International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 11 | Nov - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Food requirement Analysis Using Machine Learning

Rajashekhar G C [2] ·Huchhiresh M Chapparad[1]

[2]*Assistant Professor& Director , FCIT , GM University, Davanagere*
[1] *Student, Department of MCA, GM University, Davangere*

E-mail : veereshchapparad@gmail.com

Abstract—Understanding the diverse food requirements within a population is critical for public health policy, personalized nutrition, and efficient food supply chain management. Traditional methods often rely on broad demographic averages, failing to capture the nuanced dietary patterns of smaller subgroups. This paper proposes a data-driven framework using unsupervised machine learning to analyze and segment food requirements. We leverage clustering algorithms, specifically K-Means and Agglomerative Hierarchical Clustering, to identify distinct dietary profiles from a dataset comprising demographic, anthropometric, and lifestyle features. The dataset is pre-processed, and features are scaled to ensure algorithmic efficacy. The optimal number of clusters is determined using the Elbow method and Silhouette Score analysis. The resulting clusters reveal distinct, interpretable groups such as "Active Young Adults," "Sedentary Middle-Aged Individuals," and "At-Risk Seniors," each with unique nutritional needs. The findings demonstrate that clustering is a powerful tool for uncovering hidden patterns in food consumption, enabling targeted nutritional interventions and more effective resource allocation. This approach provides a scalable and granular alternative to conventional population analysis.

Key words : Clustering, K-Means, Hierarchical Clustering, Food Requirement Analysis, Personalized Nutrition, Data Mining, Unsupervised Learning.

I. INTRODUCTION

The global population's dietary needs are becoming increasingly complex and heterogeneous. Factors such as age, gender, physical activity level, metabolic rate, and underlying health conditions contribute to a wide spectrum of nutritional requirements. Traditional approaches to food requirement analysis, such as those based on Recommended Dietary Allowances (RDAs), provide essential guidelines but often treat the population as a few large, homogeneous groups [1]. This generalization can lead to ineffective public health campaigns, suboptimal nutritional support for specific communities, and inefficiencies in the food supply chain. With the advent of big data and advanced computational techniques, there is a significant opportunity to move beyond these broad-stroke analyses. Machine learning, particularly unsupervised learning, offers powerful tools for discovering inherent structures within data without pre-defined labels [2]. Clustering algorithms, a cornerstone of unsupervised learning, are adept at partitioning data points into groups (or clusters) such that individuals within the same group are more similar to each other than to those in other groups.

This paper explores the application of clustering algorithms for food requirement analysis. The primary objective is to segment a population into distinct groups based on their characteristics and, consequently, their likely dietary needs. By identifying these data-driven segments, we can derive actionable insights for various stakeholders:

Public Health Officials: Can design targeted intervention programs for at-risk groups (e.g., those prone to obesity or malnutrition).
Dietitians and Nutritionists: Can develop more precise personalized meal plans.
Food Retailers and Manufacturers: Can optimize inventory, marketing, and new product development to cater to specific consumer segments.

In this study, we propose a framework that utilizes K-Means and Hierarchical Clustering algorithms. We first preprocess a synthetic dataset containing demographic, anthropometric, and lifestyle features. We then apply the clustering algorithms to partition the data and use internal validation metrics like the Silhouette Score to assess the quality of the clusters. Finally, we analyze the

characteristics of each cluster to define meaningful dietary profiles. The results highlight the viability of this approach in creating a granular, evidence-based understanding of food requirements.

The remainder of this paper is structured as follows: Section II reviews related work in nutrition analysis and machine learning. Section III details the proposed methodology. Section IV describes the experimental setup and the dataset used. Section V presents and discusses the results. Finally, Section VI concludes the paper and suggests directions for future work.

II. RELATED WORK

The application of data mining and machine learning in health and nutrition is a growing field of research. This section reviews prior work in traditional nutritional analysis and the use of clustering in related domains.

A. Traditional Nutritional Analysis

For decades, nutritional science has relied on epidemiological studies and controlled trials to establish dietary guidelines. Organizations like the World Health Organization (WHO) and national health institutes publish RDAs that provide nutrient reference values for broad population groups categorized by age and gender [1]. While foundational, these guidelines do not account for other significant factors like physical activity, body composition, or metabolic health. Methods like 24-hour dietary recalls and food frequency questionnaires are common for data collection, but they are subject to recall bias and can be resource-intensive [2][3].

B. Machine Learning in Health and Nutrition

Machine learning has been successfully applied to various health-related problems, including disease prediction, medical image analysis, and drug discovery [4]. In nutrition, supervised learning models have been used to predict the risk of conditions like diabetes and cardiovascular disease based on dietary and lifestyle inputs [5]. However, supervised methods require labeled data, which is often unavailable for defining novel dietary patterns.

C. Clustering for Pattern Discovery

Unsupervised clustering has been widely used to identify patterns in various domains. In marketing, it is a standard technique for customer segmentation [6]. In bioinformatics, it is used to group genes with similar expression patterns [7].

Several studies have applied clustering to nutritional data. For instance, [8] used K-Means clustering to identify dietary patterns from food frequency questionnaire data and linked these patterns to the risk of chronic diseases. Another study [9] applied hierarchical clustering to segment older adults based on their food intake, revealing distinct groups with varying levels of nutritional risk. However, many of these studies focus solely on food intake data. Our work extends this by creating profiles based on a holistic set of personal attributes (demographics, anthropometrics, lifestyle) that fundamentally drive food requirements. By clustering based on these causal factors, we aim to define more stable and actionable population segments.

III. METHODOLOGY

The proposed framework for analyzing food requirements using clustering is depicted in Fig. 1. It consists of four main stages: Data Preprocessing, Feature Engineering and Scaling, Cluster Model Application, and Cluster Validation and Interpretation.

Fig. 1. Proposed Methodology Flowchart

A. Data Preprocessing

The initial step involves preparing the raw data for analysis. This includes:

Handling Missing Values: Missing data can lead to biased results. We use mean/median imputation for numerical features and mode imputation for categorical features to handle any missing entries.

Encoding Categorical Variables: Machine learning algorithms require numerical input. Categorical features like 'Gender' (Male/Female) and 'Activity Level' (Sedentary/Moderate/Active) are converted into numerical format using one-hot encoding.

B. Feature Engineering and Scaling

To improve model performance, we derive a new feature and scale the existing ones.

Feature Engineering: Body Mass Index (BMI) is a more informative feature for health assessment than height and weight alone. It is calculated as:

Feature Scaling: Clustering algorithms like K-Means are distance-based, making them sensitive to the scale of features. Features with larger ranges can disproportionately influence the clustering process. To mitigate this, we use StandardScaler, which transforms each feature to have a mean of 0 and a standard deviation of 1.

C. Clustering Algorithms

We employ two popular clustering algorithms to identify distinct groups in the data.

K-Means Clustering: K-Means is a partitional clustering algorithm that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean (cluster centroid) [10]. The algorithm iteratively performs two steps:

Assignment Step: Assign each data point to the closest centroid.

Update Step: Recalculate the centroid of each cluster as the mean of all points assigned to it.

The process continues until the centroids no longer move. A key challenge is selecting the optimal number of clusters, k. We use the Elbow method for this purpose.

Agglomerative Hierarchical Clustering: This is a bottom-up approach where each data point starts in its own cluster. At each step, the two closest clusters are merged until only one cluster (or a specified number of clusters) remains [11]. The process can be visualized using a dendrogram, which shows the hierarchical relationship between clusters.

D. Cluster Validation and Interpretation

After forming clusters, their quality must be evaluated.

Elbow Method: This method involves running the K-Means algorithm for a range of k values and calculating the Within-Cluster Sum of Squares (WCSS) for each. A plot of WCSS against k typically shows an "elbow," and the k value at this point is considered the optimal number of clusters.

Silhouette Score: This metric measures how well-separated the clusters are. The score ranges from -1 to 1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters [12]. The Silhouette Score for a sample i is given by:

IV. EXPERIMENTAL SETUP

A. Dataset Description

For this study, a synthetic dataset was generated to simulate a diverse population. The dataset consists of 5,000 records with the following features:

Age: Numerical, in years (18-80).

Gender: Categorical (Male/Female).

Weight: Numerical, in kilograms.

Height: Numerical, in meters.

Physical Activity Level: Categorical (Sedentary, Lightly Active, Moderately Active, Very Active).

Food Consumption (Target for analysis, not clustering): Daily intake in grams for food groups like Grains, Proteins, Vegetables, Fruits, and Fats/Sugars.

The primary clustering was performed on the personal attributes, and the food consumption data was used later to validate and describe the nutritional profile of each cluster.

B. Implementation Details

The experiments were conducted using the Python programming language (version 3.8). The following libraries were instrumental:

Pandas: For data manipulation and preprocessing.

Scikit-learn: For implementing K-Means, Hierarchical Clustering, StandardScaler, and calculating the Silhouette Score.

Matplotlib & Seaborn: For data visualization, including the Elbow plot and cluster scatter plots.

V. RESULTS AND DISCUSSION

A. Determining the Optimal Number of Clusters (k)

The Elbow method was applied to the preprocessed dataset for a range of k from 2 to 10. The resulting plot is shown in Fig. 2. A distinct "elbow" is visible at k=4, suggesting that four is an appropriate number of clusters for this dataset.

To further validate this choice, we calculated the average Silhouette Score for different values of k. The results, also pointing towards k=4 as the optimal choice, confirmed that this number of clusters provides a good balance of cohesion and separation.

B. Cluster Analysis and Interpretation

With k=4, the K-Means algorithm was used to segment the population. The centroids of the resulting clusters were analyzed to understand the characteristics of each group. The summary is presented in TABLE I.

TABLE I

CLUSTER PROFILES BASED ON FEATURE CENTROIDS

Feature	Cluster 0	Cluster 1	Cluster 2	Cluster 3
Age (Years)	25.4	45.1	68.2	38.5
BMI	22.1	26.5	24.8	32.8
Activity Level	Very Active	Moderately Active	Sedentary	Sedentary
Protein Intake (g)	120.5	85.2	65.7	90.1
Carb Intake (g)	350.6	280.4	210.1	260.5
Fat/Sugar Intake (g)	80.1	95.3	75.8	155.4
Cluster Size (%)	28%	35%	22%	15%
Profile Name	Active Young Adults	Balanced Middle-Aged	Seniors	At-Risk Group

Based on this analysis, we assigned a descriptive name to each cluster:

Cluster 0: Active Young Adults: This group consists of younger individuals with a healthy BMI and a very active lifestyle. Their food requirement is characterized by high protein and carbohydrate intake to support their energy expenditure.

Cluster 1: Balanced Middle-Aged: This is the largest segment, comprising middle-aged individuals with a moderate activity level and slightly overweight BMI. Their diet is relatively balanced.

Cluster 2: Seniors: This group includes older adults who are mostly sedentary. Their caloric requirement is lower, reflected in lower intake across most food groups.

Cluster 3: At-Risk Group: This segment, though smaller, is critical from a public health perspective. It includes individuals across a wide age range who are sedentary, have a high BMI (obese category), and show a disproportionately high intake of fats and sugars.

C. Discussion

The results successfully demonstrate that clustering algorithms can uncover meaningful and actionable segments within a population based on food requirement drivers. The identified "At-Risk Group" (Cluster 3) is a prime candidate for targeted public health interventions focusing on diet and exercise. In contrast, the "Active Young Adults" (Cluster 0) could be a target for food manufacturers developing high-protein and energy-dense products.

This granular analysis provides a significant advantage over traditional one-size-fits-all approaches. By understanding the unique profiles of these data-driven segments, policies and commercial strategies can be tailored for maximum impact and efficiency.

VI. CONCLUSION AND FUTURE WORK

This paper presented a framework for food requirement analysis using clustering algorithms. By applying K-Means and Hierarchical Clustering to a dataset of demographic, anthropometric, and lifestyle features, we successfully identified four distinct population segments with unique dietary profiles. The clusters were validated using the Elbow method and Silhouette Score and interpreted to provide actionable insights for personalized nutrition and public health planning.

The study confirms that unsupervised learning is a potent tool for moving beyond broad generalizations and achieving a more nuanced understanding of public health challenges. The data-driven profiles generated by this method can empower stakeholders to design more effective and targeted strategies.

Future work will focus on several key areas:

Enriching the Dataset: Incorporating more diverse data, such as socio-economic status, geographical location, and clinical data (e.g., blood glucose levels), could yield even more refined clusters.

Using Advanced Algorithms: Exploring other clustering algorithms like DBSCAN, which can identify arbitrarily shaped clusters and outliers, or Gaussian Mixture Models (GMM) for probabilistic clustering.

Longitudinal Analysis: Analysing how individuals move between these clusters over time could provide insights into life-stage transitions and the effectiveness of interventions.

Developing a Recommender System: The identified cluster profiles can serve as the foundation for a personalized food recommendation system.

By continuing to refine these data-driven approaches, we can make significant strides towards a future of truly personalized and proactive nutritional care.

REFERENCES

[1] Institute of Medicine (US) Committee to Review Dietary Reference Intakes for Vitamin D and Calcium, Dietary Reference Intakes for Calcium and Vitamin D. Washington (DC): National Academies Press (US), 2011.

[2] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," ACM Computing Surveys, vol. 31, no. 3, pp. 264–323, Sep. 1999.

[3] F. E. Thompson and T. Byers, "Dietary assessment resource manual," Journal of the American Dietetic Association, vol. 94, no. 5, p. 574, 1994.

[4] T. M. Mitchell, Machine Learning. New York: McGraw-Hill, 1997.

[5] S. De-la-Hoz-Correa, et al., "A Machine Learning-Based Approach for the Prediction of the Risk of Developing Type 2 Diabetes," Journal of Clinical Medicine, vol. 8, no. 9, p. 1439, Sep. 2019.

[6] M. J. A. Berry and G. S. Linoff, Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management, 3rd ed. Indianapolis, IN: Wiley Publishing, Inc., 2011.

[7] G. C. Tseng and W. H. Wong, "Tight clustering: a resampling-based approach for identifying stable and tight patterns in data," Biometrics, vol. 61, no. 1, pp. 10-16, Mar. 2005.

[8] C. L. Newby, et al., "Dietary patterns and changes in body mass index and waist circumference in adults," American Journal of Clinical Nutrition, vol. 77, no. 6, pp. 1417–1425, Jun. 2003.

[9] C. Wham, et al., "Dietary patterns of older people in a nationally representative sample," Nutrients, vol. 6, no. 4, pp. 1361–1373, Apr. 2014.

[10] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, pp. 281–297, 1967.

[11] S. C. Johnson, "Hierarchical clustering schemes," Psychometrika, vol. 32, no. 3, pp. 241–254, Sep. 1967.

[12] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," Journal of Computational and Applied Mathematics, vol. 20, pp. 53–65, Nov. 1987.