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*—**
***Index Terms*—**

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. This research fills that gap by applying clustering algorithms to discover natural patterns in nutritional data. By focusing on the relationship between fats, sugars, and sodium rather than visual appearance, this study offers a new way to understand and categorize the healthfulness of modern diets.

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.

C. 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.

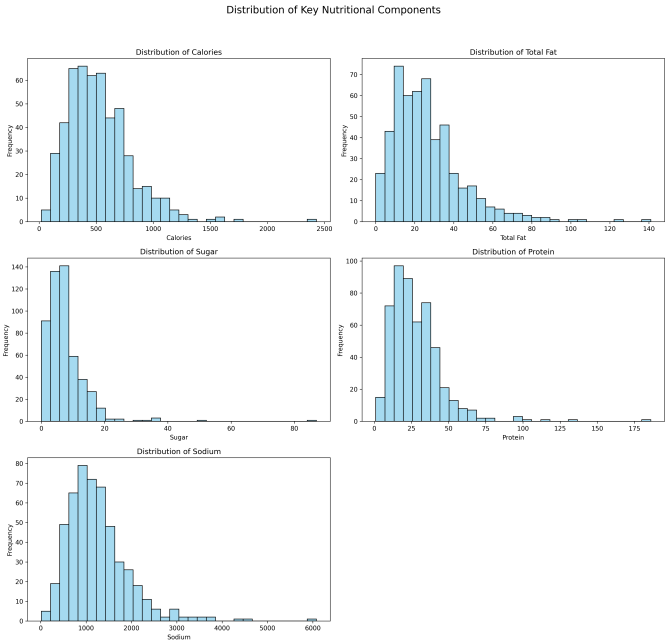

Fig. 1. Distribution of Key Nutritional Components

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.

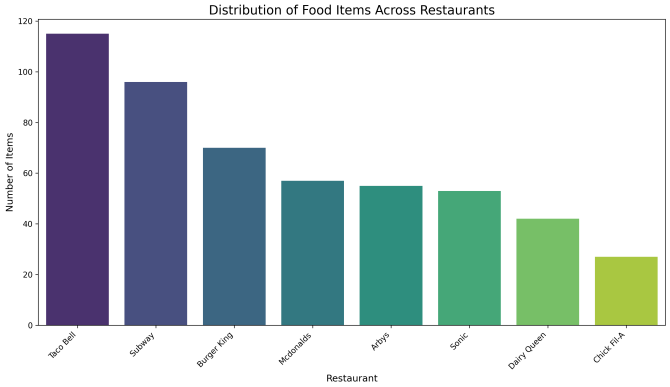

Fig. 2. Distribution of Food Items Across Restaurants

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.

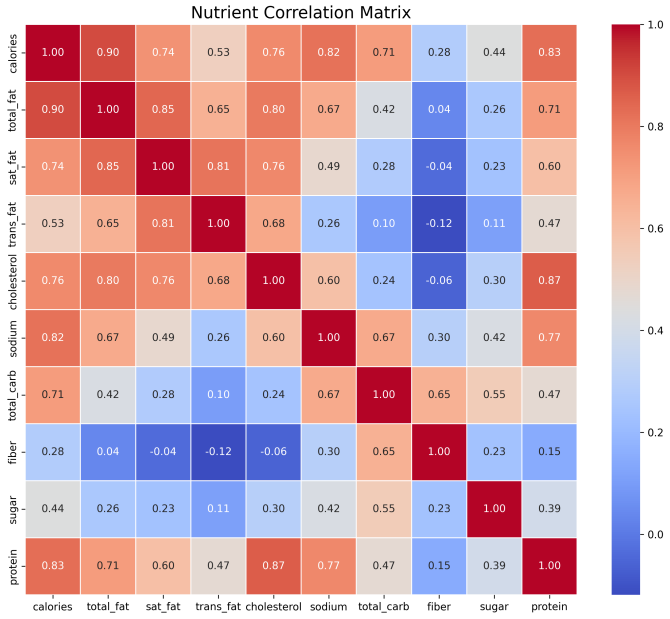

Fig. 3. Nutrient Correlation Matrix

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.

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

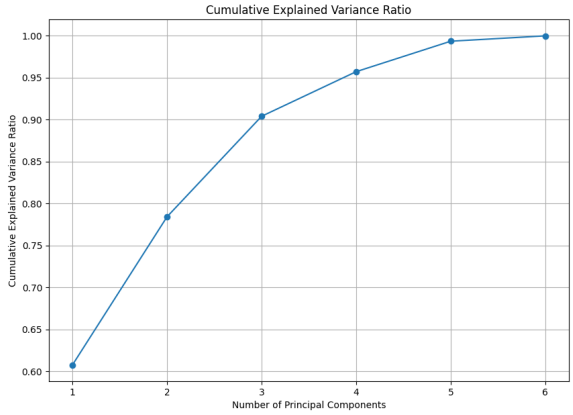

Fig. 4. Nutrient Correlation Matrix

Principal Component Analysis (PCA) was used to simplify the nutritional data while keeping the most important information. After normalizing the features, the results showed that the first two principal components explained about 78.4% of the total variation, while the first three components explained more than 90%. This means that most of the differences in nutritional values among fast food items were still preserved even after reducing the number of variables. The scatter plot of PC1 and PC2 shows that menu items with similar nutritional content such as similar amounts of calories, fat, sodium, and protein appear close to each other. By reducing the data into fewer dimensions, PCA helps remove noise, makes clustering faster and more effective, and still allows the nutritional patterns to be easily understood.

D. 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 alternatve 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.

E. Experimental Setup

F. Algorithm

G. Training Procedure

H. Evaluation Metrics

I. Comparison of Clustering Algorithms

IV. Results and Discussion

A. Limitation and Future Work

V. Conclusion

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