EPFL

Machine Learning-II
Food Nutrition Dataset

Osman ÖRNEK
370283

Coding Competition

1. Objective

The goal of this project is to apply unsupervised learning methods to analyze the nutritional profiles of food items from around the world. Without any labels, the project aims to:

- Identify natural groupings of foods based on nutritional similarity,
- Determine which features are most and least informative,
- Detect outliers that deviate from common nutritional patterns,
- Generate insights for dietary categorization or recommendation systems.

2. Dataset

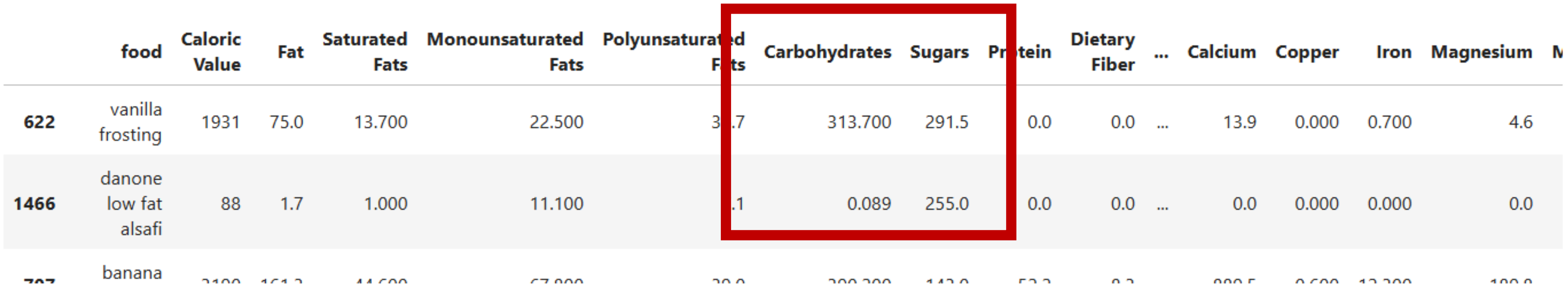
The dataset contains detailed nutritional information for a broad range of food items, including:

- Macronutrients: Calories, protein, fat, carbohydrates, fiber, sugars.
- Micronutrients: Calcium, iron, sodium, zinc, potassium, selenium, etc.
- Derived Metrics: Nutrition Density.

Units of nutrients vary (e.g., grams, milligrams); special care was taken to align comparisons meaningfully.

3. Preprocessing & Cleaning

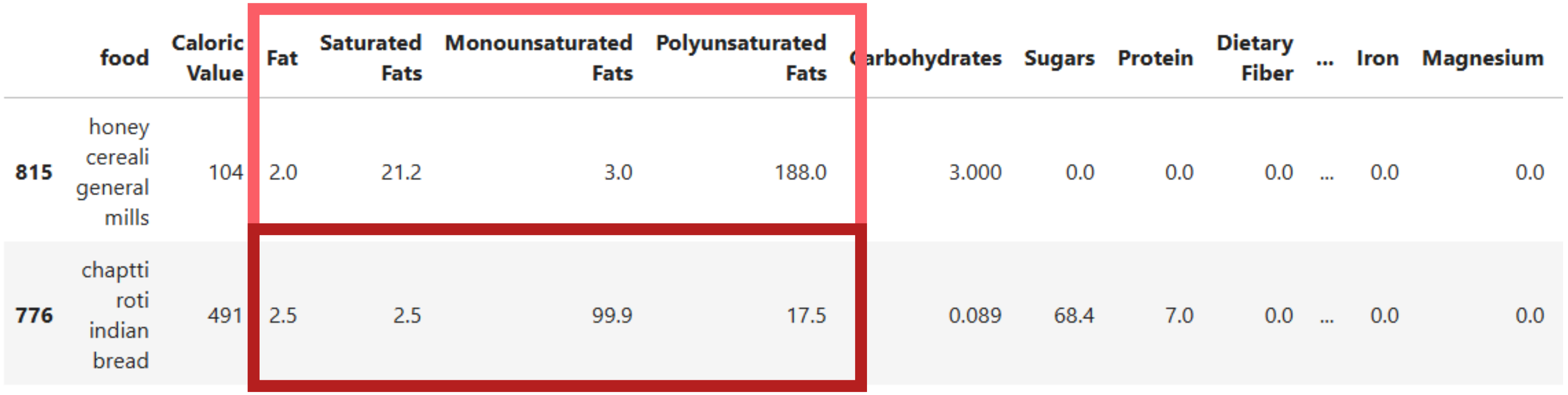
- Removed non-numeric columns (e.g., food name) for analysis.
- Handled missing or zero-dominant values.
- Mass Consistency Filtering: Removed entries where the sum of gram-based macronutrients exceeded 100g, violating physical consistency for a 100g portion.



	food	Caloric Value	Fat	Saturated Fats	Monounsaturated Fats	Polyunsaturated Fats	Carbohydrates	Sugars	Protein	Dietary Fiber	...	Calcium	Copper	Iron	Magnesium	M
622	vanilla frosting	1931	75.0	13.700	22.500	3[illegible].7	313.700	291.5	0.0	0.0	...	13.9	0.000	0.700	4.6	
1466	danone low fat alsafi	88	1.7	1.000	11.100	[illegible].1	0.089	255.0	0.0	0.0	...	0.0	0.000	0.000	0.0	
[illegible]	banana	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	

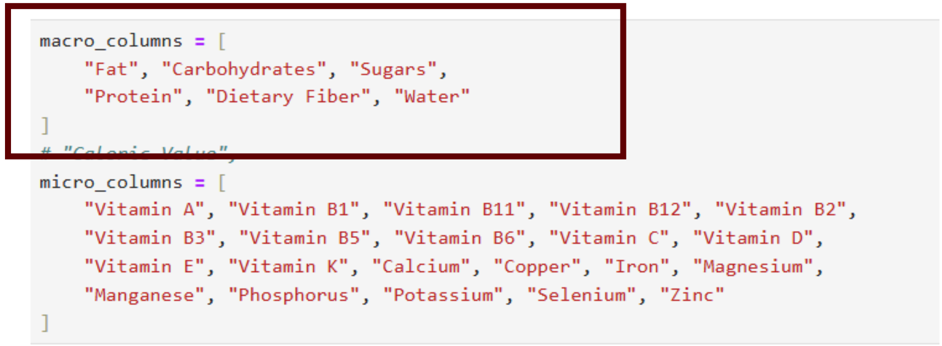
3. Preprocessing & Cleaning

- Fat Composition Consistency Check: Calculated the total of fat subtypes: Saturated Fats + Monounsaturated Fats + Polyunsaturated Fats. Removed entries where the total fat content was less than the sum of its subcomponents, indicating internal inconsistency.
- Standardized all features using z-score normalization to equalize scales.
- Investigated and optionally excluded outliers using z-score or IQR-based methods.



	food	Caloric Value	Fat	Saturated Fats	Monounsaturated Fats	Polyunsaturated Fats	Carbohydrates	Sugars	Protein	Dietary Fiber	...	Iron	Magnesium
815	honey cereali general mills	104	2.0	21.2	3.0	188.0	3.000	0.0	0.0	0.0	...	0.0	0.0
776	chaptti roti indian bread	491	2.5	2.5	99.9	17.5	0.089	68.4	7.0	0.0	...	0.0	0.0

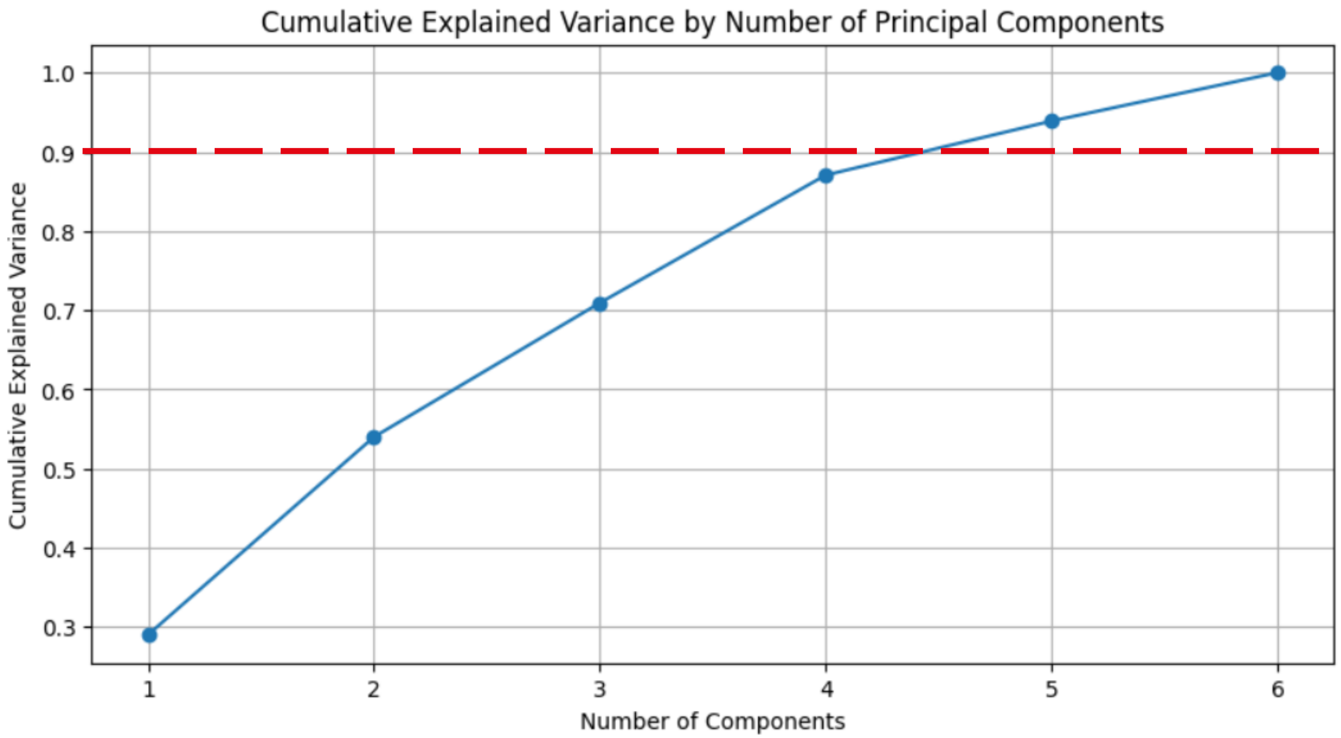
PCA + KMeans



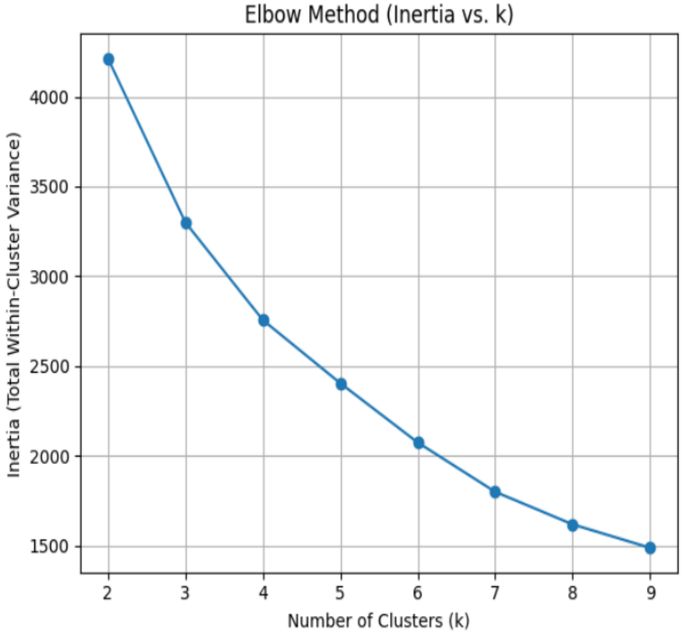
```
macro_columns = [
    "Fat", "Carbohydrates", "Sugars",
    "Protein", "Dietary Fiber", "Water"
]
# "Caloric Value",
micro_columns = [
    "Vitamin A", "Vitamin B1", "Vitamin B11", "Vitamin B12", "Vitamin B2",
    "Vitamin B3", "Vitamin B5", "Vitamin B6", "Vitamin C", "Vitamin D",
    "Vitamin E", "Vitamin K", "Calcium", "Copper", "Iron", "Magnesium",
    "Manganese", "Phosphorus", "Potassium", "Selenium", "Zinc"
]
```

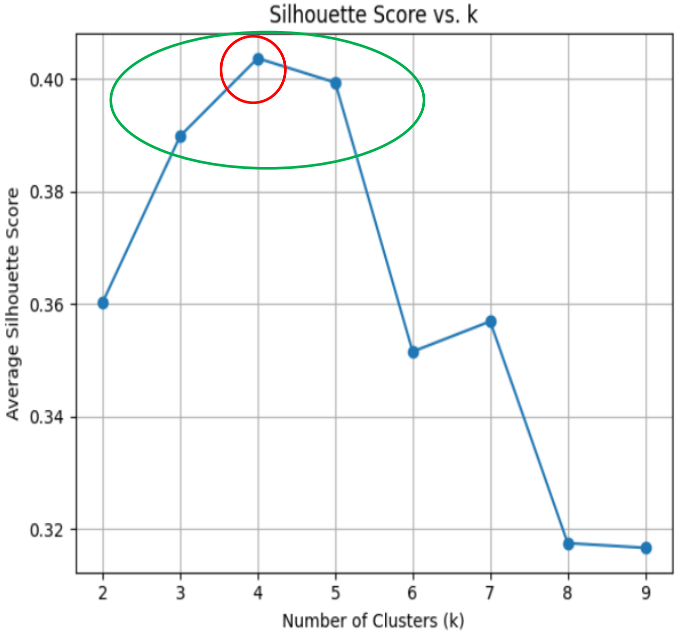
PCA

- Reduce noise and redundant information
- For covering 90% var → 5 component

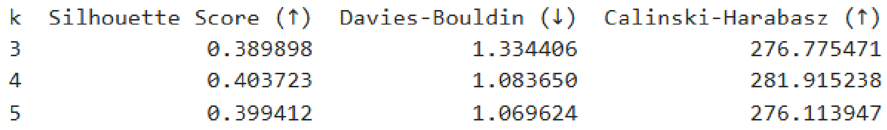


PCA + KMeans



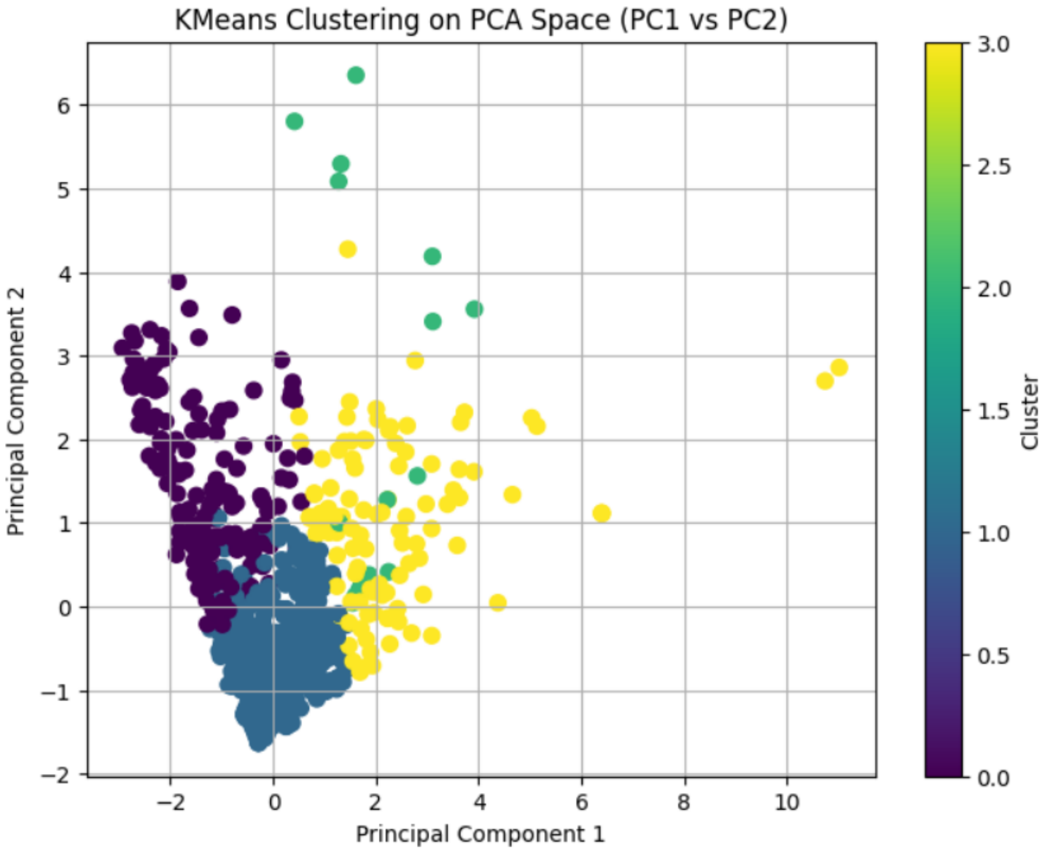


- Suitable k = 4
- Check also (k-1) and (k+1)



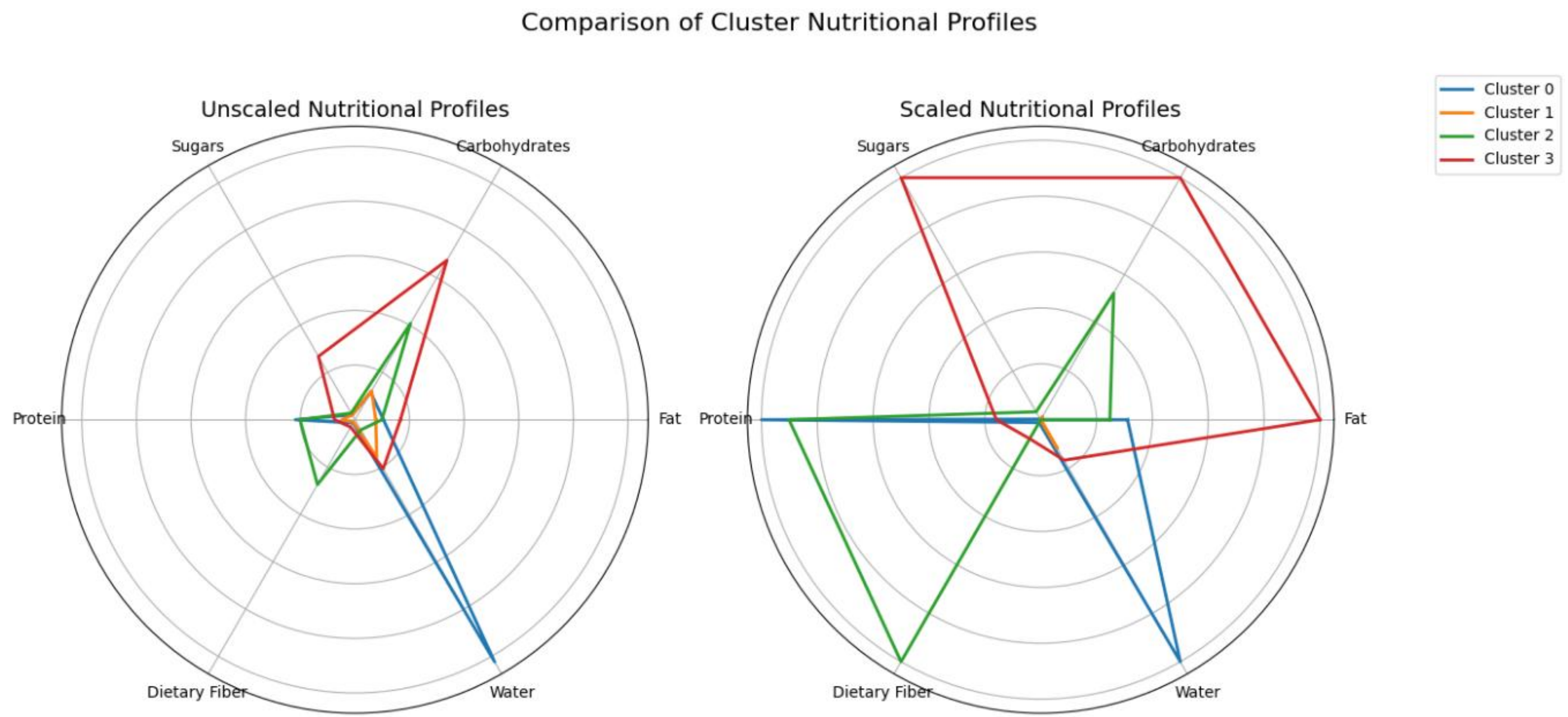
```
k  Silhouette Score (↑)  Davies-Bouldin (↓)  Calinski-Harabasz (↑)
3              0.389898            1.334406             276.775471
4              0.403723            1.083650             281.915238
5              0.399412            1.069624             276.113947
```

PCA + KMeans



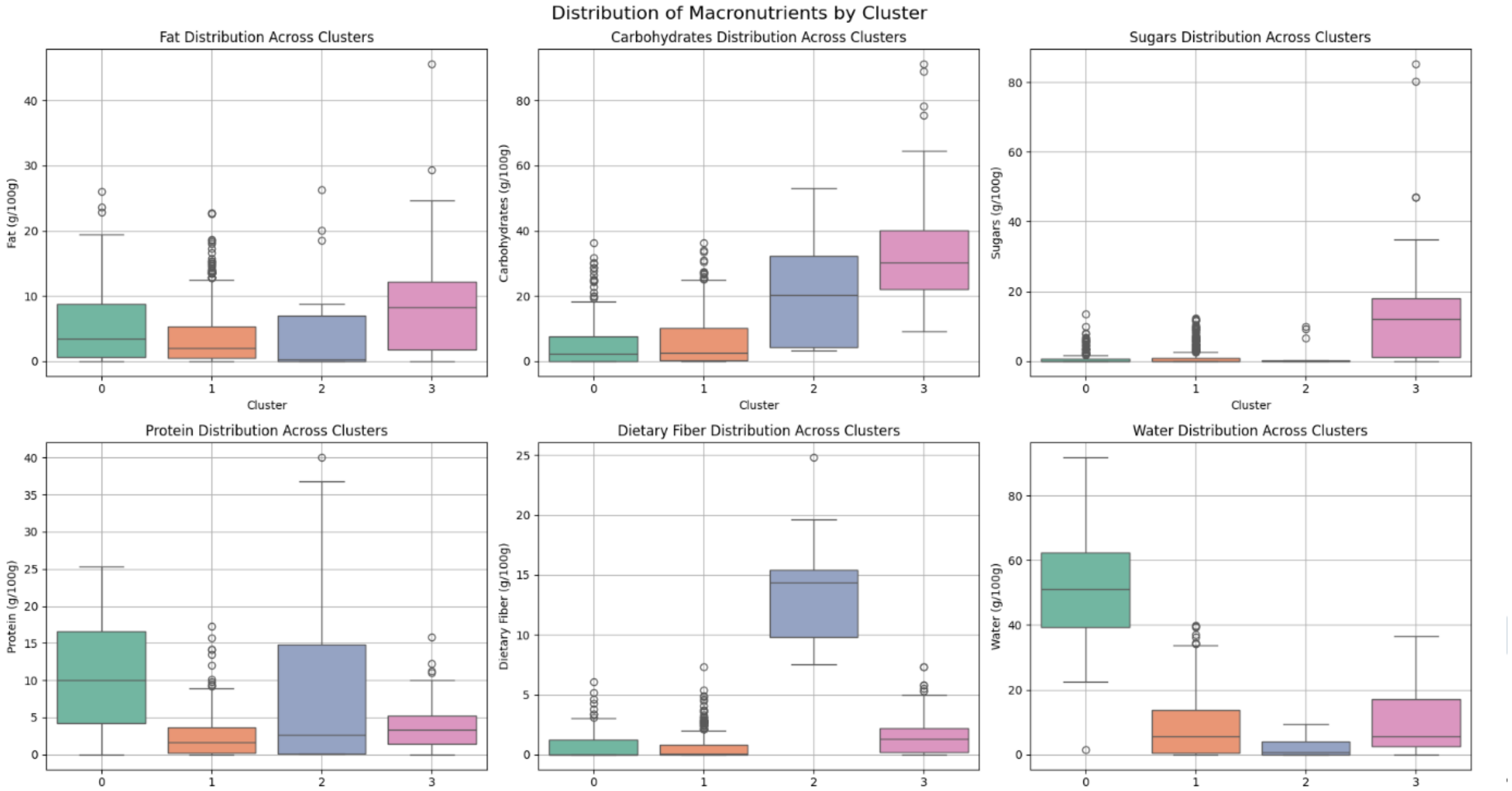
- Displays clustering result in the reduced PCA space (PC1 vs PC2)
- Shows overall structure and separation of clusters
- Helpful for visual validation of KMeans groupings
- **Limited interpretability:** PCs are combinations of features, not individual nutrients
- Does not reveal which features contribute to cluster differences
- More informative plots like radar and box plots were used for nutritional interpretation

PCA + KMeans Radar Plot



- Cluster 0 → Extreme water content and high protein — which might indicate dairy or broth-based products.
- Cluster 1 → Uniformly low across all nutrients —represents nutritionally minimal foods, possibly diet/light items or low-density products
- Cluster 2 → High fiber and protein, with minimal water — likely to represent dense snack items.
- Cluster 3 → Extremely high in fat, carbohydrates, and sugars — represents energy-dense, sweet and fatty products like chocolates, snacks, and processed desserts

PCA + Kmeans Box Plot



Distribution of Macronutrients by Cluster

Fat Distribution Across Clusters

Carbohydrates Distribution Across Clusters

Sugars Distribution Across Clusters

Protein Distribution Across Clusters

Dietary Fiber Distribution Across Clusters

Water Distribution Across Clusters

Box plots reveal distinct nutrient distributions across clusters. Cluster 3 clearly stands out with high fat and sugar content, suggesting energy-dense foods, while Cluster 2 is fiber- and protein-rich, indicative of nutrient-dense items. Cluster 0 appears to contain water-heavy, protein-rich foods, whereas Cluster 1 consistently has low nutrient values.

Transition: From Linear Clustering to Kernel-Based Methods

- Meaningful and well-separated clusters were obtained based on nutritional content using **PCA** followed by **KMeans clustering**.
- Supported both visually and statistically using **radar plots** and **box plots**.

However, these methods rely entirely on **linear assumptions**:

- **PCA** explains variance along linear axes.
- **KMeans** separates clusters based on **Euclidean distances**.
- As a result, they are only effective for **spherically and linearly separable** clusters.

Why Consider Kernel-Based Methods?

In complex datasets like nutritional profiles:

- There may be **hidden, nonlinear structures**(e.g., foods with similar protein content but varying fiber/fat composition).
- These patterns are often **missed by linear methods** like PCA and KMeans.
- **Kernel methods** can project the data into a **higher-dimensional, warped space**, where such nonlinear relationships become **linearly separable**, allowing for better clustering and interpretation.

Kernel PCA + KMeans

How to Decide on the Number of Components?

Unlike linear PCA, we cannot directly use explained variance to choose components in Kernel PCA, since the eigenvalues have a different interpretation.

However, in practice, a similar number of components as in linear PCA is typically chosen.(In our case: n_components = 5.)

How to Decide on the Number of Clusters (k)?

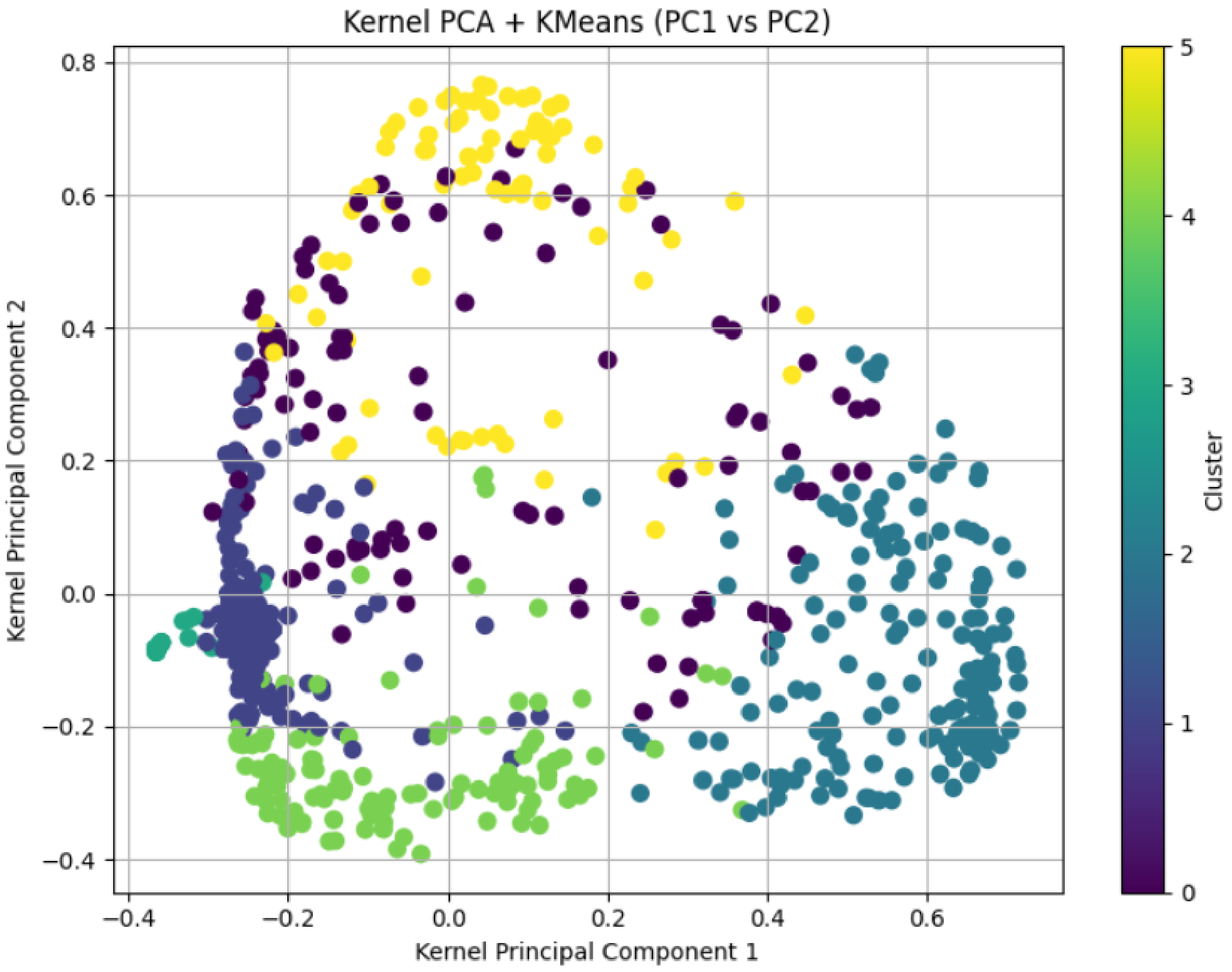
After applying KMeans to the KPCA output, used three evaluation metrics:

- Silhouette ScoreDavies
- Bouldin IndexCalinski
- Harabasz Index

To optimize both the number of clusters (k) and the kernel parameter (gamma). Performed a 2D grid search and selected the best configuration based on these metrics.

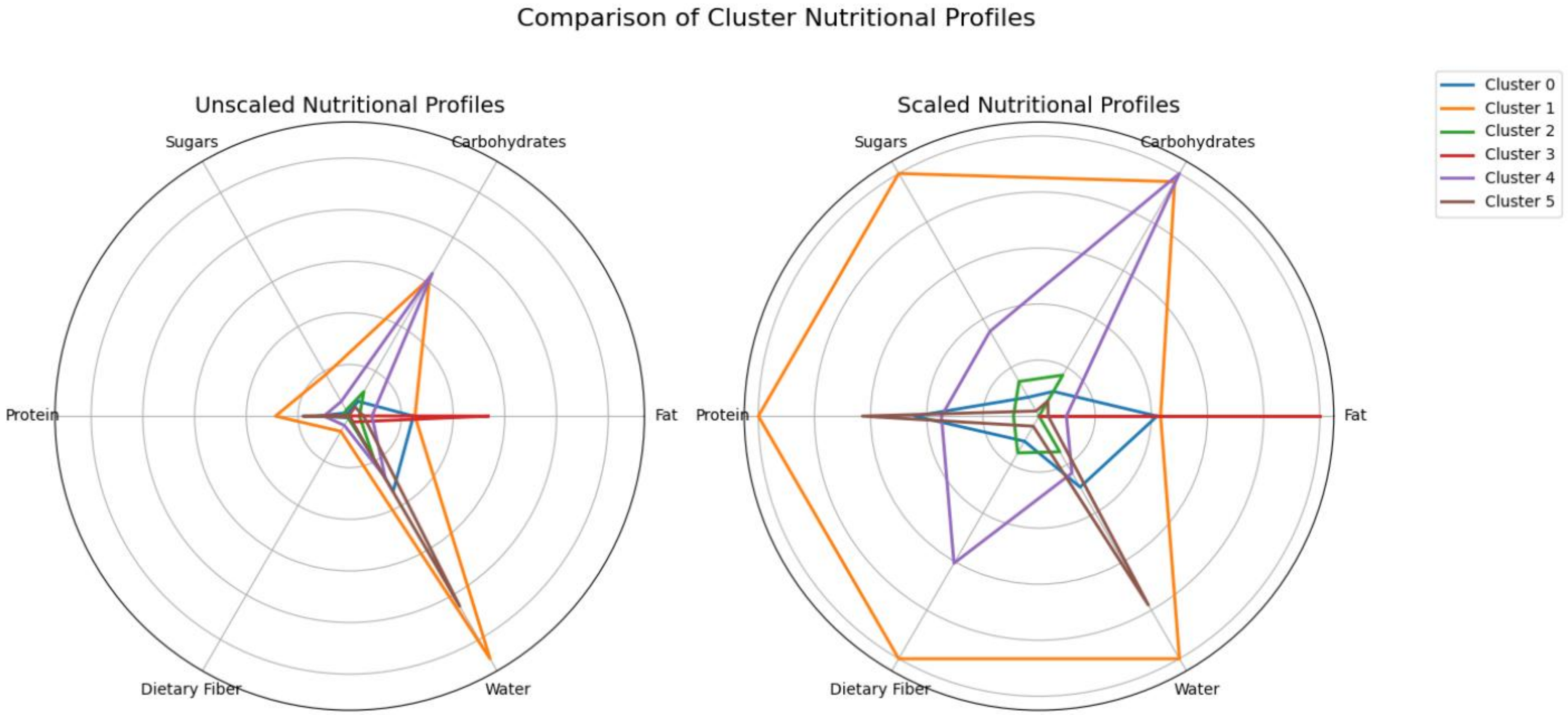
Scenario	Gamma	k	Reasoning
Balanced Choice	1	5 or 6	Metrics are strong, and the number of clusters is interpretable
Highest Scores	10.0	7 or 8	Scores are at their peak, but interpretability and cluster meaning are risky

Kernel PCA + KMeans



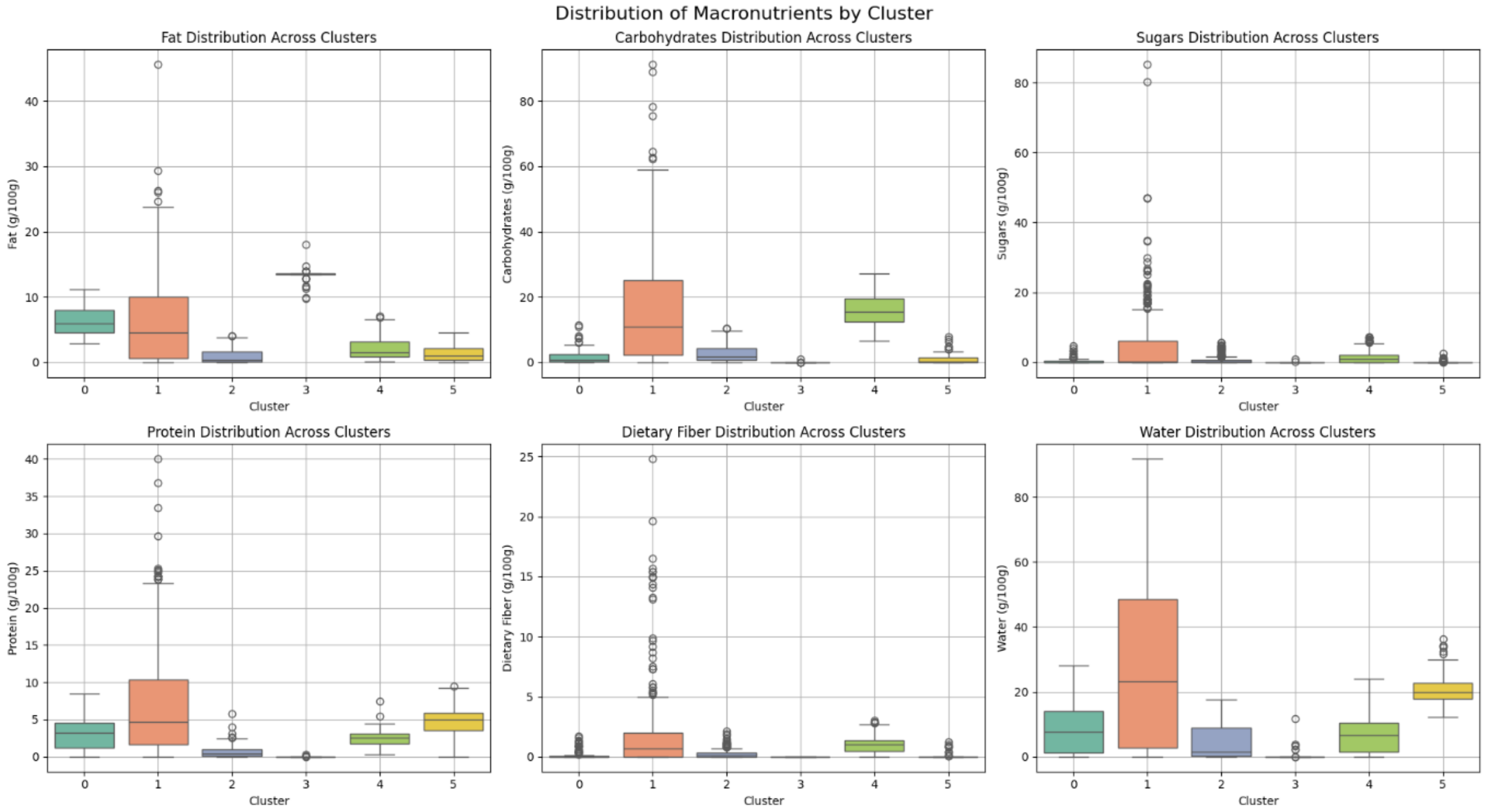
- Visualizes cluster separation after non-linear dimensionality reduction (via RBF kernel)
- Captures complex, curved relationships that linear PCA cannot reveal
- Despite visual separation, **PC1 and PC2 are abstract** and **not directly linked to original features**
- Not informative for feature-level interpretation – use radar/box plots instead

Kernel PCA + KMeans



- **Cluster 0** → Balanced nutrient profile with moderate protein, fat and fiber; likely general-purpose foods.
- **Cluster 1** → Extremely high in all nutrients; likely dense, multi-component meals or processed items.
- **Cluster 2** → Low across all nutrients; likely minimal or diet/light food products.
- **Cluster 3** → Dominated by fat; likely includes butter, oils, or fatty products.
- **Cluster 4** → High in carbohydrates; likely sweet snacks or desserts.
- **Cluster 5** → High in water and protein; likely dairy-based or hydrating protein foods.

Kernel PCA + KMeans



The box plots highlight key differences in macronutrient distributions across clusters. Cluster 1 shows the widest range and highest values overall, while Cluster 3 appears depleted. Clusters 4 and 5 point to fiber-rich and protein-hydrating profiles respectively.

Conclusion

- PCA + KMeans provided initial clustering, but with moderate separation
- Kernel PCA (RBF) + KMeans improved cluster structure significantly
- Silhouette score, DB index, and CH score supported the improvement
- Radar and box plots enabled detailed nutrient-level interpretation
- Final clusters captured meaningful nutritional profiles (e.g. high-fat, fiber-rich, water-dominant, sugar-heavy)
- Kernel-based transformation proved essential to uncover non-linear patterns in food data
- These insights can support dietary planning, recommendation engines, and nutrition-focused applications.

Thanks!