Clustering-Based Analysis of Nutritional Risk Levels in Popular Fast Food Items

Kurt Justine Almadrones
College of Computing and Information Technologies
National University
Manila, Philippines
almadroneskm@students.national-u.edu.ph

Lance Kenneth Dela Paz
College of Computing and Information Technologies
National University
Manila, Philippines
delapazlf@students.national-u.edu.ph

Abstract—**Fast food consumption is strongly associated with increased risks of obesity, cardiovascular disease, and metabolic disorders. Despite the availability of nutrition labels, consumers often lack objective tools for comparing menu items based on overall nutritional risk. This study applies unsupervised machine learning techniques to analyze the nutritional profiles of 516 fast food items collected from major international chains. After preprocessing, including feature selection, logarithmic transformation, and standardization, Principal Component Analysis (PCA) was used to reduce dimensionality while preserving approximately 89.5% of total variance. Three clustering algorithms—K-Means, Agglomerative Hierarchical Clustering, and DBSCAN—were evaluated using internal validation metrics. The results show that the K-Means algorithm, with three clusters, best balanced how closely related items were within each cluster and how different the clusters were from each other. This method identified three distinct groups of nutritional risk: low, moderate, and high. In contrast, DBSCAN was particularly good at finding extreme nutritional outliers, especially foods high in sodium and fat. These findings suggest that unsupervised learning can provide an objective, data-driven way to categorize nutritional risk. This supports dietary assessments, health education, and public health programs.**

Index Terms—**Fast food nutrition, Unsupervised learning, K-Means clustering, Agglomerative clustering, DBSCAN, Principal Component Analysis (PCA), Nutritional risk classification, Food computing, Public health analytics.**

I. Introduction

Fast food has become a major part of modern diets. The fast pace of life in cities makes fast food a common choice. Poor choices about fast food often lead to eating too many calories, sugar, salt, and fat. This increases the risk of obesity, diabetes, and heart disease. Data science applications and datasets have been created to estimate nutritional values from images of food and related data, showing how data science can be used to make better nutritional decisions [1].

The problem that this study aims to solve is the lack of available data-driven tools that allow consumers and researchers to compare fast food items based on nutritional content in a straightforward and objective way. While traditional nutritional information is available, it is often neglected. This study will use an unsupervised clustering method to categorize fast food items based on their nutritional content. Clustering allows researchers to identify patterns in large datasets without the need for prior knowledge of the data, making it an excellent tool for data exploration and pattern identification [2].

This problem is important because the consumption of fast food affects students, employees, and urban dwellers around the world. Both under nutrition and over nutrition are health hazards, and better data-driven tools can help solve long-term public health problems. The target audience for this proposed solution includes nutrition educators, diet planning services, health apps, and researchers interested in objective information about food nutritional content.

II. Literature Review

The study of nutritional patterns has evolved from simple observations of eating habits into a complex data science challenge known as food computing. For decades, health researchers have warned about the risks of modern diets. Paeratakul et al. (2003) provided early evidence that fast food consumers suffer from significantly higher intakes of fat and sodium while lacking essential vitamins [3]. This clinical concern was validated by the CARDIA study, where Pereira et al. (2005) followed subjects for 15 years and found a direct link between frequent fast food consumption and the development of insulin resistance and weight gain [4].

As processed food options became more common, Jaworowska et al. (2013) identified a major problem in the takeaway sector: the lack of clear nutritional labeling, which makes it difficult for consumers to track their intake [5]. Even with increased awareness, government data from Vikraman et al. (2015) confirmed that fast food remains a primary source of daily calories for children and adolescents in the United States [6].

To address these health challenges, researchers began moving toward automated analysis. Yin et al. (2022) explained that clustering algorithms have advanced from basic distance-based models to sophisticated tools that can process high-dimensional data [2]. These unsupervised learning methods, such as k-means and Gaussian Mixture Models (GMM), allow computers to find hidden structures in massive datasets without needing human labels. This technical progress led to the birth of food computing, which Min et al. (2019) described as an interdisciplinary field that uses machine learning to solve nutrition and food-related problems [7].

Recent work has focused on creating better datasets and more accurate estimation models. Thames et al. (2021) developed the Nutrition5k dataset, which paired thousands of top-down food images with their actual nutritional values [1]. This was a major step forward because it allowed supervised models to learn how to estimate calories directly from pictures. Following this, Qi et al. (2021) created a method that combined visual features with ingredient data to improve prediction accuracy [8]. While these studies were successful, they relied heavily on supervised learning, meaning they required huge amounts of pre-labeled data to work correctly. Previous researchers have tried to use clustering to organize food data, but their focus was usually on how food looks rather than what it contains. Pan et al. (2023) built a hierarchical model that used iterative clustering to help computers recognize different food categories more accurately [9]. While the model was good at visual classification, it did not look at nutritional attributes like sugar or sodium. In a different approach, Sucharitha and Lee (2023) used the GMM algorithm to study how people access food and what they demand in different areas [10]. Although this provided great insights into consumer behavior, it did not categorize the food items based on their actual health profiles.

There is a clear gap in the current research. Most existing studies either use supervised learning to guess calories or use clustering to identify the name of a dish. There is very little research that uses unsupervised learning to group food items by their multi-dimensional nutritional density. TThis study addresses that gap by applying clustering algorithms to identify distinct nutritional risk levels among popular fast food items. Instead of concentrating on recognizing or predicting a single nutrient, the study takes into account the overall effects of protein, fiber, calories, fat, sodium, and sugar. This study offers a machine learning approach for classifying fast food items into significant risk categories by identifying natural groupings within the data.

III. Methodology

This study investigates nutritional profiles to segment popular fast food items into clusters with distinguishing health implications. Unsupervised machine learning techniques, namely K-Means clustering, Agglomerative clustering, and Principal Component Analysis (PCA), were employed to identify unique nutritional patterns among menu items. Before modeling, the data underwent a rigorous feature engineering pipeline to preprocess the nutritional values for clustering analysis. Each process is outlined and expounded as follows

A. Data Collection

The dataset was acquired from Kaggle, an online data science platform, and was compiled by Ulrik Thyge Pedersen in 2023. The data provides a comprehensive breakdown of the nutritional content of various fast food products from popular international chains. The dataset contains 516 rows and 17 selected features representing menu items from fast food chains such as McDonald's, Burger King, Wendy's, KFC, Taco Bell, and Subway. Each row represents one distinct food item described through individual nutritional contents, including caloric content, lipid profiles (total fat, saturated fat, trans fat), cholesterol, sodium, carbohydrates, dietary fiber, sugars, and protein. Additional micronutrient data, including Vitamin A, Vitamin C, and Calcium, are also present.

B. Data Pre-processing

The nutritional profiles of fast food items were analyzed using a number of data cleaning, imputation, and normalizing techniques to guarantee data consistency, statistical validity, and computing efficiency. Converting unstructured nutritional information into a format appropriate for clustering analysis and model training was the primary objective.

Initial Sanity Check

An initial sanity check was conducted to assess data quality and structure prior to preprocessing. The analysis revealed that the dataset consisted of 516 rows and 17 columns. Several nutritional attributes contained null entries. Specifically, vitamin A, vitamin C, and calcium each contained approximately 210 missing values, validating the decision to drop these columns from further analysis. Additionally, dietary fibre contained 12 missing values, while protein had a single missing entry. Beyond missing values, two duplicate rows were also identified within the dataset. Based on these findings, preprocessing was implemented, and the vitamin and mineral columns with substantial missing data were dropped to avoid introducing bias, while the minimally incomplete features, fibre and protein, were addressed through imputation to preserve sample size and data integrity. Finally, the two duplicate rows were removed to prevent any disproportionate influence on downstream analyses and model training.

Exploratory Data Analysis

After preparing the data, a thorough exploratory data analysis was carried out to find patterns, distributions, and connections in the fast food nutritional information. The distribution of values for numerical features and the connections between them were visualized using histograms and correlation matrices.

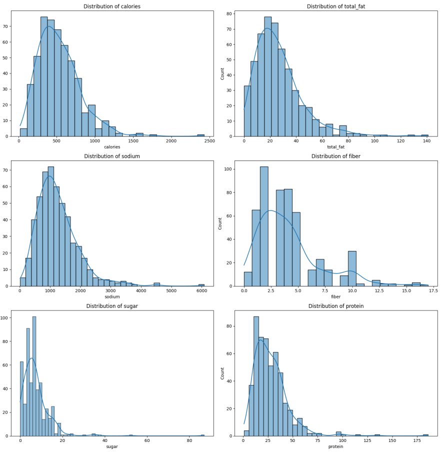

Fig. 1. Distribution of Key Nutritional Components

Histograms were plotted for the major statistical nutritional attributes like calories, total fat, sugar, protein as well as sodium (Fig. 1). The results suggest that these quantitative features are mostly right-skewed. Although most of the fast food items fall into a certain category, there are only a few having very high fat and sodium values. These results suggest that outliers exist in this data set and scaling or normalizing the data prior to clustering seems to be necessary.

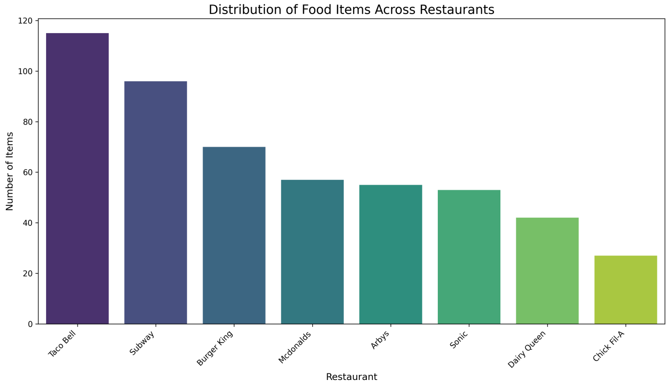

Fig. 2. Distribution of Key Nutritional Components

The prevalence of food items between different fast food chains were visualized using a count plot (Fig. 2). The data consists of 516 observations and a variety of void international restaurant. The chart suggests that chains like Taco Bell and Subway have more of these entries, while others such as Chick-fil-A are represented in smaller samples. This distribution was inspected to assure that the resulting clusters would be representative of the wider fastfood industry and not skewed towards a menu of one dominant chain.

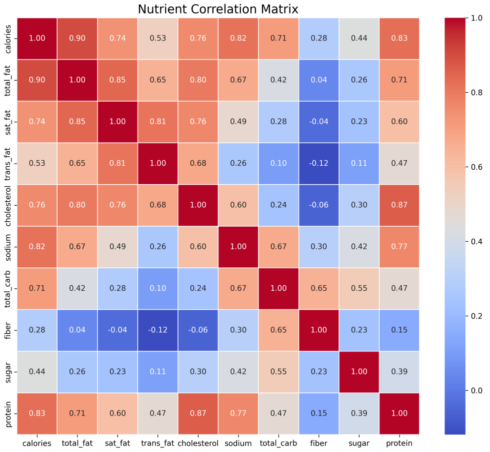

Fig. 3. Correlation Matrix

Nutrient Correlation Matrix was also produced to assess the interrelationships among various nutritional properties (Fig. 3). From the heatmap, there are highly positive relationships between caloric contents with lipid components such as total fat and saturated fat. Large correlations were also evident between sodium and other macronutrients. This will help identify features that are redundant and look at which feature provides most of the independent information to the clustering.

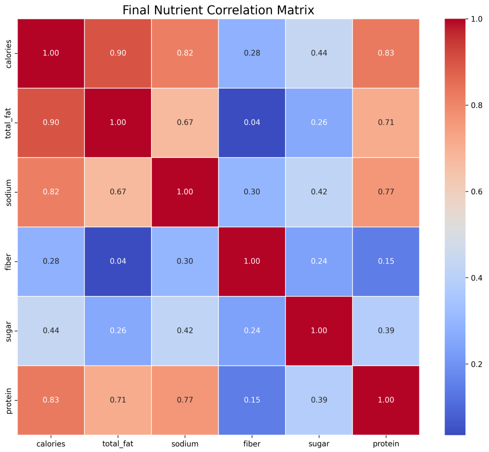

Fig. 4. Correlation Matrix Removed *sat_far, trans_fat, cholesterol, total_carb*

A number of nutritional variables were found to be extremely repetitive based on the correlation analysis. Since trans and saturated fats make up total fat, they showed a substantial correlation with lipid-related metrics. Cholesterol also showed a strong association with caloric content and total fat, suggesting that the nutritional data overlapped. Additionally, there was a moderate to significant correlation between total carbs and sugar, indicating that the two measures represent comparable effects associated to carbohydrates. Saturated fat, trans fat, cholesterol, and total carbs were eliminated from the final feature set in order to lessen multicollinearity and stop redundant characteristics from disproportionately affecting distance-based clustering. This improved interpretability while maintaining the main dietary characteristics for analysis by guaranteeing that only nutritionally unique and representative variables were kept.

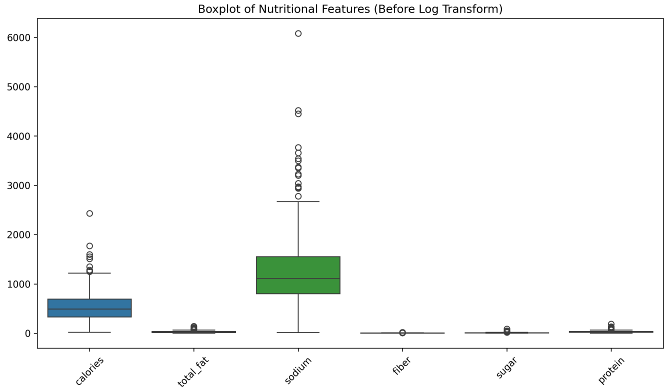

Fig. 5. Boxplot of Nutritional Features (Before Log Transform)

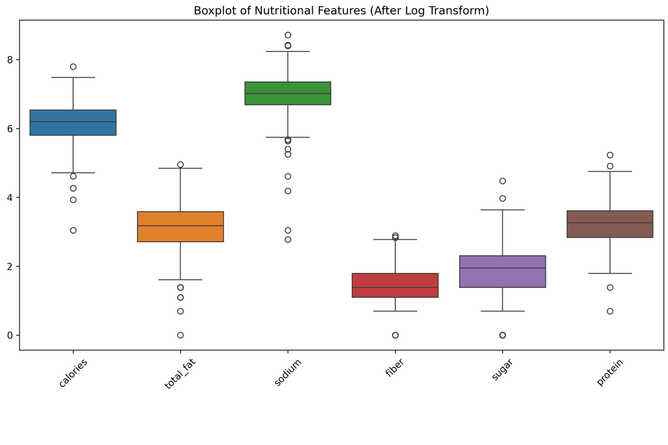

Fig. 6. Boxplot of Nutritional Features (After Log Transform)

Prior to modification, a number of nutritional indicators, especially calories and sodium, showed extreme outliers and an evident right-skew. The distributions become less dominated by extreme values and more balanced after a logarithmic transformation. This modification guaranteed that clustering was not excessively impacted by large-scale variables and enhanced comparability across characteristics.

Feature Selection and Normalization

Correlation analysis revealed that there were significant relationships between a number of nutritional factors. There was a moderate association between total carbs and sugar (0.55), but a strong link between total fat and saturated fat (0.85) and cholesterol (0.80). Only the most nutritionally unique elements were kept in order to cut down on redundancy and enhance interpretability. Calories, total fat, protein, sugar, salt, and fiber were all included in the final feature set. The main nutritional elements of every fast food item are captured by these variables. Given the right-skewed distributions observed for sugar, sodium, and total fat during exploratory analysis, a logarithmic transformation was applied to reduce skewness and stabilize variance. This transformation minimizes the influence of extreme values, ensuring that outlier items do not disproportionately affect the clustering results.

Dimensionality Reduction

Principal Component Analysis (PCA) was applied to simplify the nutritional dataset while preserving the most significant underlying information. After standardizing the selected features, the results indicated that the first two principal components explained approximately 79.3% of the total variance, while the first three principal components accounted for about 89.5%. This demonstrates that a substantial proportion of the variability in nutritional values among fast food items was retained despite reducing the dimensionality of the dataset.

Calories, total fat, sodium, and protein all are primarly captured to the first principal component, which mainly represents overall density. Items are further differentiated by the second primary component according to differences in the balance of macronutrients. Finer distinctions between menu items are made possible by the third principle component, which captures additional structural variation not fully represented by the first two components. The methodology reduces multicollinearity and redundancy while maintaining about 90% of the original data by keeping three principal components. In addition to improving clustering efficiency. This dimensionality reduction enhances clustering performance and ensures that meaningful nutritional patterns remain well represented in the transformed feature space.

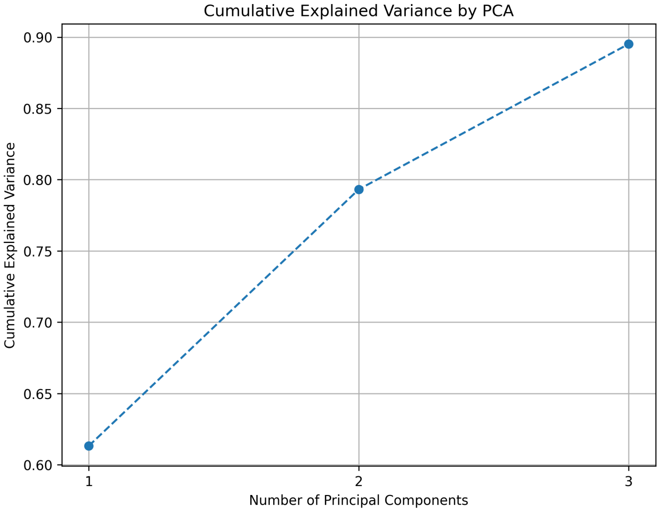

Fig. 7. Cumulative Explained Variance Ratio

C. Experimental Setup

Google Colab was utilized as the main development environment for the entirety of this study. The implementation primarily used scikit-learn (v1.6.1) for clustering algorithms, feature selection techniques, and model evaluation metrics. Data preprocessing and transformation tasks were carried out using NumPy (v1.26.4) and Pandas (v2.2.2). For data visualization, Matplotlib (v3.10.0) and Seaborn (v0.13.2) were employed. Additionally, SciPy (v1.13.1) was used to perform statistical computations for the algorithms that were conducted

Visual Studio Code was used as an alternative development environment for backup. To ensure consistency with the Colab setup, a separate Python environment was configured locally with the same primary library versions installed.

D. Algorithms

In this study, several unsupervised machine learning algorithms were applied to group fast food items according to their nutritional risk profiles. These algorithms were designed to identify natural patterns and similarities within the data and organize the food items into meaningful clusters.

The models used in the analysis were K-Means Clustering, Agglomerative Hierarchical Clustering, and DBSCAN (Density-Based Spatial Clustering of Applications with Noise).

Each algorithm was implemented using the same standardized dataset to ensure a fair and consistent comparison.

K-Means Clustering

K-Means partitions data into k distinct clusters by assigning each item to the group with the nearest centroid [11]. The algorithm updates these centers through an iterative process to minimize the variance within each cluster. We selected this model for its computational efficiency and its ability to find clear groups with similar calorie or sodium levels. A known limitation is the requirement to pre-define the number of clusters.

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \|x_i - v_j\|^2$$

Where:

- $\|x_i - v_j\|$ is the Euclidean distance between a data point and the center.
- n is the total number of data points.
- k is the total number of clusters.

Agglomerative Clustering

Agglomerative clustering is a hierarchical method that works from the bottom up [12]. Every food item starts as its own cluster, and the most similar pairs are merged until a full hierarchy is created. This approach is useful for this study as it reveals how different menu items relate at various levels of detail. The distance between clusters is defined by linkage criteria.

$$d(X, Y) = \min\{d(x, y) : x \in X, y \in Y\}$$

Single Linkage: Measures the distance between the two closest points in separate clusters.

$$d(X, Y) = \max\{d(x, y) : x \in X, y \in Y\}$$

Complete Linkage: Measures the distance between the two furthest points in separate clusters.

$$d(X, Y) = \frac{1}{|X| \cdot |Y|} \sum_{x \in X} \sum_{y \in Y} d(x, y)$$

Average Linkage: Calculates the average distance between all pairs of points across two clusters.

DBSCAN Clustering

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) identifies clusters based on the density of data points in a given area [13]. Unlike distance-based models, it does not require a fixed number of clusters and can find groups with irregular shapes. A primary advantage for this research is its ability to identify outliers as noise, preventing extreme items like high-sodium meals from distorting the centers of standard food groups.

$$N_\epsilon(p) = \{q \in D \mid dist(p, q) \leq \epsilon\}$$

Where:

- ϵ is the maximum distance for defining a neighborhood.
- $N_\epsilon(p)$ is the neighborhood around a point.
- $minPts$ is the minimum number of points needed to form a dense region.

E. Training Procedure

The dataset underwent feature selection followed by z-score standardization to ensure uniform contribution of all nutritional attributes. Subsequently, Principal Component Analysis (PCA) was applied to reduce dimensionality while preserving at least 90% of the total variance. The resulting principal components were used as input for all clustering models to improve computational efficiency and reduce feature redundancy.

The Elbow Method and Silhouette Analysis were used to identify the ideal number of clusters (k) for K-Means clustering. The PCA-transformed feature space was used for clustering.

For Agglomerative Hierarchical Clustering, Ward's linkage criterion was employed to minimize within-cluster variance. The number of clusters was selected through dendrogram analysis based on the PCA components.

Lastly, DBSCAN clustering was conducted on the PCA-reduced dataset. The epsilon (ε) parameter was determined using the k-distance graph, while the *min_samples* parameter was tuned according to dataset density characteristics.

F. Evaluation Metrics

The Silhouette Score was used to measure how well each data point fits within its assigned cluster relative to neighboring clusters. The score ranges from -1 to 1, with values closer to 1 indicating well-separated and compact clusters. Values near 0 suggest overlapping clusters, while negative values point to possible misclassification of data points.

Davies–Bouldin Index (DBI) was utilized to evaluate the average similarity between clusters based on intra-cluster dispersion and inter-cluster distance. Lower DBI values indicate better clustering performance, reflecting compact and well-separated clusters.

Calinski–Harabasz Index (CHI) was employed to measure the ratio of between-cluster variance to within-cluster variance. Higher CHI values suggest more distinct and well-defined cluster structures.

G. Comparison of Clustering Algorithms

For this project, we used three different unsupervised learning models to group fast food items into nutritional risk categories: K-Means Clustering, Agglomerative Hierarchical Clustering, and DBSCAN. Each of these models went through a tuning process to make sure the resulting clusters accurately represented the differences in calories, fat, and sodium across the dataset.

K-Means Clustering served as our baseline model. It was chosen because it is very efficient and works well when visualized through Principal Component Analysis (PCA). To find the best number of clusters, we applied the Elbow Method

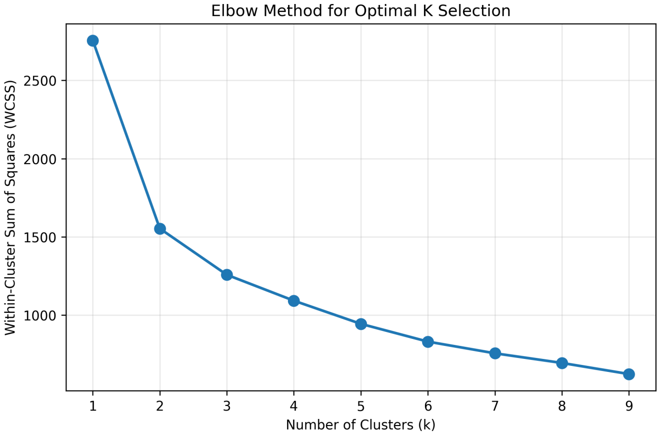

Fig. 8. Elbow Method for K Value

and Silhouette Analysis. This helped us confirm that the groups were distinct and not overlapping too much.

Agglomerative Clustering, which is a hierarchical approach, was used to see how the food items relate to each other in a nested structure. We used Ward's linkage to keep the clusters compact and minimize the variance within each group. The final number of clusters was determined by looking at the major splits in the dendrogram, which showed how different food categories naturally branched away from each other.

DBSCAN was also included because it is a density-based model. Unlike the other two, DBSCAN does not force every single item into a cluster. This was useful for identifying "noise," which in our case represented nutritional outliers like items with extremely high salt or fat content. We tuned the epsilon and min samples parameters using a k-distance graph to match the density of our specific food dataset.

When we compared the models, K-Means and Agglomerative Clustering gave us the most structured groups for general classification. However, DBSCAN was better at filtering out extreme values that might have skewed the averages in the other models. Ultimately, K-Means was the most effective for the final analysis.

IV. Results and Discussion

This section presents the findings from the unsupervised machine learning analysis of 516 fast food menu items. Following a rigorous preprocessing pipeline that included log transformations to stabilize skewed nutrient distributions , the study evaluated three clustering models: K-Means, Agglomerative Hierarchical Clustering, and DBSCAN

A. Elbow Method Plots for finding the Optimal K Value

To determine the optimal number of clusters for K-Means clustering, the Elbow Method was applied by computing the Within-Cluster Sum of Squares (WCSS) for k values ranging from 1 to 9. As shown in Fig. 6, despite the noticeable gap between k = 1 and k = 2, k = 3 was chosen to maintain significant nutritional stratification into groups at low, moderate, and high risk. This is in line with the study's goal of classifying risk levels as opposed to separating them into binary categories.

B. K-Means Clustering

The K-Means algorithm served as the baseline model for this study, utilizing an optimized cluster count of $k = 3$ derived from the Elbow Method. This model focuses on minimizing the variance within each cluster to create distinct groupings based on nutritional density.

TABLE I
K-Means Performance Metrics

Metric	Value
Number of Clusters (k)	3
Silhouette Score	0.2869
Davies-Bouldin Index (DBI)	1.1646
Calinski-Harabasz Index (CHI)	303.54

The K-Means model demonstrated the best overall performance among the three algorithms tested. With a Silhouette Score of 0.2869, it achieved the highest level of cluster cohesion and separation, suggesting that the centroid-based structure is well-suited for the log-transformed nutritional data. Furthermore, the model yielded the highest Calinski-Harabasz Index (303.54), reinforcing that the clusters are well-defined and distinct from one another. The selection of $k = 3$ successfully stratified the menu items into interpretable groups representing low, moderate, and high nutritional risk.

C. Agglomerative Clustering Model Dendrogram

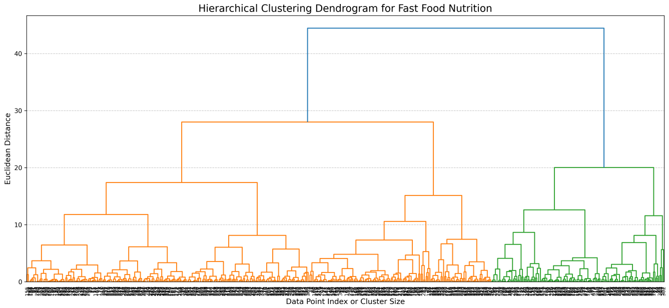

Fig. 9. Dendrogram

The dendrogram in (Fig. 9) presents a clear, two-part story. At the highest level, the wide gap between the two primary vertical branches suggests a straightforward two-cluster split. However, the larger of these branches contains a distinct internal break at a distance of approximately 28. While the initial merge is the most prominent feature, this secondary split is sharp enough to suggest that a two-cluster model would oversimplify the data's actual structure. Choosing a three-cluster configuration captures this internal nuance without adding unnecessary complexity. This more detailed view isn't just a visual preference; it aligns with the results of the

Elbow Method and other validation metrics, confirming that three clusters provide the most balanced representation of the dataset.

D. Agglomerative Hierarchical Clustering

The Agglomerative model was employed to examine the nested relationships between food items, using Ward's linkage criterion to minimize within-cluster variance. This hierarchical approach allows for a bottom-up understanding of how food items group together based on similarity.

TABLE II
AGGLOMERATIVE CLUSTERING PERFORMANCE METRICS

Metric	**Value**
Number of Clusters (k)	3
Silhouette Score	0.2288
Davies-Bouldin Index (DBI)	1.2850
Calinski-Harabasz Index (CHI)	255.52

The Agglomerative model displayed moderate performance but did not outperform the K-Means baseline. The Silhouette Score of 0.2288 indicates reasonably structured clusters, though they are less distinct than those formed by K-Means. This is further reflected in the Davies-Bouldin Index (1.285), which is higher than K-Means, indicating slightly more dispersion within clusters or closer proximity between them. While the hierarchical approach provided insight into the branching nature of food categories, the metrics suggest it is less effective for the primary goal of distinct risk categorization.

E. DBSCAN (Density-Based Spatial Clustering)

DBSCAN was utilized to identify arbitrary shaped clusters and filter out nutritional outliers. Unlike the previous models, DBSCAN does not require a pre-defined number of clusters and identifies ”noise” points that do not fit into dense regions.

TABLE III
DBSCAN PERFORMANCE METRICS

Metric	**Value**
Clusters Found	7
Noise Points (Outliers)	134
Silhouette Score	0.1305
Davies-Bouldin Index (DBI)	0.8479
Calinski-Harabasz Index (CHI)	58.82

The DBSCAN model revealed unique insights into the dataset's density but struggled with general categorization. Notably, it identified 134 noise points, effectively filtering out over 25% of the dataset as nutritional outliers. While DBSCAN achieved the lowest (best) Davies-Bouldin Index of 0.8479, this metric reflects a small number of tightly packed clusters while excluding a significant portion of the data as noise. The low Silhouette Score (0.1305) and CHI (58.82) confirm a weak overall cluster structure for the entire dataset. Consequently, DBSCAN was highly effective at outlier detection—identifying items with extreme sodium or fat—but was less suitable for comprehensive segmentation.

F. Comparison of Models

TABLE IV
PERFORMANCE EVALUATION METRICS FOR CLUSTERING MODELS

Algorithms	**Silhouette Score ↑**	**DBI ↓**	**CHI ↑**
K-Means	**0.2869**	1.1646	**303.54**
Agglomerative	0.2288	1.2850	255.52
DBSCAN	0.1305	**0.8479**	58.82

Note: ↑ indicates higher is better; ↓ indicates lower is better.

Based on the metrics presented in Table IV, K-Means (k=3) demonstrated the strongest overall performance for the fast food dataset. It achieved the highest Silhouette Score (0.2869) and the highest Calinski-Harabasz Index (303.54), indicating better cluster cohesion and separation compared to the other methods. These results suggest that the centroid-based structure of K-Means is well-suited to the general distribution of the log-transformed nutritional data.

Although DBSCAN produced the lowest (best) Davies-Bouldin Index (0.8479), its substantially lower Silhouette Score (0.1305) and CHI (58.82) indicate weak overall cluster structure. The comparatively low DBI may be influenced by DBSCAN identifying a small number of tightly packed clusters while labeling a significant portion of the 516 menu items as noise, which reduces within-cluster dispersion but limits interpretability.

Agglomerative Clustering (k=3) showed moderate performance, with metrics consistently below K-Means across all measures. While it formed reasonably structured clusters, it did not outperform K-Means in cohesion or separation.

Overall, K-Means (k=3) is selected as the primary clustering model, as it provides the most balanced and interpretable segmentation of the nutritional menu data across multiple evaluation metrics.

G. Cluster Profiling and Characteristics

TABLE V
CLUSTER CENTROIDS: NUTRITIONAL PROFILES

Cluster	**Cal. (kcal)**	**Fat (g)**	**Sod. (mg)**	**Fib. (g)**	**Sug. (g)**	**Prot. (g)**
0	398.30	18.74	967.92	3.74	5.47	20.60
1	770.00	40.00	1767.29	5.13	10.59	40.23
2	194.34	8.88	487.24	2.25	2.47	11.91

To interpret the clustering results, the mean nutritional values of each cluster were examined. The profiling analysis revealed three distinct nutritional risk levels among popular fast food items.

1) Cluster 1 - High Nutritional Risk Group

Cluster 1 exhibited the highest average values across nearly all major nutrients, with a mean caloric content of 770 kcal, total fat of 40 g, sodium of 1767 mg, and sugar of 10.59 g. These values significantly exceed those observed in the other clusters, indicating high calorie density and substantial sodium concentration.

The sodium levels in this cluster approach or surpass recommended daily intake limits. The World Health Organization (WHO) recommends limiting sodium intake to less than 2000 mg per day to reduce the risk of hypertension and cardiovascular disease [14] when consumed regularly. These items are likely highly processed and contain high amounts of calories, fat, and sodium. Therefore, Cluster 1 was classified as the High Nutritional Risk group. Nutritional Risk group.

2) Cluster 0 - Moderate Nutritional Risk Group

Cluster 0 showed intermediate nutritional qualities, with an average salt level of 968 mg and a mean calorie content of 398 kcal. These values are noteworthy even though they are far lower than the high-risk cluster, and if they are regularly ingested, they could increase cumulative dietary risk.

The menu items in this cluster are moderately high in energy content and neither extremely nor minimally nutrient-dense. Cluster 0 was therefore assigned the label of Moderate Nutritional Risk.

3) Cluster 2 - Low Nutritional Risk Group

Among all the important nutritional variables, Cluster 2 had the lowest average values: mean calories were 194 kcal, total fat was 8.88 g, salt was 487 mg, and sugar was 2.47 g. These numbers suggest menu items that are relatively lighter and have lower sodium and calorie densities.

Compared to the other groups examined, the products in this cluster provide comparatively safer nutritional options, even if they are still classified as fast food offerings. Cluster 2 was therefore categorized as the group with the Low Nutritional Risk.

The clear differences in average nutrient values show that popular fast food items are not all the same in terms of nutrition. Instead, they naturally form distinct groups based on differences in calories and sodium levels. The clustering approach helps turn raw nutritional data into clear and understandable risk levels, supporting the goal of identifying nutritional risk categories using unsupervised learning. expound more

The three-dimensional PCA visualization shows how fast food items are distributed spatially within the three nutritional risk groups that have been identified. Overall salt and calorie density have a significant impact on cluster formation, as seen by the distinct separation seen mostly along the first principal component. The Moderate Risk cluster is located between the Low Risk and High Risk clusters, which occupy opposite ends of the feature space.

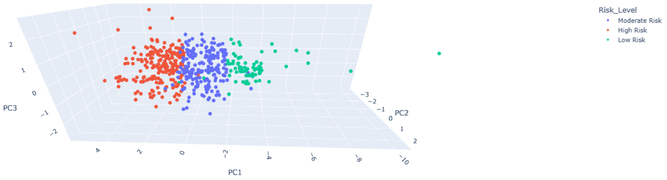

Fig. 10. PCA Visualization of Nutritional Risk Clusters

V. Conclusion

Three clustering algorithms were evaluated: K-Means Clustering, Agglomerative Hierarchical Clustering, and DBSCAN (Density-Based Spatial Clustering of Applications with Noise). Their performance was compared using internal validation metrics such as the Silhouette Score, Davies–Bouldin Index, and Calinski–Harabasz Index. Among these methods, K-Means demonstrated the most balanced and interpretable results and was supported by the Elbow Method in selecting an optimal cluster configuration of k = 3.

Based on the validation metrics and the Elbow Method, which showed an ideal cluster formation at k = 3, K-Means showed the most consistent and comprehensible results among the assessed clustering algorithms. Three different nutritional profiles that corresponded to low, moderate, and high risk levels were identified by the clustering technique. The high-risk group contained items with much higher calories, fat, and sodium levels, while the low-risk group included lighter options with lower nutritional values. The moderate group fell between these two extremes.

The results show that fast food items may be classified into distinct and quantifiable risk groups according to their nutrient composition, even though they are not nutritionally similar. This study offers an objective, data-driven basis for nutritional risk identification that does not rely on predefined labels by utilizing unsupervised learning. Future uses in dietary planning, consumer awareness, and public health research may benefit from this method, which shows the usefulness of clustering algorithms in food dataset analysis.

Despite these findings, one limitation of this study is that the dataset was restricted to selected restaurant chains and did not account for portion size variation or consumption frequency. To further improve nutritional risk modeling, future studies might use more sophisticated clustering algorithms, dietary guideline parameters, portion normalization, or larger diverse dataset

This study concludes that unsupervised machine learning techniques can effectively identify and categorize nutritional risk levels in fast food items. The findings highlight the importance of model selection, preprocessing, and dimensionality reduction in achieving meaningful clustering outcomes.

References

[1] Q. Thames, A. Karpur, W. Norris, F. Xia, L. Panait, T. Weyand, and J. Sim, “Nutrition5k: Towards automatic nutritional understanding

of generic food," 2021. [Online]. Available: https://arxiv.org/abs/2103.03375

[2] H. Yin, A. Aryani, S. Petrie, A. Nambissan, A. Astudillo, and S. Cao, "A rapid review of clustering algorithms," 2024. [Online]. Available: https://arxiv.org/abs/2401.07389

[3] S. Paeratakul, D. P. Ferdinand, C. M. Champagne, D. H. Ryan, and G. A. Bray, "Fast-food consumption among us adults and children: dietary and nutrient intake profile," *Journal of the American dietetic Association*, vol. 103, no. 10, pp. 1332–1338, 2003.

[4] M. A. Pereira, A. I. Kartashov, C. B. Ebbeling, L. Van Horn, M. L. Slattery, D. R. Jacobs, and D. S. Ludwig, "Fast-food habits, weight gain, and insulin resistance (the cardia study): 15-year prospective analysis," *The lancet*, vol. 365, no. 9453, pp. 36–42, 2005.

[5] A. Jaworowska, T. Blackham, I. G. Davies, and L. Stevenson, "Nutritional challenges and health implications of takeaway and fast food," *Nutrition reviews*, vol. 71, no. 5, pp. 310–318, 2013.

[6] S. Vikraman, C. D. Fryar, and C. L. Ogden, "Caloric intake from fast food among children and adolescents in the united states, 2011–2012," 2015.

[7] W. Min, S. Jiang, L. Liu, Y. Rui, and R. Jain, "A survey on food computing," 2019. [Online]. Available: https://arxiv.org/abs/1808.07202

[8] H. Qi, B. Zhu, C.-W. Ngo, J. Chen, and E.-P. Lim, "Advancing food nutrition estimation via visual-ingredient feature fusion," 2025. [Online]. Available: https://arxiv.org/abs/2505.08747

[9] X. Pan, J. He, and F. Zhu, "Muti-stage hierarchical food classification," in *Proceedings of the 8th International Workshop on Multimedia Assisted Dietary Management*, ser. MM '23. ACM, Oct. 2023, p. 79–87. [Online]. Available: http://dx.doi.org/10.1145/3607828.3617798

[10] R. S. Sucharitha and S. Lee, "Application of clustering analysis for investigation of food accessibility," 2019. [Online]. Available: https://arxiv.org/abs/1909.09453

[11] Y. Li and H. Wu, "A clustering method based on k-means algorithm," *Physics Procedia*, vol. 25, pp. 1104–1109, 2012.

[12] D. Müllner, "Modern hierarchical, agglomerative clustering algorithms," *arXiv preprint arXiv:1109.2378*, 2011.

[13] E. Schubert, J. Sander, M. Ester, H. P. Kriegel, and X. Xu, "Dbscan revisited, revisited: Why and how you should (still) use dbscan," *ACM Trans. Database Syst.*, vol. 42, no. 3, Jul. 2017. [Online]. Available: https://doi.org/10.1145/3068335

[14] N. Stern, A. Buch, R. Goldsmith, L. Nitsan, M. Margaliot, R. Endevelt, Y. Marcus, G. Shefer, and I. Grotto, "The role of caloric intake in the association of high salt intake with high blood pressure," Aug 2021. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC8339119/