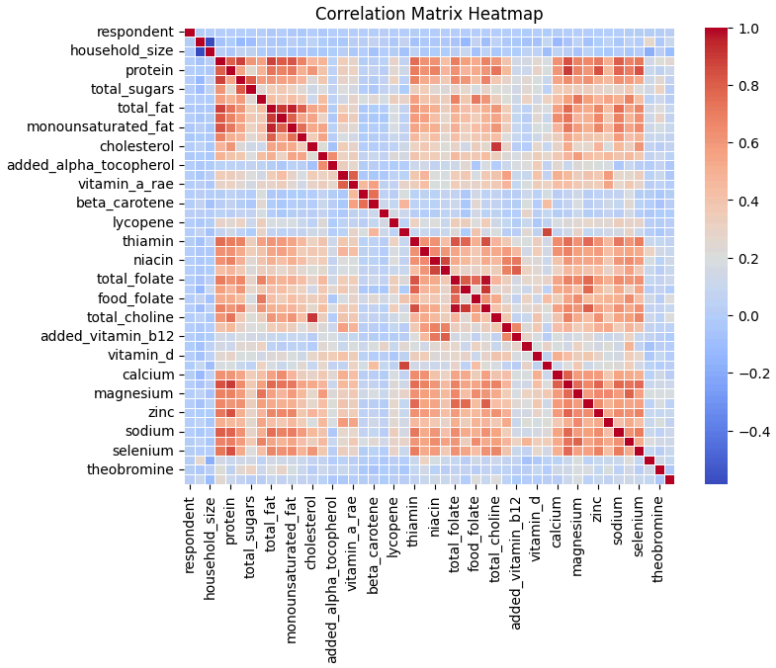



**Figure 8: Linear Correlation Matrix Heatmap.**

## k-Nearest Neighbors (kNN) Prediction

Using a k-Nearest Neighbors (kNN) algorithm, we explored the predictive power of specific variables for alcohol use prior to blood draws. For example, given a 50-year-old respondent with 0 mg caffeine and 30 grams of alcohol intake, the model predicted a low probability of alcohol consumption before the blood draw (Class No: 0.857; Class Yes: 0.143). While the general prediction aligns with the overall trend of low alcohol consumption before lab tests, the increased likelihood compared to the general population suggests the influence of specific input variables (Figure 9).

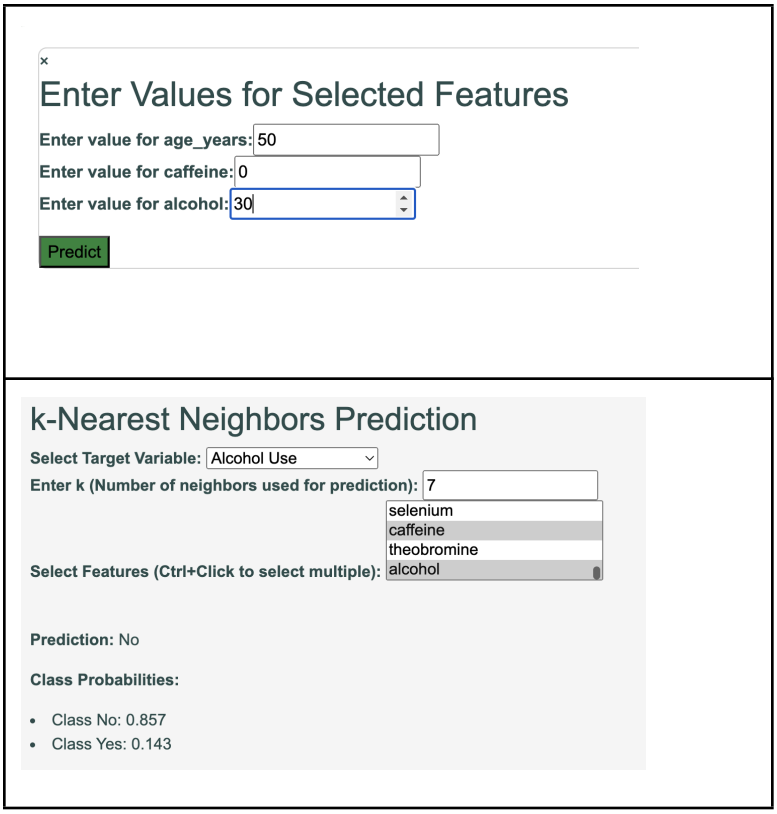



**Figure 9: Example for kNN feature selection and results.**

# Discussion

The primary objective of this API was to enhance understanding of nutritional values and their broader implications for health. Nutritional intake is widely acknowledged as a predictor of health, influencing metabolic function, immune response, and disease prevention. By analyzing detailed data on food composition alongside demographic variables, the API can foster a deeper understanding of how dietary habits intersect with population-level health trends. Users can explore how nutritional patterns vary across age groups, genders, and other demographic categories, providing valuable insights into dietary disparities and opportunities for targeted interventions.

Together with the integrated prediction tool, this approach has the potential to significantly improve the interpretation of blood test results by accounting for external factors that could otherwise lead to misinterpretation. Such insights can guide more effective and tailored healthcare decisions.

Despite its capabilities, the API underscores the complexity of diet-health relationships. While nutrition plays a crucial role in determining health outcomes, it is only one piece of a larger puzzle that includes physical activity, sleep quality, genetics, and various environmental factors. Consequently, while the API provides valuable insights into nutrition, it is important to recognize its scope and limitations.

The accuracy of the API's nutritional analysis and predictions also relies heavily on the quality and completeness of the data provided. For example, errors in self-reported food intake or omissions in demographic information could reduce the reliability of the outputs. Similarly, while the prediction model for the impact of supplements, alcohol, or gum on blood test results is based on general trends and correlations, it cannot account for every individual's unique circumstances. Users should view the API as a complementary tool rather than a replacement for professional medical advice.

In the future, the API could be expanded to include additional health metrics, such as physical activity levels, sleep patterns, and medical history, to deliver a more holistic view of individual health. Furthermore, integrating medical nutritional guidelines and comparisons with recommended dietary values could provide users with actionable insights into how their dietary habits align with optimal nutrition standards and not just how they compare to other people in the United States. These enhancements would not only improve the API's utility but also strengthen its role in advancing personalized healthcare and population-level nutritional understanding.