**SIADS Milestone II - Winter, 2025**
**Leveraging Machine Learning to Classify Junk Food vs. Regular Food**
Ayan Banerjee, Richard Chalker, Alex DeLoach

## Introduction

The U.S. food supply is heavily dominated by packaged foods and beverages, which contribute approximately 75% of daily calorie intake for the population. A significant portion of these calories derives from foods commonly perceived as “junk foods.” However, there is no universally accepted definition of junk food in academic literature, making it challenging to standardize classification. Many products high in saturated fat, added sugars, energy, or salt are not consistently labeled as junk food. The overconsumption of such foods has been linked to serious health issues, including heart disease, obesity, and reduced overall functionality.

This project aims to explore machine learning approaches to classify foods into two categories: "junk food" and "regular food." By leveraging nutritional data and advanced computational techniques, we will provide a systematic framework to analyze and categorize foods. This initiative aligns with growing efforts to address public health challenges through data-driven insights.

With respect to the Supervised approach, we determined that under both Nutrient Value and Nutrient Density Analysis, solids and liquids generated very different outcomes, including in model selection and performance. NV had one model as best performing for solids and liquids, while the NDA approach had 2 different models as the best performers.

With respect to the Unsupervised approach, the K-Means clustering model performed much better for the liquids than for the solids. There were significant differences between the average nutrient values found for each cluster for both liquids and solids; for example, the average Energy measurement for solids in Cluster 0 is about approximately **410.5 kcal**, whereas the average for solids in Cluster 1 is approximately **172.9 kcal**.

**Related Works** (See references for link information)

a. *A Real-Time Junk Food Recognition System Based on Machine Learning.* This analysis from 2022 also used machine learning in an attempt to classify junk food with a high degree of accuracy. This analysis leveraged unsupervised learning but focused on the development of a CNN to analyze multiple definitions of junk food. Our analysis focused on 2 approaches of calculating junk food and leveraged both supervised and unsupervised techniques (but not CNNs).
b. *Nutrient profiling: Is the technique family of nutrient profile models?* This analysis leveraged the Chilean study to try to evaluate the intake of different types of nutrients by Americans. This analysis, however, didn’t use any form of machine learning, but instead analyzed the NHANES data to perform a statistical analysis on the demographic characteristics in that data.
c. *Machine learning in nutrition research.* This study discusses the various techniques in which machine learning can be applied to nutritional analysis. This study didn’t run any of the analysis but provided insight as to how to evaluate different approaches.

**SIADS Milestone II - Winter, 2025**
**Leveraging Machine Learning to Classify Junk Food vs. Regular Food**
Ayan Banerjee, Richard Chalker, Alex DeLoach

This project is not a continuation of any prior work in MADS.

## Data Source

Food Surveys Research Group: Beltsville, MD under Food and Nutrient Database for Dietary Studies (FNDDS) of U.S. Department of Agriculture, collected from 2021-2023. The dataset was downloaded as a csv file. It contains 5,431 records of food including:

- A unique Food Code
- A food description
- A WWEIA Food Number
- A WWEIA Food Description
- Nutrient information – 65 different nutrients

The dataset needed little preprocessing, other than the analysis to determine liquid / solid and the junk food classification (see below).

**Data source link:** https://www.ars.usda.gov/northeast-area/beltsville-md-bhnrc/beltsville-human-nutrition-research-center/food-surveys-research-group/docs/fndds-download-databases/

## Definitions of Junk Food

To complete this analysis, we looked at different definitions of junk food.

Nutrient Value Analysis

In 2016, Chile introduced one of the most comprehensive regulatory frameworks globally to address obesity, including specific criteria for defining unhealthy foods. The Chilean nutrient profile model is a methodological approach that classifies junk foods based on their nutrient and food component composition. For one of our analyses, the criteria from phase 3 of the regulatory framework were utilized. Products exceeding the established nutrient thresholds were classified as junk foods in this study.

| | 24 months (Phase 2) | 36 months (Phase 3) |
|---|---|---|
| **A. Solid foods** | | |
| Energy, kcal/100 g | 300 | 275 |
| Sodium, mg/100 g | 500 | 400 |
| Total sugar, g/100 g | 15 | 10 |
| Saturated fat, g/100 g | 5 | 4 |
| **B. Liquids** | | |
| Energy, kcal/100 g | 80 | 70 |
| Sodium, mg/100 g | 100 | 100 |
| Total sugar, g/100 g | 5 | 5 |

| | | |
|---|---|---|
| Saturated fat, g/100 g | 3 | 3 |

*Table 1: Nutrient thresholds as defined by the Chilean nutrient profile model*

Nutrient Density Analysis

One alternative approach to evaluating whether a food is "junk food" is in using Nutrient Density Analysis (NDA). In contrast to the nutrient value approach mentioned above, NDA creates an index which evaluates the beneficial nutrients in a food, but balances (reduces) that value by the limiting values – "bad" nutrients in excess.

There are several different algorithms for NDA, varying on nutrients and serving size (see appendix for a representative sample).

For our analysis, we have selected using NRF9.3. NRF9.3 is defined as:

$$NRF9.3 = \sum(\text{beneficial nutrients per 100 kcal}) - \sum(\text{limiting nutrients per 100 kcal})$$

The higher the score, the healthier the food. In terms of the variables for calculating NRF9.3, the below nutrients are categorized as "healthy":

- Protein
- Fiber
- Vitamins A, C, and E
- Calcium
- Magnesium
- Iron
- Potassium

The limiting nutrients in NRF9.3 are:

- Saturated fat
- Added sugar
- Sodium

For our analysis using NDA, we built a model which focuses on the above listed nutrients as the beneficial and limiting nutrients.

In evaluating the 2021-2023 Food and Nutrient Database for Dietary Studies (FNDDS) database, we determined that this approach had very different results when comparing solids versus liquids (see Appendix). Of note, in looking at the outcomes, they clearly do not scale between solids and liquids. For example, is liver twice as healthy as carrots, but less healthy than coffee? For our analysis, we will differentiate between solids and liquids to minimize any scaling implications.

## Feature Engineering / EDA

We performed an exploratory data analysis (EDA) of the 2021-2023 Food and Nutrient Database for Dietary Studies (FNDDS). The dataset provides information on various food items and their nutrient composition, which is useful for dietary research and nutritional assessment. The

purpose of this analysis is to uncover insights into nutrient distributions, identify missing data, and explore relationships between different nutrient components.

Details of the analysis are in the appendix.

We derived the following insights from our analysis:

- The dataset contains a wide variety of food items with varying macronutrient compositions
- Energy content is right-skewed, with most food items having moderate caloric values
- Macronutrient distributions suggest that carbohydrate content varies widely, while protein and fat distributions are relatively constrained
- Correlation analysis reveals strong associations between energy, carbohydrate, and fat content.
- Dataset Balance:

| | **NVA** | | **NDA** | |
|---|---|---|---|---|
| | **Solid** | **Liquid** | **Solid** | **Liquid** |
| **Not Junk Food** | 4358 | 887 | 2261 | 720 |
| **Junk Food** | 114 | 72 | 2211 | 239 |

*Table 2: Data imbalance as observed in Nutrients Value and Density analysis*

**Supervised Learning**

Methods Description

Nutrients Value Based Analysis and Nutritional Density Analysis

The dataset is updated with a flag to indicate the type of food, categorizing it as either solid or liquid. This classification is determined based on the presence of specific keywords in the food description, including *milk*, *dips*, *spread*, *cream*, *butter*, *sauces*, *dressing*, *ice cream*, *yogurt*, *formula*, and *drinks*. We defined liquid as an item that, in room temperature, can take the shape of its container, flow freely, and it is not very dense.

Given that the nutrient thresholds for defining Junk Food differ between solid and liquid food types, separate datasets are created for solid and liquid foods to facilitate further analysis. Further, we have different categorizations of junk food between the Nutrients Value and the Nutrient Density analyses. NV provides a binary outcome – if the conditions are met, then the food is a junk food. For NDA, the analysis provides a score ranging from -infinity to +infinity. For this analysis, we have defined junk food as anything with a negative value (i.e., the limiting nutrients outweigh the beneficial nutrients).

As shown in the result here, in the NV analysis, a dummy classifier yields superficially high accuracy by predominantly predicting the majority class (Not Junk Food). The extremely low precision, recall, and F1-score for Junk Food (1) confirm that the model provides **virtually no predictive power** for the minority class.

**SIADS Milestone II - Winter, 2025**
**Leveraging Machine Learning to Classify Junk Food vs. Regular Food**
Ayan Banerjee, Richard Chalker, Alex DeLoach

Dummy classifier report: Liquid

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.93 | 0.92 | 0.92 | 178 |
| 1 | 0.06 | 0.07 | 0.07 | 14 |
| accuracy | | | 0.85 | 192 |
| macro avg | 0.49 | 0.49 | 0.49 | 192 |
| weighted avg | 0.86 | 0.85 | 0.86 | 192 |

Dummy classifier report: Solid

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.97 | 0.97 | 0.97 | 873 |
| 1 | 0.00 | 0.00 | 0.00 | 22 |
| accuracy | | | 0.95 | 895 |
| macro avg | 0.49 | 0.49 | 0.49 | 895 |
| weighted avg | 0.95 | 0.95 | 0.95 | 895 |

*Figure 1: Dummy classifiers for Junk food classification of Liquid and Solid food types*

Supervised Evaluation – Analysis of the Approaches and Models

Nutrients Value Based Analysis

Due to the relatively small size of the dataset, classical machine learning classification models are employed for analysis. Using 5-fold cross validation, a comparative evaluation of model performance indicates that **Gradient Boosting** outperforms other models for both solid food and liquid food categories.

Therefore, we have decided to use **Gradient Boosting Classifier** for further analysis.

| Model | Accuracy | F1 Score | Mean Accuracy | Std |
|---|---|---|---|---|
| Gradient Boosting | 0.998883 | 0.998870 | 0.998322 | 0.002712 |
| Random Forest | 0.987709 | 0.985700 | 0.985463 | 0.004295 |
| K-Nearest Neighbors | 0.978771 | 0.974398 | 0.978755 | 0.004446 |
| Support Vector Machine | 0.976536 | 0.965964 | 0.974840 | 0.002330 |
| Logistic Regression | 0.972067 | 0.964962 | 0.969528 | 0.005687 |

*Figure 2: Model performance on 'Solid' food type (NVA)*

| Model | Accuracy | F1 Score | Mean Accuracy | Std |
|---|---|---|---|---|
| Gradient Boosting | 0.989583 | 0.989212 | 0.994780 | 0.002610 |
| Random Forest | 0.979167 | 0.978424 | 0.983032 | 0.008883 |
| Logistic Regression | 0.963542 | 0.964119 | 0.949138 | 0.014592 |
| Support Vector Machine | 0.963542 | 0.964119 | 0.940031 | 0.015066 |
| K-Nearest Neighbors | 0.947917 | 0.946059 | 0.949121 | 0.023936 |

*Figure 3: Model performance on 'Liquid' food type (NVA)*

A feature importance analysis was conducted to identify the key factors influencing the classification outcomes. For solid food types, **Sodium, Total Fat** and **Sugars** were identified as the three most significant features contributing to the prediction results.

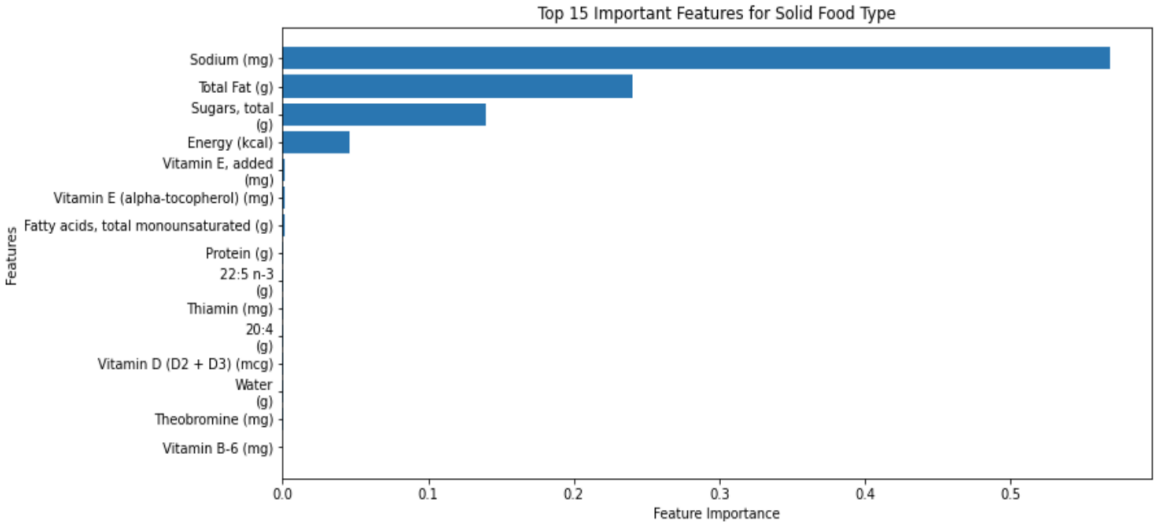


*Figure 4: Top 15 important features for 'Solid' type food*

For liquid food types, we observe the same influential features as with the solids – **Sodium**, **Total Fat** and **Sugars**.

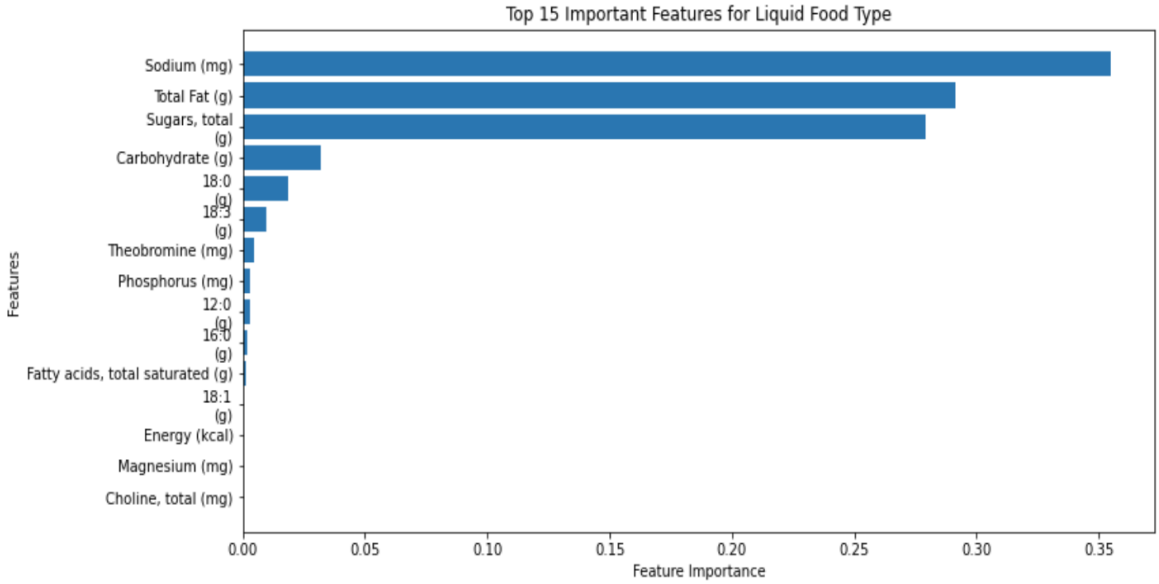


*Figure 5: Top 15 important features for Liquid type food*

The feature ablation study was conducted to assess the impact of individual features on model performance. For solid type foods, the results indicate that the removal of some features like **Alcohol, 12:0, Niacin,** and **18:0** do not affect the model's accuracy or F1 score, which remained constant at **0.998883 and 0.998870**, respectively.

However, we observe that removal of **Fatty acids, Vitamin B-12, Selenium, Zinc, Vitamin A, 14:0, Vitamin B-6, Thiamin** and **Vitamin E** reduces the performance slightly indicating that these features play more vital role in classification. This analysis highlights the potential for

feature reduction without compromising predictive performance, thereby improving model efficiency and interpretability.

For liquid type foods, the results indicate that the removal of the features, such as **Water, Lutein + zeaxanthin, Potassium, 20:5, 16:1, Vitamin E, Fatty acids, Vitamin B-12, 10:0, Copper, Choline, 16:0** and **18:0** had no effect on the model's accuracy and F1 score, which remained at approximately **0.989583 and 0.989212**, respectively.

The sensitivity analysis of the **Gradient Boosting Classifier** for solid food classification into **Junk Food** and **Not Junk Food** was conducted, with the optimal hyperparameters determined as **{'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 50}**. The best model achieved an **exceptionally high accuracy of 0.9966**, demonstrating strong predictive performance.

```
Best Hyperparameters: {'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 50}
Best Model Accuracy: 0.9966
Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       873
           1       0.95      0.91      0.93        22

    accuracy                           1.00       895
   macro avg       0.98      0.95      0.96       895
weighted avg       1.00      1.00      1.00       895
```

*Figure 6: Sensitivity Analysis and Model Performance for Solid type food*

The **classification report** indicates that the model performs equally well in classifying both **Not Junk Food (label 0) and Junk Food (label 1)**. The model achieved a **precision, recall, and F1-score of 1.00 for Not Junk Food**, suggesting perfect classification with no false positives or false negatives across 873 instances. For **Junk Food (label 1)**, the model achieved a **precision of 0.95**, meaning that **5% of the food items predicted as Junk Food were incorrect**. The **recall of 0.91** indicates that **9% of actual Junk Food items were misclassified as Not Junk Food**, leading to an **F1-score of 0.93**.

The **macro-averaged F1-score of 0.96** and **weighted F1-score of 1.00** confirm that the model achieves near-perfect classification despite the presence of **class imbalance (873 instances of Not Junk Food vs. 22 instances of Junk Food)**.

For liquid types of food, the sensitivity analysis was conducted, with optimal hyperparameters determined as **{'max_depth': None, 'min_samples_leaf': 2, 'min_samples_split': 2, 'n_estimators': 200}.** The best model achieved a high overall accuracy of **0.9896**, indicating strong classification performance.

```
Best Hyperparameters: {'max_depth': None, 'min_samples_leaf': 2, 'min_samples_split': 2, 'n_estimators': 200}
Best Model Accuracy: 0.9896
Classification Report:
              precision    recall  f1-score   support

           0       0.99      0.99      0.99       178
           1       0.93      0.93      0.93        14

    accuracy                           0.99       192
   macro avg       0.96      0.96      0.96       192
weighted avg       0.99      0.99      0.99       192
```

*Figure 7: Sensitivity Analysis and Model Performance for Liquid type food*

The **classification report** further reveals that the model performs exceptionally well. The **Not Junk Food category (label 0)** was classified with **0.99 precision, 0.99 recall, and an F1-score of 0.99**, based on **178 instances**. The **Junk Food category (label 1)** achieved **0.93 precision, 0.93 recall, and an F1-score of 0.93**, based on **14 instances**. The **macro-averaged F1-score (0.96)** indicates that the model performs well across both classes, while the **weighted F1-score (0.99)** confirms strong performance, considering class distribution. While the model is highly effective, **a slight reduction in recall (0.93) for Junk Food classification** indicates that **7% of actual Junk Food items were misclassified as Not Junk Food**. This suggests the potential for **further improvements in model sensitivity**, possibly through **data augmentation, threshold tuning, or feature engineering** to enhance the identification of Junk Food items while maintaining high overall accuracy.

Nutrients Density Analysis

Using NDA, there are some different outcomes than in the NV analysis:

| Model | Accuracy | F1 Score | Mean Accuracy | Std |
|---|---|---|---|---|
| Support Vector Machine | 0.972036 | 0.972019 | 0.950786 | 0.005934 |
| XGB | 0.96085 | 0.960851 | 0.964208 | 0.005182 |
| Gradient Boosting | 0.956376 | 0.95635 | 0.962251 | 0.007279 |
| Logistic Regression | 0.949664 | 0.94962 | 0.939321 | 0.007404 |
| Random Forest | 0.928412 | 0.928424 | 0.928694 | 0.011207 |
| K-Nearest Neighbors | 0.841163 | 0.841077 | 0.831935 | 0.002836 |

*Figure 8: Model performance on 'Solid' food type (NDA)*

| Model | Accuracy | F1 Score | Mean Accuracy | Std |
|---|---|---|---|---|
| XGB | 0.941176 | 0.940645 | 0.963803 | 0.012458 |
| Gradient Boosting | 0.935829 | 0.935829 | 0.955758 | 0.009128 |
| Support Vector Machine | 0.935829 | 0.934606 | 0.932975 | 0.008491 |
| Random Forest | 0.930481 | 0.929397 | 0.946425 | 0.018291 |
| Logistic Regression | 0.930481 | 0.928908 | 0.937029 | 0.015475 |
| K-Nearest Neighbors | 0.882353 | 0.881657 | 0.898174 | 0.021051 |

*Figure 9: Model performance on 'Liquid' food type (NDA)*

In looking at the models for NDA, solids performed better using the **Support Vector Machine Classifier**. However, the liquid data performed better using the **XGB Classifier**.

A feature importance analysis was conducted to identify the key factors influencing the classification outcomes. For solid food types, like the NV, **Sodium** was identified as the most significant features contributing to the prediction results. All other features had minimal impact.

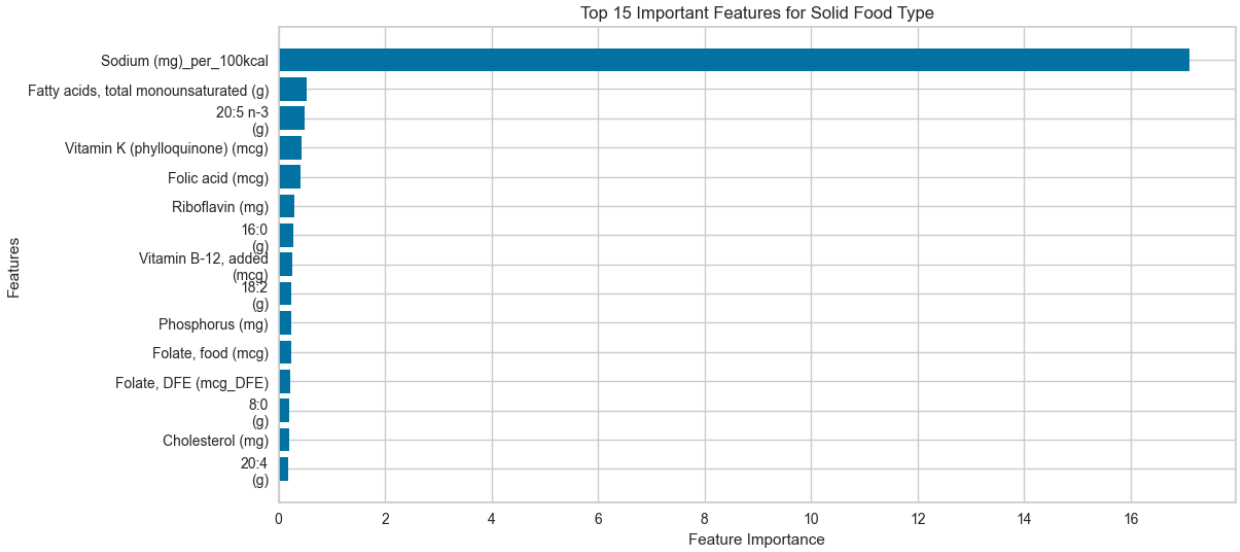


*Figure 10: Top 15 important features for 'Solid' type food*

For liquid food types, **Potassium and Sodium** were the most significant factors, although other nutrients contributed to the overall model performance.

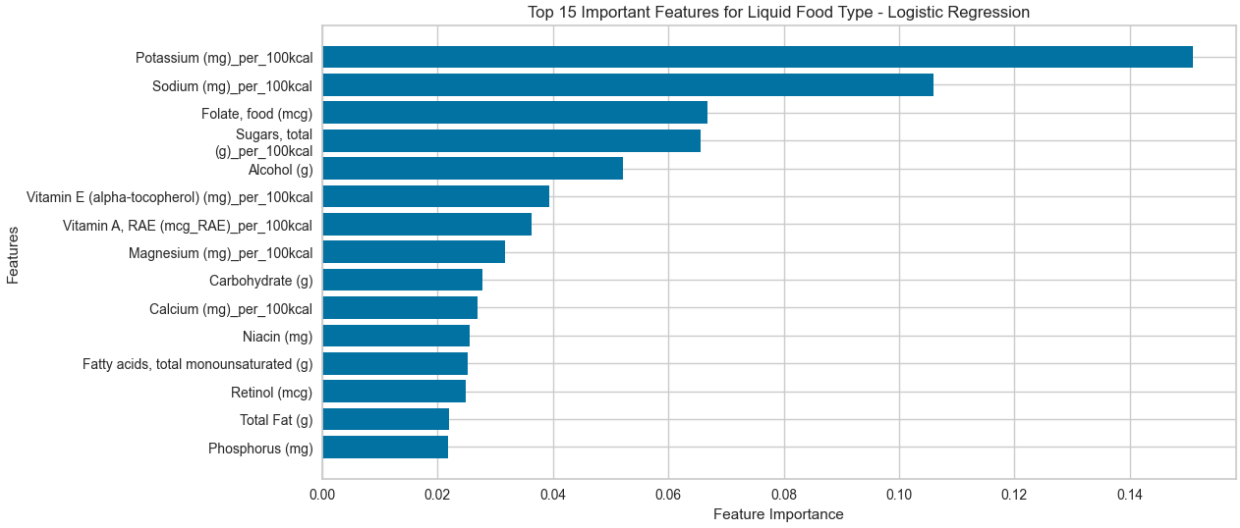


*Figure 11: Top 15 important features for Liquid type food*

For NDA, the feature ablation study was conducted to assess the impact of individual features on model performance. The analyses for both solids and liquids indicate that the removal of each of the nutrients impacts the overall performance of the models. Reducing the top 12 features reduces both models from accuracies in the mid – 90s to the low 70s (solids) / 80s (liquids).

**SIADS Milestone II - Winter, 2025**
**Leveraging Machine Learning to Classify Junk Food vs. Regular Food**
Ayan Banerjee, Richard Chalker, Alex DeLoach

***References:*** Feature ablation study on Solid type food (NDA)**,** Feature ablation study on Liquid type food (NDA)

For NDA, the sensitivity analysis of the **SVM Classifier** for solid food classification into **Junk Food** and **Not Junk Food** was conducted, with the optimal hyperparameters determined as **{'C': 10, 'gamma': 'scale', 'kernel': 'linear'}**. The best model achieved an **exceptionally high accuracy of 0.9855**, demonstrating strong predictive performance.

Best Hyperparameters: {'C': 10, 'gamma': 'scale', 'kernel': 'linear'}
Best Model Accuracy: 0.9855
Classification Report:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.99 | 0.98 | 0.99 | 435 |
| 1 | 0.98 | 0.99 | 0.99 | 459 |
| accuracy | | | 0.99 | 894 |
| macro avg | 0.99 | 0.99 | 0.99 | 894 |
| weighted avg | 0.99 | 0.99 | 0.99 | 894 |

*Figure 12: Sensitivity Analysis and Model Performance for Solid type food*

The **classification report** indicates that the model performs consistently in classifying both **Not Junk Food (label 0) and Junk Food (label 1)**. The model achieved a **precision, recall, and F1-score of .97 - .99 for Not Junk Food and Junk Food**. While not as accurate as the NV approach, the results suggest that the NDA approach is consistent with respect to classifying **Junk Food** and **Not Junk Food**.

However, for the liquid foods, the performance of the **XGB** model isn't as strong as the solid. The optimal hyperparameters were similarly determined as **{'colsample_bytree': 0.8, 'gamma': 1, 'learning_rate': 0.1, 'max_depth': 5, 'n_estimators': 50, 'subsample': 0.8}}**.

Best Hyperparameters: {'colsample_bytree': 0.8, 'gamma': 1, 'learning_rate': 0.1,
'max_depth': 5, 'n_estimators': 50, 'subsample': 0.8}
Best Model Accuracy: 0.9198
Classification Report:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.93 | 0.96 | 0.95 | 134 |
| 1 | 0.90 | 0.81 | 0.85 | 53 |
| accuracy | | | 0.92 | 187 |
| macro avg | 0.91 | 0.89 | 0.90 | 187 |
| weighted avg | 0.92 | 0.92 | 0.92 | 187 |

*Figure 13: Sensitivity Analysis and Model Performance for Liquid type food*

However, the **classification report** reveals that the model performs better in identifying **Not Junk Food** category (label 0), with scores for **precision (.93), recall (.96),** and **F1-score (.95)** than in the **Junk Food** category (label 1), with scores for **precision (.90), recall (.81),** and **F1-score (.85)**.

**SIADS Milestone II - Winter, 2025**
**Leveraging Machine Learning to Classify Junk Food vs. Regular Food**
Ayan Banerjee, Richard Chalker, Alex DeLoach

Failure Analysis (best model)

Nutrients Value Based Analysis

The **model failure analysis** for solid food classification into **Junk Food** and **Not Junk Food** was conducted to identify misclassified instances and understand potential reasons for model errors. The analysis revealed **three misclassified samples.** In one case, a **Not Junk Food (label 0)** has been misclassified as **Junk Food (label 1)** and in two cases **Junk Foods (label 1)** were misclassified as **Not Junk Food (label 0)**.

```
Misclassification Counts for solid food type:
   Actual  Predicted  Count
0       0          1      1
1       1          0      2
```

*Figure 14: Misclassification report for Solid food type*

Examining the nutritional composition of the misclassified samples, it was observed that the sample that was misclassified as **Junk Food** has **relatively high sugar (13.08 g)** and **Total Fat (22.85 g)** content whereas the two samples misclassified as **Not Junk Food** have **comparatively low Total Fat (17.05 g and 4.01 g) content** which may have influenced the model's misclassification. This result confirms the outcome of feature importance analysis (reference: Figure 4).

Model failure analysis for liquid food classification identified **two instances of misclassification**. In one instance, **a Not Junk Food (label 0) has been misclassified as Junk Food (label 1)** and in the other instance **a Junk Food has been misclassified as Not Junk Food**.

```
Misclassification Counts for liquid food type:
   Actual  Predicted  Count
0       0          1      1
1       1          0      1
```

*Figure 15: Misclassification report for Liquid food type*

Based on the nutritional composition of the misclassified samples, it is observed that the sample misclassified as Junk Food has high carbohydrate quantity (23.11 g) whereas the sample misclassified as Not Junk Food is low on fat content (4.40 g). As per the feature importance analysis of liquid food type, carbohydrate is the 4$^{th}$ most influential feature and therefore the model is likely to have misclassified a regular food as Junk food whereas misclassified a Junk food as regular food due to low fat content (reference: Figure 5).

Nutrient Density Analysis

The **model failure analysis** for solid food classification into **Junk Food** and **Not Junk Food** was conducted to identify misclassified instances and understand potential reasons for model errors. The analysis revealed **25 misclassified samples**, for both **Junk Food** and **Not Junk Food** classifications:

Misclassification Counts:

| | Actual | Predicted | Count |
|---|---|---|---|
| 0 | 0 | 1 | 19 |
| 1 | 1 | 0 | 6 |

*Figure 16: Misclassification report for Solid food type*

**Magnesium** and **Potassium** differences seem to have contributed the most of the model's misclassification.

Misclassification Counts:

| | Actual | Predicted | Count |
|---|---|---|---|
| 0 | 0 | 1 | 4 |
| 1 | 1 | 0 | 7 |

*Figure 17: Misclassification report for Liquid food type*

Model failure analysis for liquid food classification had fewer instances of **misclassification**. Much of those misclassifications can be attributed to variances with **Calcium** and **Magnesium**.

## Unsupervised Learning

Methods Description

K-Means Clustering To Identify Natural Groupings Based on Nutritional Values

For both liquids and solids, the datasets were first standardized using the StandardScaler method from scikit-learn. After the standardization, Principal Component Analysis (PCA) was used for dimensionality reduction. Twenty principal components were constructed from the original dataset and these twenty components were used in a K-Means Clustering model to identify any groupings based on the nutritional values. K-Means was selected as opposed to a DBSCAN clustering model because we already have an idea of the number of potential clusters in the dataset (Junk Food & Non-Junk Food). Silhouette scores were then used to evaluate the model's sensitivity to a varying number of clusters, and to select the optimal number of clusters. Labels were created for each of the entries in the dataset, and these labels were used to produce summary statistics for comparison between the different clusters.

Unsupervised Evaluation

Overall Results

The initial dataset contained 70 features, each describing a measured nutrient such as **Energy (kcal)**, **Protein (g)**, and **Carbohydrate (g)**. Because each of these nutrients may use different metrics, standardization was necessary. StandardScaler standardizes features by removing the mean and scaling to unit variance to make each of the individual features look more or less like standard normally distributed data. After both datasets had been normalized, PCA was performed to reduce the dimensionality of the dataset. Each of the datasets were reduced to **20 Principal Components**, with a **Total Explained Variance** of approximately **82.84%** and **86.91%** for solids and liquids, respectively.

**SIADS Milestone II - Winter, 2025**
**Leveraging Machine Learning to Classify Junk Food vs. Regular Food**
Ayan Banerjee, Richard Chalker, Alex DeLoach

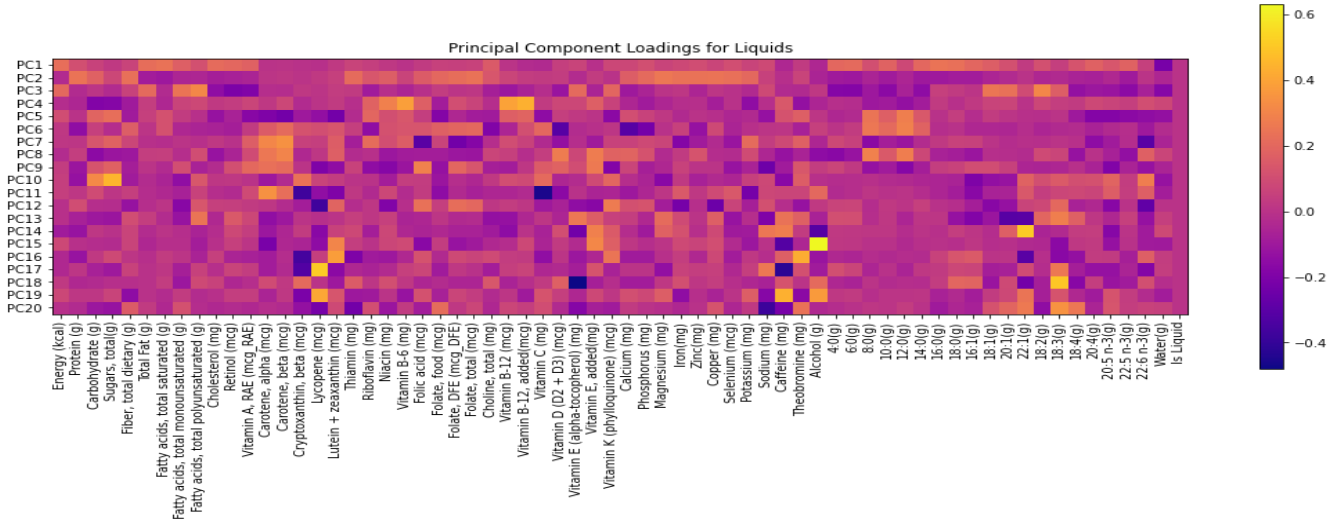


*Figure 18: Principal Component Loadings for Liquids*

The heatmap shows how each feature contributes to the principal components; for the first three liquid principal components (which are responsible for explaining approximately 20.41%, 14.49%, and 8.78% of the variance, respectively), the most important features, or the features having the greatest weight in each of the principal components, are **16:0 (g)**, **Magnesium (mg)**, and **Fatty acids, total polyunsaturated (g).**

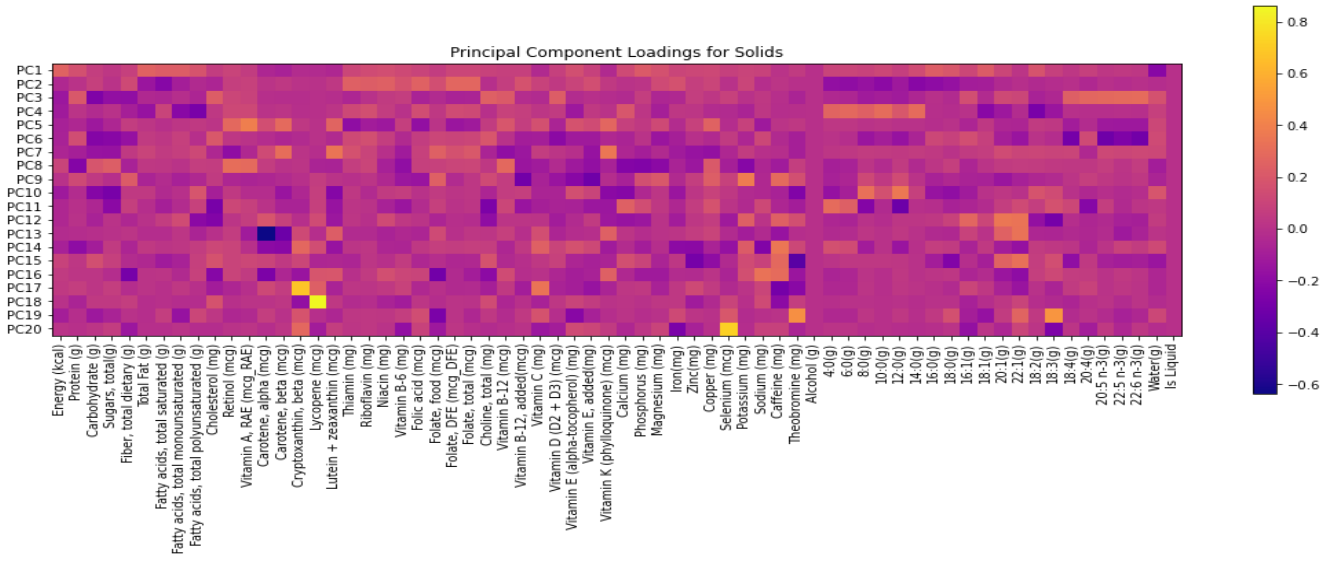


*Figure 19: Principal Component Loadings for Solids*

For the first three solid principal components (which are responsible for explaining approximately 15.80%, 10.53%, and 8.06% of the variance, respectively), the most important features are **Total Fat (g), Folate, DFE (mcg_DFE)**, and **22:6 n-3(g)**.

The twenty principal components were then used as inputs in a K-Means clustering model (all default parameter settings other than **max_iter = 1000**).

**SIADS Milestone II - Winter, 2025**
**Leveraging Machine Learning to Classify Junk Food vs. Regular Food**
Ayan Banerjee, Richard Chalker, Alex DeLoach

Sensitivity Analysis

For the sensitivity analysis of the K-Means model, the number of clusters, ***k***, was changed to analyze the impact of different values of ***k***, and to select the optimal number of clusters.

***References:*** Silhouette Scores for Solids, Silhouette Scores for Liquids

As the number of clusters increases above 2, the average silhouette score continues to decrease for both solids and liquids, so 2 clusters appears to be optimal in this instance. Using 2 clusters, the average silhouette score for liquids is approximately **0.70**, indicating moderately strong clustering. For solids, the average silhouette score with 2 clusters is approximately **0.42**, indicating moderate clustering but suggesting that there is room for improvement.

The solids and liquids datasets were then split apart based on the cluster labels produced from the K-Means model, and summary statistics were used to compare and contrast the different clusters found in the data.

***References:*** Liquid Statistics - Cluster 0 (Left) vs. Cluster 1 (Right), Solid Statistics - Cluster 0 (Left) vs. Cluster 1 (Right)

Model Evaluation

Due to the lack of 'ground truth' labels for data, the Davies-Bouldin Index (DBI) and the Calinski-Harabasz Index (CHI) were the two selected metrics for cluster quality evaluation.

The DBI measures the average similarity measure of each cluster with its most similar cluster, where similarity is the ratio of intra-cluster similarity (compactness) and inter-cluster difference (separation); lower values for the DBI indicate better clustering quality. The DBI for the Solids dataset is **2.11**, suggesting that the clusters may not be very well defined and there is potentially some overlapping between clusters. For the Liquids, the DBI is **0.91**, suggesting much better cluster quality for liquids than for solids.

The CHI (also known as the Variance Ratio Criterion) measures the ratio of the sum of between-cluster dispersion and of within-cluster dispersion; higher values for the CHI indicate better clustering quality. The CHI for solids is approximately **560.95**, while the CHI for liquids is approximately **164.32**. The CHI values indicate that while both solids and liquids appear to have relatively strong clustering quality, the CHI for solids is significantly higher, indicating that the clusters are better separated from each other and internally cohesive. This likely needs additional analysis, given the higher DBI and lower average silhouette scores when clustering.

## Discussion

Part A:

The most surprising part about the results was how the two analyses yielded significantly different results (models, strength of successes, failure analyses). One challenge we encountered was defining what was a solid vs. a liquid, as discussed above. A second challenge was in

understanding the differing definitions of junk food. As noted, NVA identified significantly fewer foods that were defined as junk food, as opposed to NDA. With more time, more of the differences between the models would be explored. For example, a question that could be explored is 'Why were certain features more important than others under the different supervised models?' It also would be interesting to include data that looked at packaged foods.

Part B:
The most surprising part about the results was the Principal Component loadings, specifically for the liquids; based on prior knowledge on the topic, it was not expected that magnesium would be one of the more important features in constructing the components. Some challenges were selecting the optimal number of clusters for the K-Means model; this was mitigated by using silhouette scores to determine the best number of clusters. With more time, a deeper analysis into the differences found in the clusters for liquids and solids could be done, such as why the cluster quality metrics for the solids yield significantly different results than for the liquids. Additionally, an analysis on the effect of additional Principal Components could be conducted.

## Ethical Considerations

For both Supervised and Unsupervised approaches, from a public health perspective, labeling food as junk may contribute to food stigma and disordered eating behaviors, negatively impacting individuals' relationships with food. Additionally, socioeconomic factors must be considered, as many communities rely on processed foods due to affordability and accessibility. A classification model that does not account for these disparities may reinforce food inequity. Ethical concerns also extend to data privacy, industry influence, and commercial misuse, where biased classifications could be exploited for marketing or regulatory decisions. Moreover, reliance on thresholds for sugar, fat, or sodium content without considering portion size and dietary context may lead to misleading classifications.

## Statement of Work

Ayan, Alex, and Richard all performed EDA on the dataset. This included the basic understanding of what data we had available and how to analyze that data. Most importantly, all 3 of us worked to come up with a common application of the liquid vs solid classification. With the data we had, there was a significant amount of analysis to understand how best to classify. **Ayan Banerjee** led the Supervised analysis of the Nutrient Value approach. **Richard Chalker** led the Supervised analysis of the Nutrient Density approach. **Alex DeLoach** led the Unsupervised analysis.

## Future Scope

Future work could include expanding the dataset to incorporate a broader variety of food items, including regional and specialty cuisines, to improve the model's generalizability. Additionally, integrating more detailed nutritional data (e.g., processing level) could enable more nuanced categorizations and health recommendations.

**SIADS Milestone II - Winter, 2025**
**Leveraging Machine Learning to Classify Junk Food vs. Regular Food**
Ayan Banerjee, Richard Chalker, Alex DeLoach

**References**


a. **American Heart Association.** (n.d.). *Unhealthy foods.* Heart.org. Retrieved February 22, 2025, from www.heart.org/en/healthy-living/go-red-get-fit/unhealthy-foods
b. **Drewnowski, A., & Fulgoni, V. L. III.** (2014). *Nutrient density: Principles and evaluation tools. The American Journal of Clinical Nutrition, 99*(5), 1223S-1228S. ScienceDirect. Retrieved from https://www.sciencedirect.com/science/article/pii/S0002916523050748
c. **Dunford, E. K., Popkin, B., & Ng, S. W.** (2022). *Nutrient profiling: Is the technique family of nutrient profile models? The Journal of Nutrition, 152*(2), 492-500. ScienceDirect. Retrieved from https://www.sciencedirect.com/science/article/pii/S0022316622005399
d. **European Food Information Council.** (n.d.). *What is nutrient density?* Retrieved February 22, 2025, from https://www.eufic.org/en/understanding-science/article/what-is-nutrient-density#:~:text=The%20nutrient%20density%20of%20foods%20is%20determined%20by%20what%20we,recommendations%2C%20assigning%20them%20a%20score
e. **Kirk, D., Kok, E., Tufano, M., Tekinerdogan, B., Feskens, E. J. M., & Camps, G.** (2022). Machine learning in nutrition research. *Advances in Nutrition, 13*(6), 2573–2589. https://doi.org/10.1093/advances/nmac103
f. **Kumar, N.** (2022, March 22). *A real-time junk food recognition system based on machine learning.* Academia.edu. Retrieved from https://www.academia.edu/109150798/A_Real_Time_Junk_Food_Recognition_System_Based_on_Machine_Learning
g. **National Institutes of Health.** (n.d.). *Junk food intake among adults in the United States.* National Library of Medicine. Retrieved February 22, 2025, from https://pmc.ncbi.nlm.nih.gov/articles/PMC8826924/#abstract1
h. **United States Department of Agriculture (USDA).** (n.d.). *USDA Food Surveys Research Group.* Agricultural Research Service. Retrieved February 22, 2025, from https://www.ars.usda.gov/northeast-area/beltsville-md-bhnrc/beltsville-human-nutrition-research-center/food-surveys-research-group/docs/fndds-download-databases/
i. **WebMD.** (n.d.). *Junk food facts.* Retrieved February 22, 2025, from www.webmd.com/diet/features/junk-food-facts

**SIADS Milestone II - Winter, 2025**
**Leveraging Machine Learning to Classify Junk Food vs. Regular Food**
Ayan Banerjee, Richard Chalker, Alex DeLoach

## Appendix

1) EDA Analysis

We found that there are no missing values in the dataset.

Figure 20 shows the distribution of energy (kcal) in different food items. The distribution appears right-skewed, indicating that most food items have a moderate caloric content, with a few high-calorie outliers.

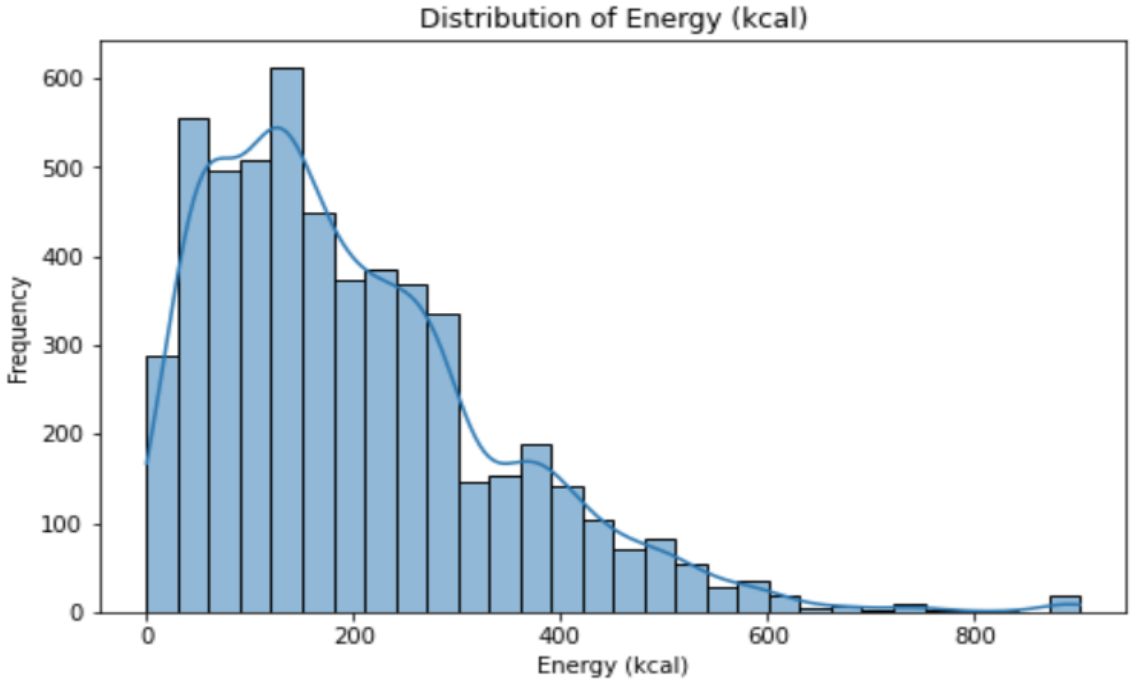


*Figure 20: Distribution of Energy (kcal)*

Figure 21 illustrates the distribution of key macronutrients: protein, carbohydrates, and fat. The distributions suggest that carbohydrate content varies widely across different food items, whereas protein and fat have relatively narrower distributions.

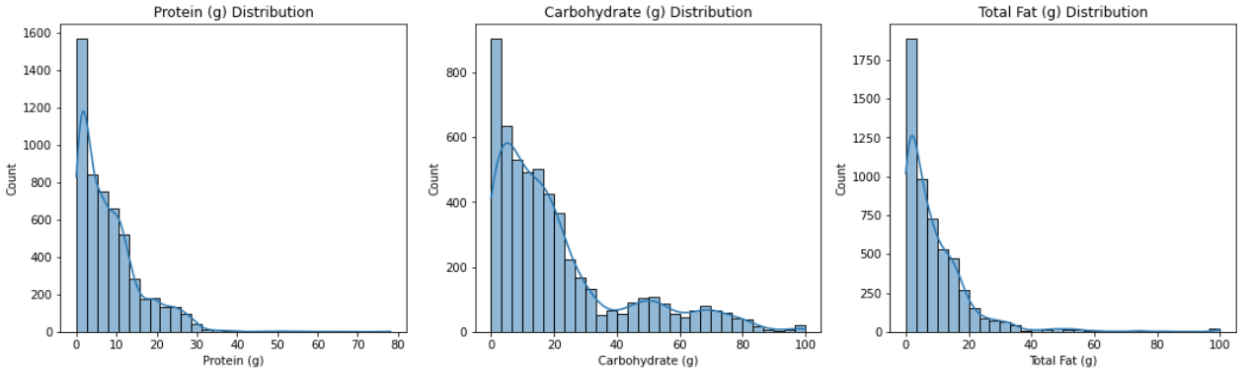


*Figure 21: Distribution of Macronutrients*

Figure 22 presents the correlation matrix of various nutrients. Strong correlations can be observed between carbohydrate and energy content, as well as fat and energy content, which is expected. These correlations help understand the interdependence among nutrients in different food items.

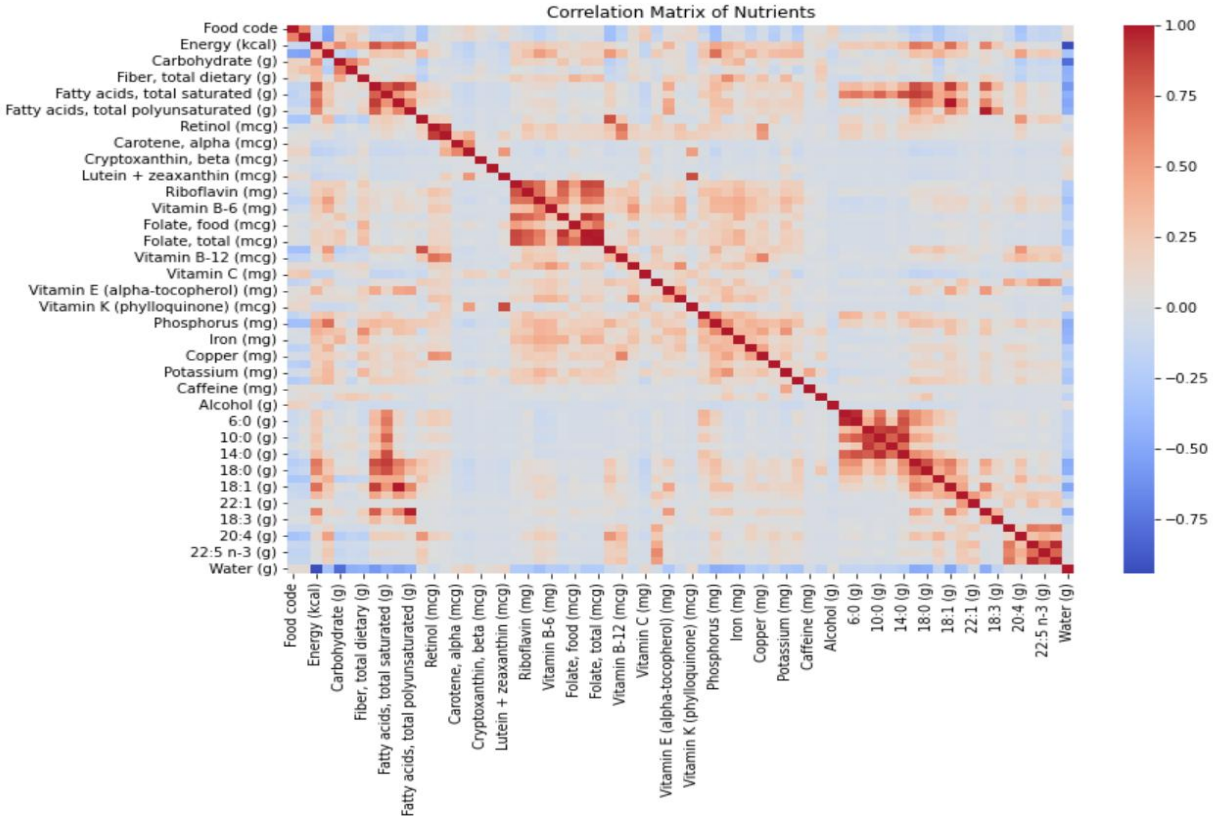


*Figure 22: Correlation Matrix of Nutrients*

2) Representative models for Nutrient Density Analysis (Drewnowski et al)

TABLE 1. Algorithms for NR*n* and LIM subscores and for the composite NRF nutrient profile models[1]

| Model | Algorithm | Reference amount | Comment |
|---|---|---|---|
| Subscores NR*n* | | | |
| NR*n*_100 g | $\sum_{1-n}(\text{Nutrient}_i/\text{DV}_i) \times 100$ | 100 g | $\text{Nutrient}_i$ = content of nutrient *i* in 100 gDV = daily value |
| NR*n*_100 kcal | (NR*n*_100 g/ED) × 100 | 100 kcal | ED = energy density (kcal/100 g) |
| NR*n*_RACC | (NR*n*_100 g/100) × RACC | Serving | RACC = FDA serving size |
| Subscores LIM | | | |
| LIM_100 g | $\sum_{1-3}(L_i/\text{MRV}_i) \times 100$ | 100 g | $L_i$ = content of limiting nutrient *i* in 100gMRV = maximum recommended value |
| LIM_100 kcal | (LIM_100 g/ED) × 100 | 100 kcal | ED = energy density (kcal/100 g) |
| LIM_RACC | (LIM_100 g/100) × RACC | Serving | RACC = FDA serving size |
| Composite NRF*n*0.3 | | | |
| NRF*n*.3_sum | NR*n*_100 kcal – LIM_100 kcal | 100 kcal | Difference between sums |
| NRF*n*.3_mean | NR*n*/*n* – LIM/3 | 100 kcal | Difference between means |
| NRF*n*.3_ratio | NR*n*/LIM[2] | None | Ratio |

*Figure 23: Representative models for Nutrient Density Analysis*

3) Features in the data

| **Features** | | |
|---|---|---|
| Energy (kcal) | Folate, food (mcg) | Alcohol (g) |
| Protein (g) | Folate, DFE (mcg_DFE) | 4:0(g) |
| Carbohydrate (g) | Folate, total (mcg) | 6:0(g) |
| Sugars, total(g) | Choline, total (mg) | 8:0(g) |
| Fiber, total dietary (g) | Vitamin B-12 (mcg) | 10:0(g) |
| Total Fat (g) | Vitamin B-12, added(mcg) | 12:0(g) |
| Fatty acids, total saturated (g) | Vitamin C (mg) | 14:0(g) |
| Fatty acids, total monounsaturated (g) | Vitamin D (D2 + D3) (mcg) | 16:0(g) |
| Fatty acids, total polyunsaturated (g) | Vitamin E (alpha-tocopherol) (mg) | 18:0(g) |
| Cholesterol (mg) | Vitamin E, added(mg) | 16:1(g) |
| Retinol (mcg) | Vitamin K (phylloquinone) (mcg) | 18:1(g) |
| Vitamin A, RAE (mcg_RAE) | Calcium (mg) | 20:1(g) |
| Carotene, alpha (mcg) | Phosphorus (mg) | 22:1(g) |
| Carotene, beta (mcg) | Magnesium (mg) | 18:2(g) |
| Cryptoxanthin, beta (mcg) | Iron(mg) | 18:3(g) |
| Lycopene (mcg) | Zinc(mg) | 18:4(g) |
| Lutein + zeaxanthin (mcg) | Copper (mg) | 20:4(g) |
| Thiamin (mg) | Selenium (mcg) | 20:5 n-3(g) |
| Riboflavin (mg) | Potassium (mg) | 22:5 n-3(g) |
| Niacin (mg) | Sodium (mg) | 22:6 n-3(g) |
| Vitamin B-6 (mg) | Caffeine (mg) | Water(g) |
| Folic acid (mcg) | Theobromine (mg) | |

*Figure 24: Dataset features*

4) NDA Analysis Food Distribution

In running the NDA analysis, we see the top foods to be:

| Most Healthy Foods | | |
|---|---|---|
| | **Main food description** | **NRF9.3** |
| 5107 | Tea, hot, leaf, green, decaffeinated | inf |
| 5142 | Tea, iced, brewed, green, decaffeinated, unswe... | inf |
| 5328 | Water, enhanced, diet | inf |
| 4255 | Watercress, raw | 5,788.85 |
| 4989 | Coffee, NS as to type | 5,213.80 |
| 4990 | Coffee, NS as to brewed or instant | 5,213.80 |
| 4991 | Coffee, brewed | 5,213.80 |
| 4996 | Coffee, brewed, flavored | 5,213.80 |
| 5086 | Coffee and chicory, brewed | 5,213.80 |
| 4186 | Beet greens, raw | 4,902.05 |

*Figure 25: Top healthy foods as per NDA*

The least healthy foods are:

| Least Healthy Foods | | |
|---|---|---|
| | **Main food description** | **NRF9.3** |
| 5147 | Tea, iced, bottled, black, unsweetened | inf |
| 4367 | Buffalo sauce | (24,437.36) |
| 1073 | Fish sauce | (20,966.58) |
| 2058 | Soy sauce | (9,322.80) |
| 2059 | Soy sauce, reduced sodium | (5,501.80) |
| 1748 | Soup, broth | (5,470.87) |
| 4711 | Peppers, hot, pickled | (5,417.01) |
| 1766 | Oyster sauce | (5,178.65) |
| 4694 | Pickles, dill | (4,488.78) |
| 4713 | Pickles, NFS | (4,488.78) |

*Figure 26: Top unhealthy foods as per NDA*

5) Feature ablation study for nutrient value-based analysis

| Removed Features | Accuracy | F1 Score |
|---|---|---|
| Alcohol (g) | 0.998883 | 0.998870 |
| 12:0\n(g) | 0.998883 | 0.998870 |
| Niacin (mg) | 0.998883 | 0.998870 |
| 18:0\n(g) | 0.998883 | 0.998870 |
| Fatty acids, total polyunsaturated (g) | 0.997765 | 0.997713 |
| Vitamin B-12, added\n(mcg) | 0.997765 | 0.997713 |
| Selenium (mcg) | 0.997765 | 0.997713 |
| Zinc\n(mg) | 0.997765 | 0.997713 |
| Vitamin A, RAE (mcg_RAE) | 0.997765 | 0.997713 |
| 14:0\n(g) | 0.997765 | 0.997713 |
| Vitamin B-6 (mg) | 0.997765 | 0.997713 |
| Thiamin (mg) | 0.997765 | 0.997713 |
| Vitamin E, added\n(mg) | 0.997765 | 0.997713 |

*Figure 27: Feature ablation study on Solid type food (NVA)*

| Removed Features | Accuracy | F1 Score |
|---|---|---|
| Water\n(g) | 0.989583 | 0.989212 |
| Lutein + zeaxanthin (mcg) | 0.989583 | 0.989212 |
| Potassium (mg) | 0.989583 | 0.989212 |
| 20:5 n-3\n(g) | 0.989583 | 0.989212 |
| 16:1\n(g) | 0.989583 | 0.989212 |
| Vitamin E (alpha-tocopherol) (mg) | 0.989583 | 0.989212 |
| Fatty acids, total monounsaturated (g) | 0.989583 | 0.989212 |
| Vitamin B-12 (mcg) | 0.989583 | 0.989212 |
| 10:0\n(g) | 0.989583 | 0.989212 |
| Copper (mg) | 0.989583 | 0.989212 |
| Choline, total (mg) | 0.989583 | 0.989212 |
| 16:0\n(g) | 0.989583 | 0.989212 |
| 18:0\n(g) | 0.989583 | 0.989212 |

*Figure 28: Feature ablation study on Liquid type food (NVA)*

6) Feature ablation study for nutrient density-based analysis

| Removed Features | Accuracy | F1 Score |
|---|---|---|
| Potassium (mg)_per_100kcal | 0.949664 | 0.94962 |
| Magnesium (mg)_per_100kcal | 0.883669 | 0.883669 |
| Vitamin E, added\n(mg) | 0.884787 | 0.884756 |
| 22:1\n(g) | 0.872483 | 0.872442 |
| Vitamin B-6 (mg) | 0.869128 | 0.869104 |
| Caffeine (mg) | 0.845638 | 0.845514 |
| Lutein + zeaxanthin (mcg) | 0.817673 | 0.817457 |
| 12:0\n(g) | 0.815436 | 0.815347 |
| 20:1\n(g) | 0.777405 | 0.777398 |
| Selenium (mcg) | 0.781879 | 0.781857 |
| 20:4\n(g) | 0.788591 | 0.788406 |
| Phosphorus (mg) | 0.777405 | 0.777445 |
| Folic acid (mcg) | 0.729306 | 0.729289 |

*Figure 29: Feature ablation study on Solid type food (NDA)*

| Removed Features | Accuracy | F1 Score |
|---|---|---|
| Theobromine (mg) | 0.930481 | 0.928908 |
| 10:0\n(g) | 0.925134 | 0.923163 |
| Vitamin D (D2 + D3) (mcg) | 0.930481 | 0.928908 |
| Lycopene (mcg) | 0.925134 | 0.922581 |
| 8:0\n(g) | 0.925134 | 0.922581 |
| 18:3\n(g) | 0.925134 | 0.922581 |
| Zinc\n(mg) | 0.925134 | 0.922581 |
| Vitamin K (phylloquinone) (mcg) | 0.919786 | 0.917368 |
| Iron\n(mg)_per_100kcal | 0.930481 | 0.927231 |
| Riboflavin (mg) | 0.935829 | 0.933108 |
| Phosphorus (mg) | 0.925134 | 0.922581 |
| Calcium (mg)_per_100kcal | 0.914439 | 0.912186 |
| Alcohol (g) | 0.871658 | 0.862571 |

*Figure 30: Feature ablation study on Liquid type food (NDA)*

7) Silhouette scores

```
The average silhouette score for solids with 2 clusters is: 0.41839000241258756
The average silhouette score for solids with 3 clusters is: 0.32694560423837643
The average silhouette score for solids with 4 clusters is: 0.19213908097580468
The average silhouette score for solids with 5 clusters is: 0.20811089117087211
```

*Figure 31: Silhouette Scores for Solids*

```
The average silhouette score for liquids with 2 clusters is: 0.6971584512797478
The average silhouette score for liquids with 3 clusters is: 0.37693195745711194
The average silhouette score for liquids with 4 clusters is: 0.3734705480457939
The average silhouette score for liquids with 5 clusters is: 0.2332851183882263
```

*Figure 32: Silhouette Scores for Liquids*

## 8) Summary Statistics

| | Energy (kcal) | Sodium (mg) | Sugars, total\n(g) | Total Fat (g) |
|---|---|---|---|---|
| count | 942.0000 | 942.0000 | 942.0000 | 942.0000 |
| mean | 108.8885 | 138.2389 | 7.4765 | 5.6150 |
| std | 137.0140 | 296.0826 | 7.4947 | 14.0411 |
| min | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| 25% | 40.0000 | 8.0000 | 1.4600 | 0.0800 |
| 50% | 66.0000 | 36.0000 | 6.6500 | 1.1500 |
| 75% | 125.7500 | 104.7500 | 10.3475 | 3.7675 |
| max | 900.0000 | 5843.0000 | 73.4000 | 100.0000 |

| | Energy (kcal) | Sodium (mg) | Sugars, total\n(g) | Total Fat (g) |
|---|---|---|---|---|
| count | 17.0000 | 17.0000 | 17.0000 | 17.0000 |
| mean | 651.4118 | 424.6471 | 3.9706 | 69.8924 |
| std | 191.8538 | 476.7967 | 11.6893 | 23.9862 |
| min | 343.0000 | 0.0000 | 0.0000 | 33.4500 |
| 25% | 499.0000 | 21.0000 | 0.0000 | 55.1000 |
| 50% | 683.0000 | 450.0000 | 0.5800 | 75.3300 |
| 75% | 750.0000 | 524.0000 | 0.7700 | 82.2000 |
| max | 902.0000 | 2039.0000 | 48.5400 | 100.0000 |

*Figure 33: Liquid Statistics - Cluster 0 (Left) vs. Cluster 1 (Right)*

| | Energy (kcal) | Sodium (mg) | Sugars, total\n(g) | Total Fat (g) |
|---|---|---|---|---|
| count | 821.0000 | 821.0000 | 821.0000 | 821.0000 |
| mean | 410.5457 | 521.5079 | 11.7602 | 24.1872 |
| std | 112.5209 | 403.3034 | 15.3644 | 13.7030 |
| min | 74.0000 | 0.0000 | 0.0000 | 0.0000 |
| 25% | 319.0000 | 280.0000 | 0.7800 | 16.4100 |
| 50% | 403.0000 | 438.0000 | 4.1600 | 21.1800 |
| 75% | 487.0000 | 704.0000 | 20.5200 | 29.3400 |
| max | 892.0000 | 3668.0000 | 84.8300 | 100.0000 |

| | Energy (kcal) | Sodium (mg) | Sugars, total\n(g) | Total Fat (g) |
|---|---|---|---|---|
| count | 3651.0000 | 3651.0000 | 3651.0000 | 3651.0000 |
| mean | 172.9474 | 345.7998 | 5.5870 | 6.6095 |
| std | 93.0198 | 306.9351 | 11.9859 | 5.1132 |
| min | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| 25% | 107.0000 | 188.0000 | 0.6200 | 2.5800 |
| 50% | 160.0000 | 323.0000 | 1.7500 | 5.5800 |
| 75% | 235.0000 | 443.0000 | 4.1000 | 10.1250 |
| max | 465.0000 | 7851.0000 | 99.8000 | 24.7800 |

*Figure 34: Solid Statistics - Cluster 0 (Left) vs. Cluster 1 (Right)*