Unsupervised Nutritional Clustering using PCA and Kernel Methods

Osman ÖRNEK 370283 EPFL, Department of Mechanical Engineering osman.ornek@epfl.ch

Abstract

This project explores unsupervised learning techniques to group food items based on their nutritional content. Dimensionality reduction (PCA and Kernel PCA) was combined with KMeans clustering to uncover meaningful patterns. Internal evaluation metrics and visualization tools such as radar and box plots were used to analyze and interpret the results. The final cluster profiles revealed distinct food categories like high-fat, high-sugar, protein-rich, and water-heavy items, offering valuable insights for nutritional systems.

1. Introduction

Understanding the nutritional composition of foods is essential for dietary planning, health monitoring, and food education. The dataset used in this project, sourced from Kaggle[1], provides detailed macro- and micronutrient information for a wide variety of food items commonly consumed across the globe. It is designed to support nutritional analysis and promote informed dietary choices.

The objective of this project is to apply unsupervised machine learning techniques to uncover patterns within this nutritional dataset. Since the data is unlabeled, the analysis focuses on discovering intrinsic structure and similarity between food items based solely on their nutritional profiles.

Specifically, the project aims to:

- Identify natural groupings of foods based on nutritional similarity
- Determine the most and least informative nutrient features
- Detect nutritional outliers and anomalies
- Generate insights that may support dietary categorization or recommendation systems

By combining dimensionality reduction methods with clustering techniques, and supporting them with visual analytics, the project seeks to explore whether meaningful and interpretable nutritional categories can emerge from the data.

2. Methodology

2.1. Initial Data Analysis

The dataset contains nutritional profiles for a total of 2,395 food items from a wide variety of categories. Each entry includes quantitative nutritional values per 100g of product. The features fall into the following categories:

- **Macronutrients:** Calories, Protein, Fat, Carbohydrates, Fiber, Sugars
- **Micronutrients:** Calcium, Iron, Sodium, Zinc, Potassium, Selenium, and others
- **Derived Metric:** Nutrition Density

The nutrients are expressed in varying units (e.g., grams, milligrams, micrograms), requiring careful consideration during comparison and analysis.

	Fat	Carbohydrates	Sugars	Protein	Dietary Fiber	Water
count	2395.000000	2395.000000	2395.000000	2395.000000	2395.000000	2395.000000
mean	10.176276	18.589021	4.457459	13.400777	2.235790	83.814906
std	29.008915	29.406134	13.339929	32.294246	5.404483	117.121124
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
50%	2.100000	6.800000	0.086000	3.500000	0.200000	39.600000
max	550.700000	390.200000	291.500000	560.300000	76.500000	1875.900000

Figure 1: Summary statistics for macronutrients

Initial statistical exploration included computation of key descriptive statistics (mean, median, standard deviation, min, max) for all macronutrients. These results, summarized in Figure 1, revealed wide variation in nutrient values across food items.

Notably, several macronutrient values exceeded 100g per 100g of food, which is physically implausible and suggests inconsistencies in data entries. This observation motivated further data cleaning steps to ensure consistency and reliability prior to clustering.

Although the dataset contains both macro- and micronutrients, only macronutrients were retained for clustering: *Fat, Carbohydrates, Sugars, Protein, Dietary Fiber, and Water.* These features have greater interpretability in dietary contexts and allow for more coherent and meaningful grouping of food items based on nutritional content.

2.2. Data Preprocessing

To ensure the quality and physical plausibility of the data, a series of preprocessing steps were applied:

- **Column Selection:** Removed non-numeric columns such as food names for numerical analysis

[1] https://www.kaggle.com/datasets/utsavdey1410/food-nutrition-dataset/data

- **Missing Values:** Excluded rows with missing or zero-dominant nutrient values
- **Mass Consistency Filtering:** Removed entries where the sum of all gram-based macronutrients exceeded 100g, violating physical mass balance
- **Fat Composition Check:** Computed the sum of fat subcomponents (Saturated + Monounsaturated + Polyunsaturated Fats) and removed rows where this exceeded total fat, indicating internal inconsistency
- **Standardization:** Applied z-score normalization to all numeric features to ensure equal contribution to distance metrics during clustering
- **Outlier Treatment:** Outliers were identified using z-score thresholds and IQR ranges; some were excluded for robustness

These cleaning steps were essential to guarantee meaningful and interpretable results in the subsequent clustering stages. They also helped ensure that the derived patterns reflect real-world nutritional structure rather than data artifacts.

2.3. Dimensionality Reduction and Clustering Strategy

To uncover structure in the nutritional dataset, a two-stage pipeline was adopted:

1. Dimensionality reduction via PCA or Kernel PCA
2. Clustering in the reduced space using KMeans

Case 1: PCA + KMeans applies linear projection to preserve maximum variance along orthogonal axes before clustering. This serves as a baseline for evaluating structure in the original linear space.

Case 2: Kernel PCA + KMeans uses a non-linear mapping via the Radial Basis Function (RBF) kernel to uncover potentially hidden curved structures. Data are projected into a higher-dimensional space where clustering via KMeans may better separate complex groups.

In both cases, **KMeans clustering** was applied to the reduced representations. The number of clusters k was treated as a hyperparameter and selected using internal evaluation metrics:

- **Elbow Method:** Based on KMeans inertia (within-cluster variance)
- **Silhouette Score:** Assesses average intra-cluster cohesion vs. inter-cluster separation (↑ better)
- **Davies-Bouldin Index:** Measures cluster compactness and overlap (↓ better)
- **Calinski-Harabasz Index:** Captures between- vs. within-cluster dispersion (↑ better)

For Kernel PCA, a grid search over the kernel parameter γ was performed. Each (γ, k) pair was evaluated using the above metrics, with Silhouette Score guiding final selection.

All features were standardized prior to clustering to ensure scale comparability. The final hyperparameter choices and metric outcomes are discussed in the Results section.

3. Results and Discussion

The analysis focuses on macronutrient-based clustering, as outlined in Section 2.1.

3.1. PCA + KMeans Clustering

To reduce redundancy and noise in the nutritional feature space, PCA was applied, projecting the data into a lower-dimensional space while preserving most of its variance. The cumulative explained variance indicated that the first five components captured over 90% of the total variance, and these were retained for clustering.

KMeans clustering was then applied to the PCA-reduced space. A grid search over k values from 2 to 9 was performed, and the optimal number of clusters was selected using the elbow method and silhouette score.

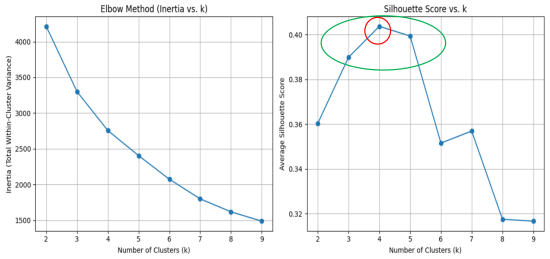

Figure 2: Elbow method and silhouette score analysis for PCA-reduced space.

As shown in Figure 2, the best performance was observed at $k = 4$, based on the combination of internal metrics. For robustness, $k = 3$ and $k = 5$ were also considered in the analysis.

k	Silhouette Score (↑)	Davies-Bouldin (↓)	Calinski-Harabasz (↑)
3	0.389898	1.334406	276.775471
4	0.403723	1.083650	281.915238
5	0.399412	1.069624	276.113947

Figure 3: Internal evaluation metrics (Silhouette, Davies-Bouldin, Calinski-Harabasz) for $k \in \{3, 4, 5\}$.

The final number of clusters was selected as $k = 4$, based on internal evaluation metrics. Clustering results were first visualized using the first two principal components (PC1 vs PC2), as shown in Figure 4. This 2D projection provides a general sense of cluster separation and structure.

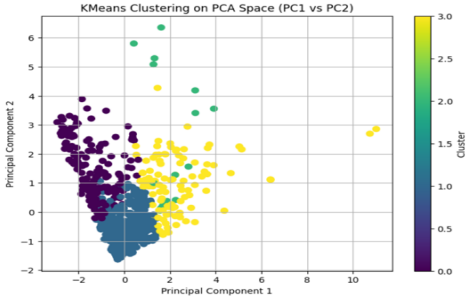

Figure 4: KMeans clustering result visualized in PCA space (PC1 vs PC2).

While useful for confirming the presence of structure, this plot has limited interpretability because the principal components are linear combinations of original features. It does not reveal which specific nutrients drive cluster differences. Therefore, more informative plots such as radar and box plots were used for nutritional interpretation.

To provide a clearer, nutrient-level understanding of each cluster, radar plots were generated using both unscaled and scaled nutrient values (Figure 5). The unscaled plot reflects absolute nutrient quantities, while the scaled version normalizes feature ranges, which may better account for differences in biological relevance or unit disparity.

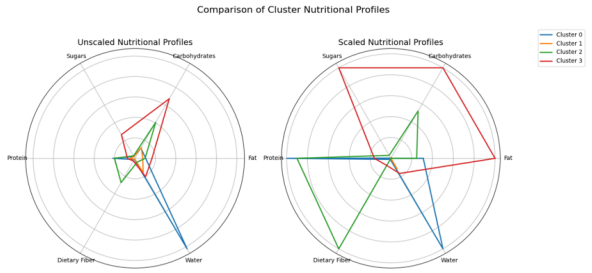

Figure 5: Radar plots showing cluster-level nutrient profiles (left: unscaled, right: scaled).

Based on the scaled radar plot, the clusters can be interpreted as follows:

- **Cluster 0:** High in water and protein, suggesting hydrating, protein-rich items such as dairy or broth-based foods.
- **Cluster 1:** Uniformly low across all nutrients, likely representing diet/light foods or nutritionally minimal items.
- **Cluster 2:** High in fiber and protein with minimal water, indicative of dry, dense foods such as snack bars or cereals.
- **Cluster 3:** Extremely high in fat, carbohydrates, and sugars — likely energy-dense, sweet and fatty products such as chocolates, desserts, or processed snacks.

To further validate these interpretations, macronutrient distributions were visualized using box plots for each cluster (Figure 6). Box plots show the median, quartiles, and outliers for each nutrient, offering a more detailed view of within-cluster variability.

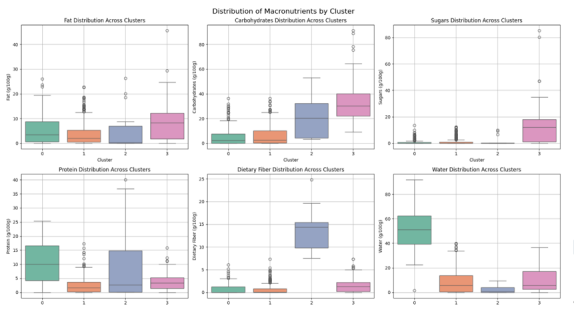

Figure 6: Box plots of macronutrient distributions across clusters.

Box plots confirm the trends observed in the radar charts. Cluster 3 clearly stands out with high fat and sugar content, supporting its identification as a group of energy-dense foods. Cluster 2 is fiber- and protein-rich, while Cluster 0 contains water-heavy items. Cluster 1 consistently shows low nutrient values. These insights provide a concrete nutritional basis for the discovered structure.

3.2. Kernel PCA + KMeans Clustering

While PCA-based clustering yielded interpretable groups, it relied on linear assumptions—PCA captures variance along linear axes, and KMeans identifies spherical clusters. Such assumptions may not hold in complex datasets like nutritional profiles, which may exhibit nonlinear relationships.

To capture these patterns, Kernel PCA (KPCA) with an RBF kernel was applied, projecting the data into a higher-dimensional space where non-linear structures could become linearly separable. The RBF kernel was chosen for its general applicability and effectiveness.

Since KPCA does not yield interpretable variance ratios, the number of components was fixed to five, matching the PCA case for consistency.

A grid search over the kernel parameter γ and cluster count k was conducted, evaluated using Silhouette, Davies-Bouldin, and Calinski-Harabasz scores. The best-performing configurations are shown in Figure 7.

Scenario	Gamma	k	Reasoning
Balanced Choice	1	5 or 6	Metrics are strong, and the number of clusters is interpretable
Highest Scores	10.0	7 or 8	Scores are at their peak, but interpretability and cluster meaning are risky

Figure 7: Grid search results for kernel parameter γ and cluster number k in KPCA + KMeans.

The best trade-off was achieved using $\gamma = 1$ and $k = 6$, balancing interpretability and clustering performance. Subsequent analysis and interpretation are based on this configuration.

The result of clustering in the first two kernel principal components is shown in Figure 8.

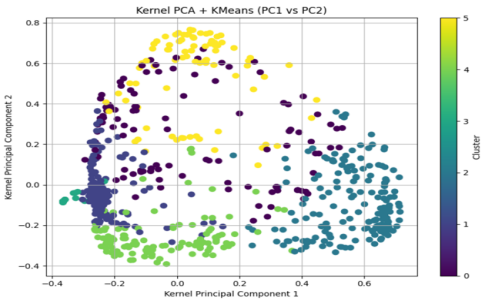

Figure 8: Cluster separation visualized in Kernel PCA space (PC1 vs PC2).

While the visual separation of clusters is evident, these principal components remain abstract constructs not di-

rectly tied to specific nutrients. Hence, feature-level interpretation was again performed using radar plots.

Figure 9 displays the scaled and unscaled nutrient profiles across clusters. This allows clearer identification of dominant nutrient characteristics in each group.

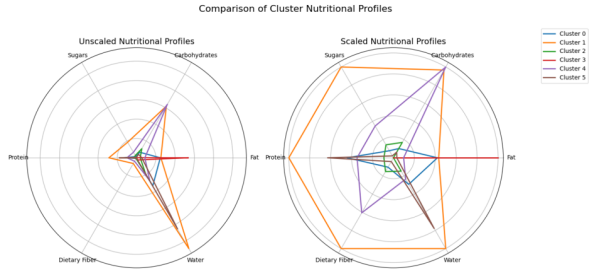

Figure 9: Radar plots showing nutrient profiles per cluster (left: unscaled, right: scaled) for KPCA-based clustering.

Cluster interpretation based on the scaled profiles is as follows:

- **Cluster 0:** Balanced nutrient profile with moderate levels of protein, fat, and fiber. Likely represents general-purpose food items.
- **Cluster 1:** Extremely high in all nutrients — likely multi-component, dense meals or processed products.
- **Cluster 2:** Uniformly low across all nutrients — represents light or minimal foods.
- **Cluster 3:** Dominated by fat — indicative of oils, butters, or fatty condiments.
- **Cluster 4:** High in carbohydrates — consistent with sweet snacks, baked goods, or desserts.
- **Cluster 5:** High in water and protein — likely dairy-based or hydrating, protein-rich items.

To complement the radar plots, box plots were generated to visualize the distribution of macronutrients across the six clusters (Figure 10). These plots reveal not only the central tendency (median) but also the variability and presence of outliers for each nutrient within each cluster.

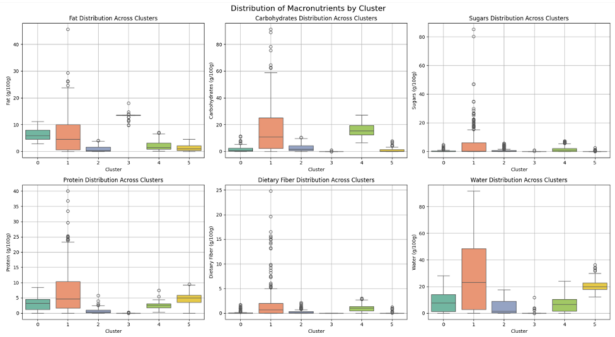

Figure 10: Box plots showing macronutrient distributions across clusters (KPCA + KMeans).

The box plots highlight key differences in nutrient composition across the clusters:

- **Cluster 1** shows the highest median and the widest range across almost all macronutrients, confirming its identification as a nutrient-dense group.
- **Cluster 3** appears significantly depleted across most macronutrients, likely representing minimal or fat-dominant items.
- **Cluster 4** is distinguished by its elevated carbohydrate and fiber content.
- **Cluster 5** is rich in both protein and water, in line with hydrating protein-rich foods.

These trends are consistent with the cluster interpretations obtained from the radar plots, providing additional validation for the nutrient-based groupings derived from KPCA and KMeans.

4. Conclusion

This study applied unsupervised learning methods to discover meaningful structures in food nutrition data. The initial approach, based on PCA followed by KMeans clustering, yielded moderately distinct clusters that were interpretable but limited by linear assumptions. Kernel PCA with an RBF kernel was subsequently employed to capture non-linear relationships, significantly enhancing cluster separation and structure.

The use of internal evaluation metrics — Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Score — quantitatively supported the improvement in clustering quality. Radar and box plots further enabled nutrient-level interpretation, revealing distinct profiles across clusters (e.g., fat-dominant, fiber-rich, protein-heavy, water-rich).

The final clusters reflect coherent nutritional categories, providing insight into underlying patterns in dietary composition. Kernel-based dimensionality reduction proved essential for uncovering complex, nonlinear trends not apparent through linear techniques.

These findings highlight the potential of kernel-based clustering in food analytics. The resulting insights can inform dietary planning, power recommendation systems, and support nutrition-focused applications aimed at personalized health and wellness.