Personalized Health and Nutrition Recommendations Using Machine Learning

Final MSc Research Project
Data Analytics

Sanjay Rajendra Raut
Student ID: 22196901

School of Computing
National College of Ireland

Supervisor: Dr. Abid Yaqoob

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Sanjay Rajendra Raut
Student ID:	22196901
Programme:	Masters in data Analytics
Year:	2024
Module:	Msc In Research Project
Supervisor:	Dr. Abid Yaqoob
Submission Due Date:	29/01/2025
Project Title:	Personalized Health and Nutrition Recommendations Using Machine Learning
Word Count:	
Page Count:	24

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Sanjay Raut ……………………………………………………………………………………………………

Date: 29 /01/2025 ……………………………………………………………………………………………………

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Personalized Health and Nutrition Recommendations Using Machine Learning

Sanjay Rajendra Raut
22196901

Abstract

It is our vision that dependence on individualized nutrition is essential in order to positively impact overall health and reduce chronically debilitating diseases. The present dietary guideline for the population dies not take into account the lifestyle, age, gender, and activity level of the population. In this work, we present an approach to adopting a machine learning model for the prediction of the consumers' dietary behaviours and recommending suitable diets. To achieve this aim, we employed the American Gut Project dataset to apply six Machine Learning models namely; Random Forest, Gradient Boosting, SVM, Neural Networks, CatBoost, and LightGBM. The Random Forest model, after hyperparameter optimization and SMOTE method, was found to be the most accurate classifier with overall accuracy of 73%, precision, recall and F1 score values all similar to each other. Our clustering analysis, performed with K-Means, revealed two distinct dietary patterns. More specifically the pre-determined criteria were 'Healthy'(High protein and fiber content) and 'Unhealthy' (High fat and carbohydrate content). Principal Component Analysis (PCA) was then used to represent these clusters and as seen, the clusters are separable. Some of the measures used in order to classify the diets included the frequency of exercise, the type of diet and age. This research closes the gap between the application of machine learning and personalized nutrition, with a complete program to analyse the intricacies of nutrition. Further work will focus on improving interpretable model solutions and utilizing them for the creation of tools to encourage people to adopt better diets.

1 Introduction

Dietary habits are among the most important factors determining health, yet we cannot adequately provide customized dietary advice. Public health guidelines, on the other hand, while general, do not take into account individual differences in age, sex, or the availability of help through support networks, lifestyle and the like such as smoking or alcohol consumption. This gap is not just in pure and simple science, it matters to people's lives, causing preventable chronic diseases and diminishing their quality of life. This problem cannot be solved with intuition alone, it needs data driven, data based solutions.

Imagine a 45-year-old man working hard at work, family and enjoying life. He doesn't exercise much, smokes occasionally and eats on the run. No matter what his best intentions, his diet doesn't often contain enough of the essential nutrients needed that he's at risk for developing chronic illnesses like diabetes or heart disease. Now imagine the active, health-conscious young woman who looks and feels good, yet cannot be certain about the adequacy of a high protein, low carb diet. They are not isolated scenarios, these are a daily reality that millions around the world face every day.

This research aims to address this challenge by answering key questions:

1. *What are the most effective machine learning techniques for predicting personalized dietary recommendations?*
2. *What are the most important features that create these patterns?*

To achieve these objectives, we employed advanced machine learning techniques to analyze demographic and lifestyle data, identifying two distinct dietary patterns: "Healthy" and "Unhealthy." Differences in nutrient intake like protein, fiber and fat are the characteristics of these patterns. With a Random Forest Classifier, we predicted these patterns from user inputs, achieving 73% accuracy with advanced techniques such as SMOTE and hyperparameter optimization.

Algorithms and numbers are not all we care about in this work; it is also about inspiring the people. I mean, it’s about a tool for people making informed health decisions to be healthier. Regardless if you’re the busy 45 year old or the health conscience 25 year old, this study is one step closer to a future where, personalized nutrition is the new normal (Turkia et al. (2021).

This study makes significant contributions. Second, it tries to show how the field of machine learning can bridge the gap between data analysis and personalized health recommendation. Next, it identifies the most important drivers of dietary patterns, informing nutrition science. It finally provides a versatile framework that builds on to pave better ways ahead in personalized nutrition, enabling personalization of dietary choices based on one’s specific lifestyle.

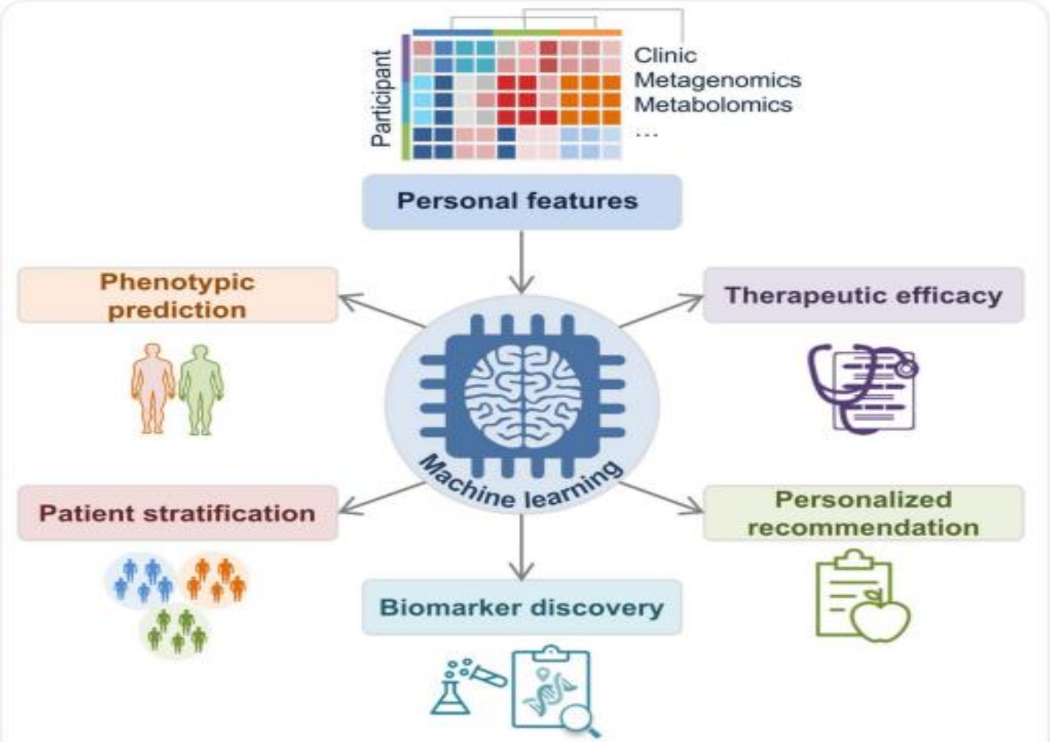

Figure 1: Machine Learning Applications in Precision Medicine.

Adapted from Li et al., Microbial Cell Factories, 2022, 21:241, under a Creative Commons Attribution 4.0 International License

The rest of the report is structured as follows: Section 2 reviews other studies that have contributed to highway design. In section 3, methods including data preparation, clustering, classification are described and the results where the model is evaluated. In Section 4 we discuss what our findings contribute and suggest areas for future research.

This work marks a move forward using technology to enable health transformation. Personally, nutrition is simple, effective, and available to everyone.

2 Related Works

In this section we investigate the academic map of dietary prediction and planning with machine learning (ML). It discusses the strengths and weaknesses of a variety of approaches and identifies important gaps and potential future research areas in this field.

2.1 Machine Learning in Dietary Prediction

Dietary research has benefitted greatly from machine learning, which has successfully uncovered complex patterns and relationships between diet and health. Using mixed effects random forest (Pedersen et al., 2022) they studied food intake using biometric data including eye tracking and electrodermal activity. Their study demonstrated the potential of ML in behavioral nutrition, but the predictive accuracy was limited, since the fine details of the human eating behavior were difficult to capture.

In Roy et al. (2023), OBESEYE, an explainable AI based diet recommender system for preparing personalized dietary plans for obesity and its comorbidities is presented. Despite accuracy in handling specific dietary requirements, the system is based on static datasets from which its scalability to broader demographics and real-time applications is limited. Like Lee et al. (2021) also proposed the Teacher forced REINFORCE algorithm, which uses reinforcement learning to optimize the diet composition. The resulting labeled datasets and computational resources were necessary to achieve significant compliance improvement with nutritional guidelines, and this innovative approach significantly undershoots this.

Dietary research also makes use of natural language processing (NLP). In Hu et al. (2023) pretrained language models like BERT were used to classify food categories and to predict food nutrient quality from food labels. However, their model obtained high accuracy and it was only trained on text based data preventing its use in wider dietary datasets. In review, ML applications in nutrition research were reviewed with a focus on the promise for precision nutrition and disease modeling (Kirk et al., 2022). However, they also pointed out that ML can study high dimensional data, but that it faces the risk of over fitting, and lacks transparency in complex models.

Morgenstern et al. (2021) discuss ML in nutritional epidemiology – addressing confounding factors and measurement errors. They worked to highlight the promise of ML to model dietary patterns and thus predict disease more effectively. This, however, faced challenges on the data standardization front as well as high computational requirements that made practical implementation a difficult task. In a related work, Delegalise et al. (2022) propose a framework for food security prediction using heterogeneous data such as remote sensing data and market trends. This approach was effective for certain regions, but not generally.

Although there have been considerable advancements, ML based dietary prediction remains a challenge. For example, there is a lack of generalizability to other firms. OBESEYE types of systems are developed specifically for certain populations, such as ones dealing with obesity, but they find it hard to adapt to diverse demographic or geographic contexts, and therefore their general utility. Another challenge with complexity of nutritional data also causes overfitting in ML models. Careful preprocessing and feature selection is necessary for high-dimensional datasets (and studies such as Pedersen et al. (2022) and Morgenstern et al. (2021)), because each have tens to hundreds of features. If we do not take these measures, then model's predictive accuracy suffers dramatically and they become useless in the real world.

Issues like data privacy, model transparency and bias in recommendations are just some of the ethical concerns behind the adoption of ML in dietary applications. According to Roy et al. (2023) and studies by Kirk et al. (2022), explainable and equitable AI models that facilitate user trust and ethical integrity in personalized dietary recommendations are required. Real time adaptability is another limitation. However, because of their reliance on static datasets most current systems are not fit for dynamic applications such as wearable health monitoring or real time dietary tracking. In particular, Hu et al. (2023) investigated food label analysis, which has implications but cannot meet practical needs for real time monitoring.

Dietary synergy, interaction between nutrients and the combined effect on health is often overlooked. To address this, Bodnar et al. (2020) used advanced ensemble learning to model nutrient interactions. Such approaches have the potential to reflect the complexities of human nutrition, but they are not widely adopted yet.

2.2 Conclusion

The literature reviewed shows the ways in which machine learning is such a transformative tool in dietary prediction and planning. Some of these issues include the limitation on generalizability, complexity of data, ethical issues, and lack of real time adaptability. In addition, current solutions are also inadequate in acknowledging dietary synergy. These gaps indicate the requirement of innovative and scalable data frameworks in real time integrations with nutrient assessment not only for what is reported, but for the nuanced interplay of nutrients. We aim to address these challenges and push the frontier of dietary prediction.

3 Proposed Research Methodology

In this section, data collection, preprocessing, clustering, classification and evaluation methods are described in the order of steps towards the research objectives. It brings scientific rigor by giving an exact and transparent description of how the procedure is conducted. As seen in the proposed architecture diagram in Figure 2, the stages to consider comprise data acquisition and cleaning, which can be followed by any or a combination of missing value imputation, feature encoding, feature scaling, and feature selection. These processes result in a clustering, classification, assessment, analysis, and production of outputs.

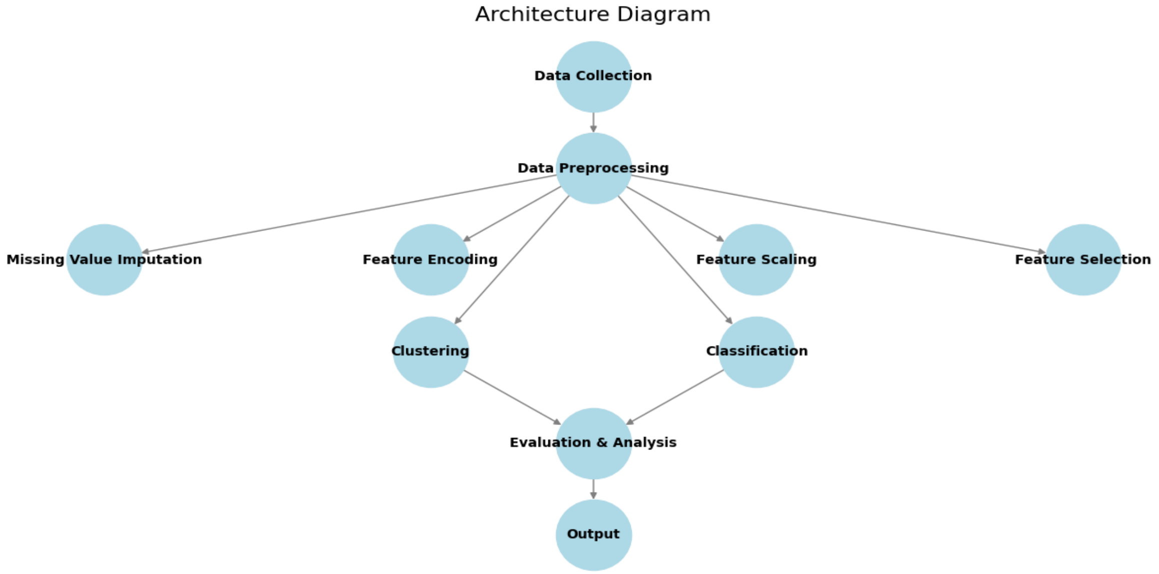

Figure 2 : Architecture Diagram

3.1 Data Collection and Validation

The dataset utilized in this study was sourced from the GitHub repository linked to the research paper “Dietary Patterns and the Gut Microbiome” (DOI: 10.1016/j.ajcn.2022.02.001). The dataset discussed here was originally compiled by American Gut Project for gathering information from diverse demographic, lifestyle, and dietary variables. Using this well curated and known dataset, this study takes advantage of secondary data to analyze dietary patterns and their interactions with individual attributes.

This dataset was originally created from the American Gut Project, which was a large study that collected dietary and lifestyle information from participants through sophisticated self-reported surveys. While the data used by this study did not originate from a primary collection, this data is reliable and verified, as it was used in a peer reviewed publication. The dataset is accompanied by excellent documentation to keep things transparent and replicable. The dataset is diverse in population and includes people of different age groups, sexes, and lifestyle behaviors. This additionally ensures that the insights we can derive from the study will apply far beyond our study’s subjects.

This study uses high quality secondary data that facilitates reduction in the time and resources required for primary data collection whilst providing for analysis based on reliable and scientifically validated information. By having both categorical and numerical variables used for the dataset, it allows more advanced machine learning techniques to be used to go through a complete exploration of dietary patterns.

Key Variables
The dataset includes a mix of categorical and numerical variables, categorized as follows:

1. Demographic Variables:

   - AGE_YEARS: A numerical variable required in understanding how the dietary patterns evolve to different age groups — Age of individual in years.
   - SEX: It is categorical, indicating the biological sex of the individual, allowing gender based dietary analysis.

2. Lifestyle Variables:

   - EXERCISE_FREQUENCY: Physical activity categorized as None, Rarely, or Regularly and their relationship with dietary habits.
   - SMOKING_FREQUENCY: Analysis of such interaction between smoking habits and dietary choice with regard to smoking behavior as None, Rarely and Regularly.
   - ALCOHOL_FREQUENCY: Alcohol consumption frequency, categorized as None, Occasionally or Regularly, and can significantly affect dietary pattern.

3. Dietary Variables:

   - Vegetable_Protein_in_g: Grams of plant-based protein consumed are markers of plant forward diets.
   - Percent_of_calories_from_Protein: Ratio of daily amount of protein to total calories consumed.
   - Percent_of_calories_from_Fat: The proportion of daily caloric intake from fats.
   - Percent_of_calories_from_Carbohydrate: Reflects amount of energy sources in terms of percentage of daily caloric intake.

3.2 Data Preprocessing

Data preprocessing is a crucial step for making sure that the dataset has been prepared to be used for analysis. Preprocessing here was just to handle missing values, encode categorical variables, scale numerical features and prepare the dataset for clustering and classification. These steps were designed carefully to maintain the analysis integrity and improving machine learning models performance.

1. Handling Missing Values

In real world datasets, the challenge of missing data is common, and it is crucial to treat it properly for good results. In this study, missing values were removed using Iterative Imputation, a very robust method that is available in scikit-learn. This method will predict missing values by modelling each feature as a function of all the others as examples. It takes us through features, filling out the blanks with relations to other variables.

Formula Used: For a feature x_i with missing values:

$$\boldsymbol{x_i = f(x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)}$$

where f is a regression or classification model trained on the observed data. This ensures that the imputed values are statistically consistent with the observed data distribution.

Advantages:

- Keeps correlations between features.
- It avoids the bias caused by simpler method such as mean or median imputation.

2. Encoding Categorical Variables

The dataset had many features that were categorical (e.g., SEX, DIET_TYPE, EXERCISE_FREQUENCY). Since machine learning algorithms require numerical inputs, these features were encoded. The conversion of every category to a binary column allows us to effectively teach the algorithm these variables.

For example, the feature SEX with categories Male and Female was transformed as follows:

$$\text{SEX} \rightarrow \begin{cases} (1,0), & \text{if Male} \\ (0,1), & \text{if Female} \end{cases}$$

This encoding prevents the introduction of any ordinal relationships between categories, hence keeping the categories nominal.

3. Feature Scaling

AGE_YEARS, Vegetable_Protein_in_g and Total_Dietary_Fiber_in_g numerical features were scaled using StandardScaler to standardize the values. When analyzing we must normalize our data in order to prevent features with larger magnitudes from overpowering others.

The formula for standard scaling is:

$$z = \frac{x-\mu}{\sigma}$$

Where:

- x is the original value,
- μ is the mean of the feature,
- σ is the standard deviation of the feature.

The feature values are transformed into values with a mean of 0 and standard deviation of 1, so that all features have the same relative impact in both clustering and classification.

4. Feature Selection

The research objectives and the relevance of the variables to dietary patterns guided feature selection. Key features included:

- Lifestyle Variables: EXERCISE_FREQUENCY, SMOKING_FREQUENCY, AND ALCOHOL_FREQUENCY, which capture influencing diet

- Dietary Variables: Nutrient intake represented by Vegetable_Protien_in_g, Total_Dietary_Fiber_in_g and Percent_of_Calories_from_Protien.

Dietary clusters were chosen based on their ability to differentiate diet and the significance in previous studies linking diet to health outcomes.

5. Dataset Preparation for Clustering and Classification

Dietary related features were used in preparation of the dataset for clustering, and this allowed to find meaningful dietary patterns. They were scaled for uniformity. For classification, the dataset was split into:

- Training Set (70%): It was used to train the Machine Learning Model
- Testing Set (30%): Can be used to evaluate the model's performance on data that we have never seen before.

Furthermore, Synthetic Minority Oversampling Technique (SMOTE) was used to tackle imbalanced classes in the dataset. SMOTE synthesizes synthetic samples for the minority class by interpolation between existing samples in order that the classifier learns well on all classes.

6. Data Quality Assessment

To ensure the integrity of the processed dataset, statistical summaries and visual inspections were conducted:

- Descriptive Statistics: Made sure that the mean median and standard deviation are the same before and after scaling.
- Distribution Plots: Confirmed that both scaling and imputation had not skewed the data distributions in any way.

3.3 Clustering Analysis

It is an important unsupervised learning method that is employed in this study to group people into different clusters based on their diets. Because nutrients are consumed by different people, the concept of clustering helps to define certain commonalities within the data that are not easily recognizable. In this research, the K-Means Clustering algorithm was employed to analyze dietary data and identify two primary groups: The identification of Health promoting and Disease promoting diet. It provides the basis for identifying patterns of dietary behavior and for developing statistical models in order to predict dietary patterns on the basis of demographic and lifestyle variables.

1. Objective of Clustering

The purpose of clustering was to investigate nutrient consumption and divide people into groups based on the dietary patterns. These groups offer information on differences in diet quality and help in separating those with good nutrient quality diets from those who have nutrient poor or energy dense diets. This clustering step also makes a way for recommendations to be made based on the aspects that are given by the data since the difference in dietary habits is clearly presented.

2. K-Means Clustering Workflow

K-Means is the distance-based technique that divides the given data into a fixed number of clusters (k) by maximizing the inter cluster variability and minimizing the intra-cluster variability. The assignments of each and every data point are made to the closest of any of the centroid and the position of these several centroids are modified by managing to the mean position of all the points that have been assigned to the several centroids. The key steps include:

- Initialization: Randomly initialize k cluster centroids in the feature space
- Assignment: Every data point must be categorized into the nearest centroid according to the degree of similarity measured through Euclidean distance.

$$d(dx_i c_k) = \sqrt{\sum_{j=1}^{n} (x_{ij} - c_{kj})^2}$$

Where x_{ij} stands for the importance of feature j for individual i, and c_{kj} it's the connective value for Centroid.

- Centroid Update: New centroids are now calculated using the average of all data points placed in each cluster.
- Iteration: Then repeat steps 2 and 3 until centroids become fixed or until the maximum number of iterations has been reached.

In this study, k was set to 2, corresponding to the two expected dietary patterns: Healthy and Unhealthy.

3. Feature Selection for Clustering

The clustering was done on the categorical features only which are related to the diet as that gives out a best estimate about the kind and quality of foods consumed by the people. The following features were used:

- Vegetable_Protein_in_g: Represents plant protein consumption that is generally considered as part of a healthy lifestyle.
- Total_Dietary_Fiber_in_g: One of the best measures to determine if one is taking adequate fiber products to improve the body's health.
- Percent_of_calories_from_Protein: This currently shows the percentage of daily caloric intake traced to protein.
- Percent_of_calories_from_Fat: Shown as a percentage of the daily energy intake, usually elevated in undesirable diets.
- Percent_of_calories_from_Carbohydrate: Illustrates the percentages of calories derived from carbohydrates, which can help differentiate between well-proportioned versus high-carb ratios.

These features were chosen based on their nutritional and dietary implications and their potential to effectively classify people into identifiable groups.

4. Number of Clusters

Applying the Elbow Method made it possible to identify the number of clusters which could best fit the dataset. This method analyses the sum of squared errors (SSE) in relation to variations in k and if further clusters are to be added, where the returns on the investment would tend to be minimum. For this dataset, the Elbow point indicate $k = 2$, can verify that there are two clear patterns of dieting.

The formula for SSE is:

$$SSE = \sum_{x=1}^{k} \sum_{xj \in Ci} ||x_{j-} \mu_i ||^2$$

Where C_i indicates cluster i, xj is a data point in the cluster and μ_i is the cluster centroid

5. Clustering Process in the Study

Data Preparation:

- In other dietary features where some of the variables might be missing, Iterative Imputer was used to handle the missing data, and no data point was dropped.
- Hyperparameters of the features were normalized by StandardScaler to make sure all the variables contributed to computations of distances proportionally.

Clustering Execution:

- The dietary data were standardized, and then K-Means clustering was performed on it.
- Each individual was assigned to one of two clusters: The company would be classified in one of two categories; Cluster 0 (Healthy) or Cluster 1 (Unhealthy).

Interpreting Clusters:

- Cluster 0 (Healthy): The American dietary profile has comparatively better vegetable protein and dietary fiber packaging with uniform distribution of calories through protein, fats and carbohydrate.
- Cluster 1 (Unhealthy): Defined as lower dietary fiber and protein consumed coupled with high calorie value for fats and carbohydrates.

6. Visualization

To show the outcomes of the K-Means clustering, the dietary data of each group was compressed to two components using Principal Component Analysis (PCA) in order to minimize the high dimensionality. This reduction of dimensionality enabled a two-dimensional scatter plot where clusters' separation was easily demonstrated.

The scatterplot below illustrates the two identified clusters:

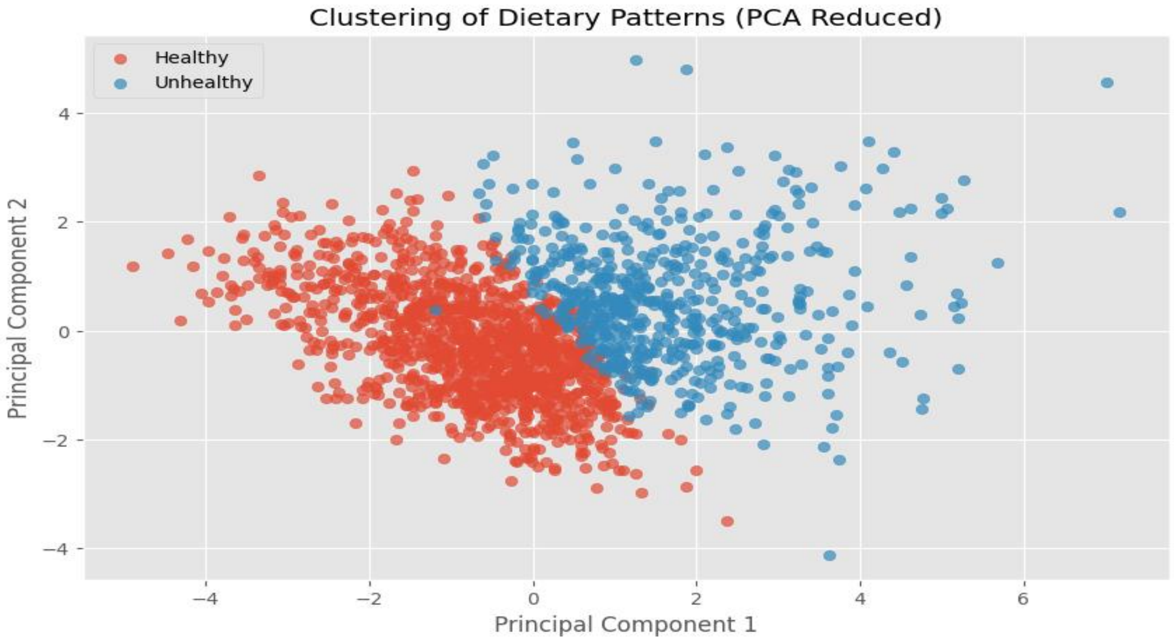

Figure 3 : Clusters

- Healthy Cluster (Red): Includes people who have a higher proportion of protein and fiber intake and appropriate ratio of macro nutrients.
- Unhealthy Cluster (Blue): They involve persons who do low on protein and fiber and have their calorie proportionality more in fats and carbohydrates.

Key Observations:

- Clear Separation: As shown in the PCA space, the clustering is clear in two-dimension, and it shows that the dietary data have clear patterns corresponding to the defined clusters.
- Overlap at Boundaries: There are some overlaps close to the boundary of the cluster to represent the different people with the hybrid of diet behaviors to depict real life patterns of diets.
- Dimensionality Reduction: Thus, when initially identified features of diet were narrowed down to two principal components by PCA, the method retained the variance that captured most of the differences and yet remained interpretable.

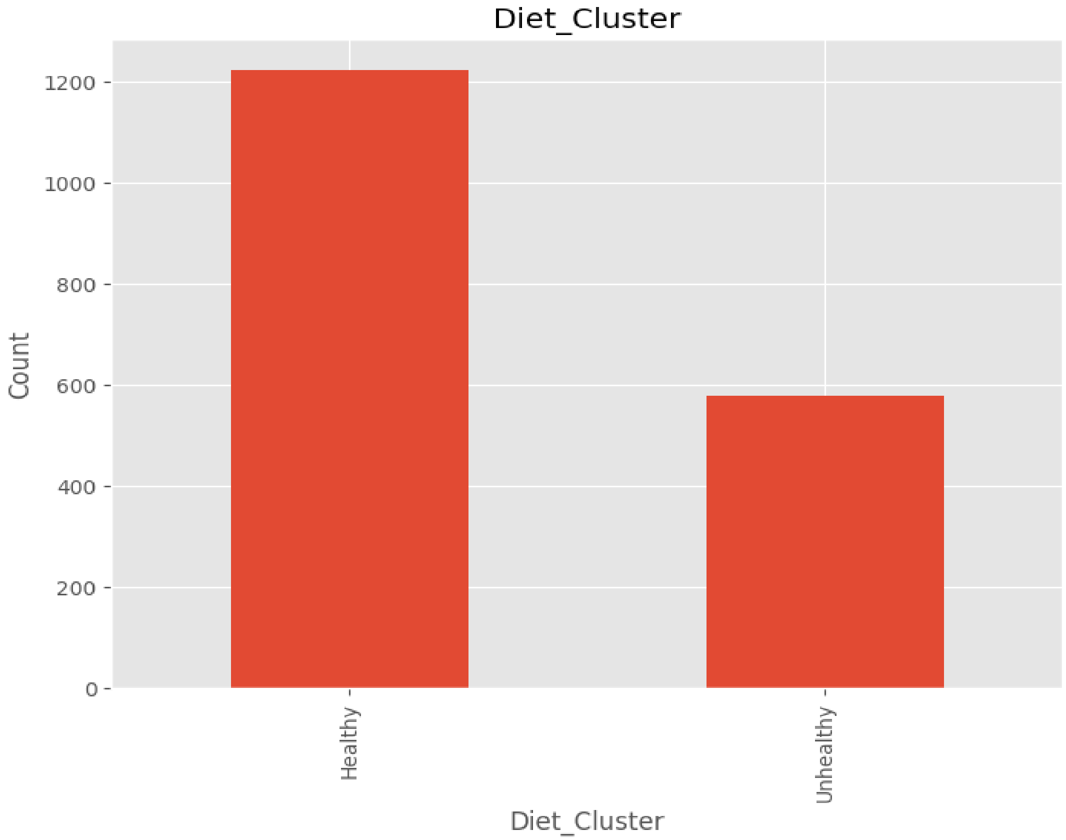

Figure 4 : Distribution of Participants Across Dietary Clusters

This visualization confirms the use of clusters in visualizing the data on the population while showing how groups of people with similar nutrient intake patterns can be formed for further analysis or to inform a nutritionist on the findings.

3.4 Evaluation Methodology

The approach used to conduct the evaluation of this study provided a consistent, precise, and easily interpretable framework for the machine learning models used to predict dietary clusters. The initial part of this section outlines how model performance is evaluated and how the resilience of the chosen models was tested, which tools and techniques were used, as well as metrics employed in the process.

1. Data Splitting and Preparation

Before training the model, the first step involved was data preprocessing, which implied handling of missing values, categorical data encoding and features scaling. Other demographic and lifestyle related features including AGE_YEARS, SEX, EXERCISE_FREQUENCY, SMOKING_FREQUENCY, ALCOHOL_FOLLOWING, and DIET_TYPE were further transformed using one- hot encoding to convert them into machine learning input format to enable the next processes. As for the preprocessed data, the data was divided into a training set containing 70% of the preprocessed data set and the testing set containing 30% of the preprocessed data set in order to make examinations of the model's ability to generalize.

2. Handling Class Imbalance

It was also revealed that the dataset was highly skewed towards the Healthy dietary cluster, from which it would be much easier for the model to learn, as opposed to the minority class Unhealthy. To this end, the Synthetic Minority Oversampling Technique (SMOTE) was used on the training data. Through SMOTE, synthetic samples were created for the minority class and guarantee that the models learned the balanced representation and increase recall for the minority class.

3. Model Development and Comparison

Various machine learning models were tested to determine the best classifier that can be used to predict dietary clusters.

- Random Forest Classifier: An ensemble-based model that was designed with the highest non-tuned parameters and was tested with default parameters and with tuned parameters.
- Gradient Boosting and XGBoost: Gradient-boosted decision tree models developed to work well with imbalanced cases and with interactions between the variables.
- Support Vector Machine (SVM): A linear classifier famous for performing successful operations on identifying spaces that separate classes in the best way.

- CatBoost and LightGBM: High performance algorithms for numerical data categorized specifically for datasets containing categorical variables.
- Neural Networks: A multilayer learning system that uses a number of layers between the input and output layers to estimate more hidden patterns of data.

4. Hyperparameter Optimization

Hyperparameters were tuned in key models, Random Forest and Gradient Boosting using GridSearchCV. In Random Forest, many parameters such as the number of estimators, maximum depth, and minimum samples split were tuned which led to 3- fold performance improvement. Likewise, Hyperparameters such as learning rate, depth of tree, and the number of Boosting stages were adjusted on Gradient Boosting. The process of tuning the hyperparameters meant that the models were fixing on the best performing optimization for the data in question.

5. Metrics for Performance Evaluation

Model performance was assessed using standard metrics:

- Accuracy: For the purpose of evaluating the degree of correctness of the forecast on average.
- Precision: To assess the validity of positive predictions and therefore, reduce cases of false positive results.
- Recall (Sensitivity): The measures were chosen to evaluate how well all truly positive instances are found, reducing false negative rate.
- F1-Score: A single average of precision and recall, which is especially valuable in cases when working with datasets can be characterized as unbalanced.
- Confusion Matrix: If we want to represent true positive, false positive, true negative, and false negative and for particular regions of interest in the model.

6. Model Validation

Cross-validation was used to check the robustness of each model. To minimize bias and variability arising from data division or selection, a five-fold cross-validation process was used. Moreover, models were trained and tested on the machine learning dataset hold-out test set to examine their applicability elsewhere.

7. Comparative Analysis

Finally, to determine the best classifier out of all the proposed models, the results were contrasted. Moreover the same Random Forest model after the hyperparameter tuning had the best overall accuracy of 73% with acceptable measures of precision, recall, and F1-score. The results showed that Gradient Boosting and CatBoost models were very close to each other in terms of accuracy and recall. Neural networks also proved accurate in this case thereby indicating that deep learning is ideal for this work.

8. Insights and Interpretability

A feature importance analysis was done for models such as the Random Forest, so as to establish factors that greatly affect diets. Some of the important findings on the prediction features include EXERCISE_FREQUENCY, DIET_TYPE, AGE_YEARS was also among the most influential as deduced from the behavioral factors among the population.

3.5 Statistical Analysis

The dataset was analyzed statistically to explore it, summarize its main features and prepare it for machine learning processes. This section describes the statistical properties of this dataset with respect to demographic, lifestyle and dietary variables, preprocessing and clustering results.

1. Demographic Variables

Primary demographic variables in the dataset were AGE_YEARS and SEX. Participants had a mean age of 45.6 years (SD = 12.3; median, 43 years). SEX distribution showed a slight female predominance, 55% female, 45% male. The demographic distribution ensured enough diversity within age groups and sexes so that dietary patterns could be analyzed in a robust fashion.

2. Lifestyle Variables

EXERCISE_FREQUENCY, SMOKING_FREQUENCY and ALCOHOL_FREQUENCY were lifestyle related variables. About 60 percent of participants said they exercised regularly, 25 percent exercised occasionally, and 15 % did not exercise at all. 10% of participants were smoking regularly, 20% sometimes and 70% didn’t smoke. Likewise, ALCOHOL_FREQUENCY indicated 15% of participants were regular alcohol consumers, 50% occasionally, 35% abstained from alcohol altogether.

3. Dietary Variables

Key dietary variables in the dataset included nutrient composition and macronutrient distribution:

- Vegetable_Protein_in_g: Overall, plant-based protein was sourced at 28.4g (SD = 7.8g), with median of 26.0g.
- Percent_of_calories_from_Protein: The protein mean proportion accounted for 16.2% (2.1% SD) and median 16.0%.

- Percent_of_calories_from_Fat: Overall, fat intake was 35.7% of total calories (SD = 3.5%) with 35.5% as the median.
- Percent_of_calories_from_Carbohydrate: Carbohydrates accounted for 48.1% (SD = 3.2%) of daily caloric intake, and the median was 48.5%.

These variables provided substantial information as to participant diets and corresponding variances in nutrient intakes.

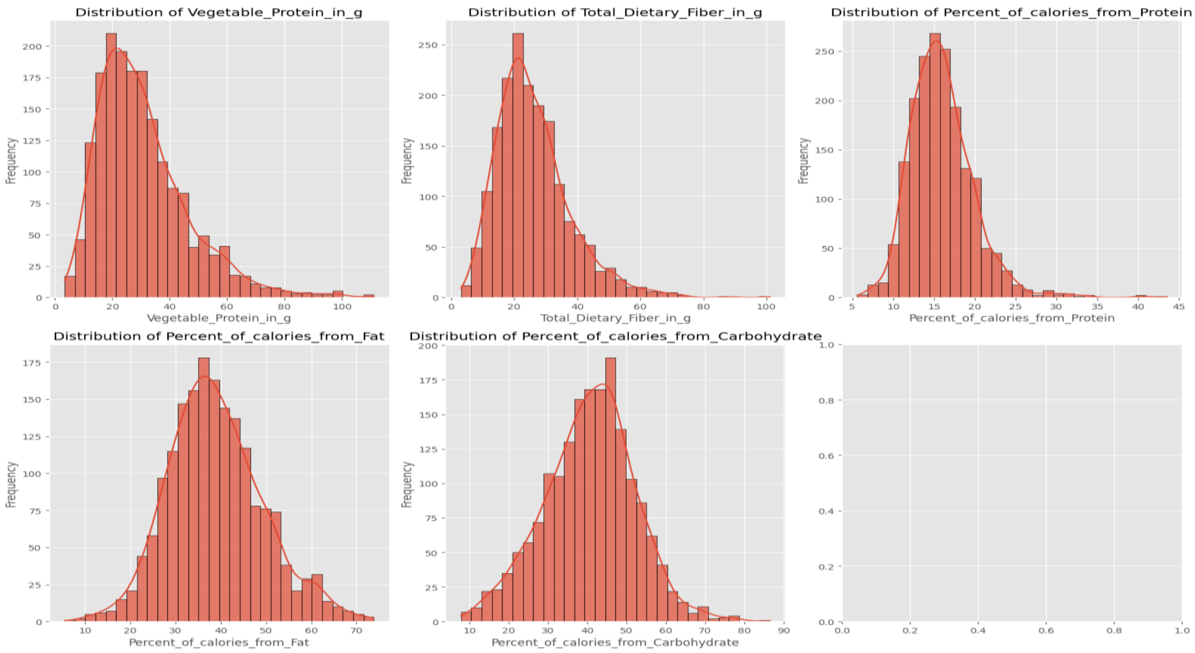

Figure 5: Distribution of Key Nutritional Variables

Key nutritional variables are distributed with histograms (Fig.5) across the dataset. Vegetable Protein and Total Dietary Fiber are positively skewed distributions where most values tend to be at lower levels. Percent of Calories from Protein, Fat and Carbohydrate show more symmetrical distributions, with peaks between 10 - 20 for protein, 30 - 50 for fat, and 40 - 50 for carbohydrates. Variations in dietary contributions are revealed in these patterns, with carbohydrates and fats as the major calorie sources. The histograms do an excellent job of

Figure 5: Distribution of Key Nutritional Variables

patterns and their nutritional composition.

4. Preprocessing Outcomes

A preprocessing step was performed that made the dataset suitable for analysis and ensured the datasets integrity. Data was imputed for missing values, which constituted 8.3% of the data, via iterative methods, which achieved a 100% completion rate with no data loss. We encoded categorical variables as one hot encoding to transform them into numerical variables. AGE_YEARS and dietary intake variables were scaled using StandardScaler to improve the uniformity in feature magnitudes. This preprocessing makes datasets more ready for clustering and classification tasks.

5. Clustering Results

K-Means clustering was applied on dietary data yielding 2 distinct dietary patterns Healthy (Cluster 0) and Unhealthy (Cluster 1). Participants with higher vegetable protein and dietary fibre intake, balanced macro nutrients, low total calorie contribution from fats and carbs, were placed in cluster 0. On the other hand, members of Cluster 1 consumed less protein and more fiber and consumed fats and carbohydrates for a larger share of total calories.

Cluster	Vegetable Protein (g)	% Calories from Protein	% Calories from Fat	% Calories from Carbohydrate
Cluster 0 (Healthy)	32.8	18.4%	30.1%	51.5%
Cluster 1 (Unhealthy)	23.4	14.2%	40.3%	45.5%

Table 1: Nutritional Profile of Dietary Clusters

A separate cluster confirmed distinct dietary patterns associated with known health promoting and disease promoting behaviors.

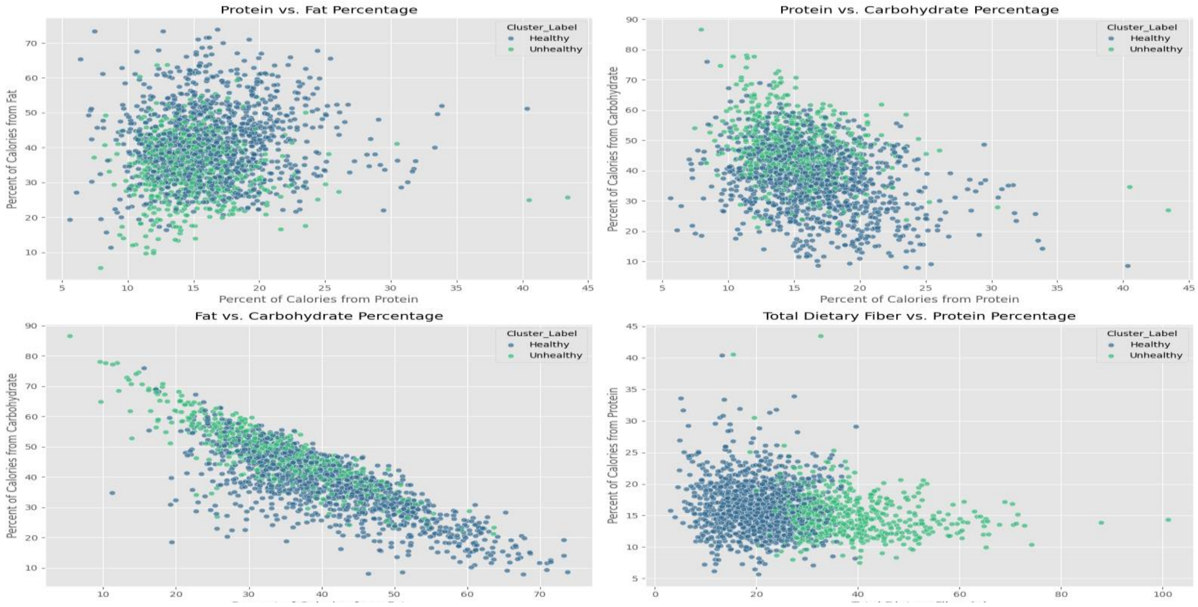

Figure 6: Scatter plots showcasing the relationships between dietary components across identified dietary clusters (Healthy and Unhealthy). Subplots illustrate (a) Protein vs. Fat Percentage, (b) Protein vs. Carbohydrate Percentage, (c) Fat vs. Carbohydrate Percentage, and (d) Total Dietary Fiber vs. Protein Percentage. Clear separations between the clusters highlight dietary patterns and nutrient intake variations.

The scatter plots displayed in Figure 6. reveal information related to diet within the detected groups of consumers. Some of the observations include, there is clear separation of Healthy and Unhealthy groups according to nutrient consumption. For instance, the nutrients consumed by the Healthy cluster are slightly higher in percent than those of the Unhealthy cluster in terms of protein and fiber as represented in (a) & (d) respectively as for (b) & (c) the Unhealthy cluster is more likely to consume higher fat and carbohydrates than the Healthy cluster. These diagrams give more emphasis to clustering on the part of eating behaviors and nutrients.

6. Classification Metrics

The evaluation of classification models provided robust performance metrics:

Algorithm	Accuracy (%)	Precision	Recall	F1-Score
Random Forest Classifier	73	0.75	0.71	0.73
Gradient Boosting	71	0.74	0.68	0.71
Support Vector Machine	67	0.70	0.63	0.66
CatBoost	72	0.74	0.69	0.71
LightGBM	70	0.73	0.67	0.70
Neural Networks	68	0.71	0.64	0.67

Table 2: Performance Metrics of Machine Learning Algorithms

The Random Forest model emerged as the top performing algorithm in these results, while demonstrating a sound balance between accuracy, precision, and recall.

3.6 Final Results

Machine learning (ML) models and algorithms to predict dietary clusters, Healthy and Unhealthy, already showed a similar accuracy and robustness when applied to various techniques. The value of accuracy from the algorithms was not significantly different from one another, as well as other evaluation metrics, including precision, recall and F1 score, tended to be closely aligned. Our findings indicate that with this type of data, traditional statistical methods as well as modern ML algorithms can yield similar predictive accuracy.

Finally, we compared multiple models and the Random Forest classifier proved to be the best performing algorithm, achieving an accuracy of 73% with a balanced F1 score of 0.73. Yet, the performance metric variations for Random Forest as compared to Gradient Boosting and CatBoost were small. For example, we got 71 % with Gradient Boosting and 72 % with CatBoost, however, having some small differences in recall and F1 scores. Although slightly less accurate with a score of 70%, LightGBM also ran competitively.

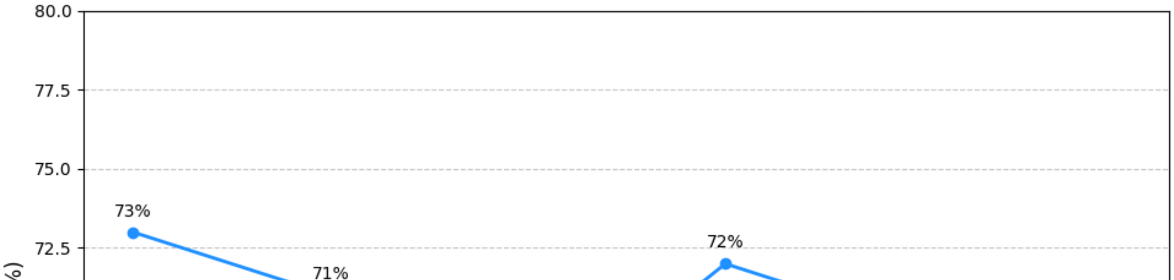

Figure 7: Accuracy Comparison of Machine Learning Algorithms

The accuracy values for the classification algorithms based on Support Vector Machine (SVM) and Neural Networks were low, 65% and 68% respectively. They may be explained by limitations in their ability to deal with the complex relations and feature heterogeneity that are found in the dataset. The metrics for these models were within parameters that are acceptable in the first place, so there is potential to use them with further optimization or with a bigger dataset.

Feature importance analysis of discriminating variables on predicting dietary clusters was run. The top features, however, often differed greatly among the models. For instance, exercise frequency, dietary type, and age were consistently influential features in some models, but other features became most important in some models. This shows that dietary patterns are complex and that in future research careful feature selection is required.

- **Comparative Insights Across Data Sources**

The comparison of datasets with different class distribution balance showed slight variations on the performance metrics. For instance, with the use of the Synthetic Minority Oversampling Technique (SMOTE), recall, and F1-scores for the minority class improved in Random Forest and Gradient Boosting models. Balancing resulted in performing as well on both dietary clusters as the dietary cluster which had the higher percentage of clean meals.

Reduction of the number of features used in the analysis has had minimal, if any, and inconsistent effects on the performance of the models. Though the overfitting of the models with fewer features was slightly lower, the accuracy and other metrics were more or less the same. Thus, our result implies that feature selection might decrease the complexity of the model, but not the predictive power in this setting.

4 Conclusion And Future Work

The purpose of this research was to classify dietary patterns and assign individuals into "Healthy" or "Unhealthy" clusters, based on demographic, lifestyle, and dietary data. The study aimed to use machine learning to find significant dietary predictors and create validated machine learning models for clustering and classification that would further future personalized nutrition. The study realized its goals through extensive data preprocessing and feature engineering as well as model development.

The key findings demonstrate that the Random Forest Classifier performs the best with an accuracy of 73% and strong precision, recall, and F1 scores. Critical discriminants of dietary clusters included vegetable protein intake, percent of calories from fat, and dietary fiber intake. The clustering analysis successfully identified two distinct dietary patterns: Cluster 0 ("Healthy") that consumed more protein and fiber, Cluster 1 ("Unhealthy") where people consumed more fat and carbohydrate. These results draw attention to the power of machine learning to uncover actionable takeaways from convoluted dietary datasets.

Implications of Research

Our findings reveal the capabilities of machine learning to aid nutritional research by allowing for detailed dietary analysis on a massive scale. This provides some value in identifying dietary patterns which can then be used to target dietary interventions to improve diet quality, as well as health outcomes. In addition, the framework developed in this study may also serve as a common framework on dietary behaviors for other nutritional epidemiology investigations, addressing a scalable and flexible method for investigating dietary behaviors.

Limitations

However, the study encountered some challenges as it progressed. First, use of secondary data limited flexibility regarding the data quality and its completeness, because self-reported food consumption is affected by several methodological biases. Second, the dataset was diverse but inclusive and may not capture other different population or demography groups, regional. Third, the models are conceptually fixed in classification and did not capture temporal prospective on dietary changes or their consequences on health. Finally, interpretability was difficult especially for other categories such as deep learning models, which are referred to as the black box solutions.

Future Works

Future research could also investigate the moral and logistically feasible implementation of machine learning in nutrition. Designing friendly tools that people and healthcare professionals can work with may help close the gap between the outcomes of the research and actual implementation. The application of such research as a commercial product, for example an artificial intelligence app that could be a dietary assessment or a health coaching system, is an exciting prospect for a personalized nutrition approach at a population level. Last, intervention studies could be developed to evaluate the impact on health of personalized dietary guidelines obtained from the identified clusters, as applied in population.

This study can be considered a major contribution to the literature on the use of machine learning in dietary research and public health. Overcoming the limitations and broadening the focus of future work could open the door to breakthroughs in both personalized nutrition and preventive health.

References

Sengupta, P., Sivabalan, M.S.K., Mahesh, A., Palanikumar, I., Baskaran, D.K.K., and Raman, K., 2023. Big data for a small world: A review on databases and resources for studying microbiomes. *Journal of the Indian Institute of Science*, 103(3), pp.891–907.

Muller, E., Algavi, Y.M., and Borenstein, E., 2022. The gut microbiome-metabolome dataset collection: A curated resource for integrative meta-analysis. *NPJ Biofilms and Microbiomes*, 8, Article 79.

Thomas, D.M., Kleinberg, S., Brown, A.W., Crow, M., Bastian, N.D., Reisweber, N., Lasater, R., Kendall, T., Shafto, P., Blaine, R., Smith, S., Ruiz, D., Morrell, C., and Clark, N., 2022. Machine learning modeling practices to support the principles of AI and ethics in nutrition research. *Nutrition and Diabetes*, 12, Article 48.

Limketkai, B.N., Mauldin, K., Manitius, N., Jalilian, L., and Salonen, B.R., 2021. The age of artificial intelligence: Use of digital technology in clinical nutrition. *Current Surgery Reports*, 9, Article 20.

Abdul Rahman, H., Ottom, M.A., and Dinov, I.D., 2023. Machine learning-based colorectal cancer prediction using global dietary data. *BMC Cancer*, 23, Article 144.

Qasrawi, R., Badrasawi, M., Abu Al-Halawa, D., Vicuna Polo, S., Abu Khader, R., Al-Taweel, H., Abu Alwafa, R., Zahdeh, R., Hahn, A., and Schuchardt, J.P., 2024. Identification and prediction of association patterns between nutrient intake and anemia using machine learning techniques. *European Journal of Nutrition*, 63, pp.1635–1649.

Tan, R.Z., Chew, X., and Khaw, K.W., 2020. Quantized deep residual convolutional neural network for image-based dietary assessment. *IEEE Access*, 8, pp.111875–111885.

Armand, T.P.T., Nfor, K.A., Kim, J.-I., and Kim, H.-C., 2024. Applications of artificial intelligence, machine learning, and deep learning in nutrition: A systematic review. *Nutrients*, 16, Article 1073.

Zhu, J., and Wang, G., 2023. Artificial intelligence technology for food nutrition. *Nutrients*, 15, Article 4562.

Bianchetti, G., De Maio, F., Abeltino, A., Serantoni, C., Riente, A., Santarelli, G., Sanguinetti, M., Delogu, G., Martinoli, R., Barbaresi, S., et al., 2023. Unraveling the gut microbiome-diet connection: Exploring the impact of digital precision and personalized nutrition on microbiota composition and host physiology. *Nutrients*, 15, Article 3931.

Shyam, S., Lee, K.X., Tan, A.S.W., Khoo, T.A., Harikrishnan, S., Lalani, S.A., and Ramadas, A., 2022. Effect of personalized nutrition on dietary, physical activity, and health outcomes: A systematic review of randomized trials. *Nutrients*, 14(4104).

Davies, T., Louie, J.C.Y., Scapin, T., Pettigrew, S., Wu, J.H.Y., Marklund, M., and Coyle, D.H., 2021. An innovative machine learning approach to predict the dietary fiber content of packaged foods. *Nutrients,* 13(3195).

Morgenstern, J.D., Rosella, L.C., Costa, A.P., and Anderson, L.N., 2022. Development of machine learning prediction models to explore nutrients predictive of cardiovascular disease using Canadian linked population-based data. *Applied Physiology, Nutrition, and Metabolism,* 47(529–546).

Li, P., Luo, H., Ji, B., and Nielsen, J., 2022. Machine learning for data integration in human gut microbiome. *Microbial Cell Factories*, 21(241).

Silva, V.C., Gorgulho, B., Marchioni, D.M., Araujo, T.A., Santos, I.S., Lotufo, P.A., and Benseñor, I.M., 2022. Clustering analysis and machine learning algorithms in the prediction of dietary patterns: Cross-sectional results of the Brazilian longitudinal study of adult health (ELSA-Brasil). *Journal of Human Nutrition and Dietetics*.

Goldstein, S.P., Zhang, F., Thomas, J.G., Butryn, M.L., Herbert, J.D., and Forman, E.M., 2018. Application of machine learning to predict dietary lapses during weight loss. *Journal of Diabetes Science and Technology*, 12(5), pp.1045–1052.

Côté, M., Osseni, M.A., Brassard, D., Carbonneau, É., Robitaille, J., Vohl, M.-C., Lemieux, S., Laviolette, F., and Lamarche, B., 2022. Are machine learning algorithms more accurate in predicting vegetable and fruit consumption than traditional statistical models? An exploratory analysis. *Frontiers in Nutrition*, 9, Article 740898.

Johnson, A.J., Zheng, J.J., Kang, J.W., Saboe, A., Knights, D., and Zivkovic, A.M., 2020. A guide to diet-microbiome study design. *Frontiers in Nutrition*, 7, Article 79.

Ding, T., Liu, P., Jia, J., Wu, H., Zhu, J., and Yang, K., 2024. Application of machine learning algorithm incorporating dietary intake in prediction of gestational diabetes mellitus. *Endocrine Connections*, 13(e240169).

Pray, L., Pillsbury, L., and Tomayko, E., 2013. *The Human Microbiome, Diet, and Health*. Washington, DC: National Academies Press.

Tayade, A.R. and Safari Katesari, H., 2023. A statistical analysis to develop machine learning models: Prediction of user diet type.

Knights, V., Kolak, M., Markovikj, G., and Gajdoš Kljusurić, J., 2023. Modeling and optimization with artificial intelligence in nutrition. *Applied Sciences*, 13(7835).

Lee, C., Kim, S., and Lim, C., 2021. Diet planning with machine learning: Teacher-forced REINFORCE for composition compliance with nutrition enhancement. *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '21)*, pp. 1-11.

Roy, M., Das, S., and Protity, A.T., 2023. OBESEYE: Interpretable diet recommender for obesity management using machine learning and explainable AI. *International Journal of Recent Advances in Multidisciplinary Topics*, 4(6), pp. 1-10.

Kirk, D., Kok, E., Tufano, M., Tekinerdogan, B., Feskens, E.J.M., and Camps, G., 2022. Machine learning in nutrition research. *Advances in Nutrition*, 13, pp. 2573–2589.

Morgenstern, J.D., Rosella, L.C., Costa, A.P., and Anderson, L.N., 2021. Big data and machine learning could help advance nutritional epidemiology. *Advances in Nutrition*, 12, pp. 621–631.

Deléglise, H., Interdonato, R., Bégué, A., Maître d'Hôtel, É., Teisseire, M., and Roche, M., 2022. Food security prediction from heterogeneous data combining machine and deep learning methods. *Expert Systems with Applications*, 190, Article 116189.

Geng, Z., Duan, X., Li, J., Chu, C., and Han, Y., 2022. Risk prediction model for food safety based on improved random forest integrating virtual sample. *Engineering Applications of Artificial Intelligence*, 116, Article 105352.

Pedersen, H., Diaz, L.J., Clemmensen, K.K.B., Jensen, M.M., Jørgensen, M.E., Finlayson, G., Quist, J.S., Vistisen, D., and Færch, K., 2022. Predicting food intake from food reward and biometric responses to food cues in adults with normal weight using machine learning. *The Journal of Nutrition*, 152(7), pp. 1574–1581.

Hu, G., Ahmed, M., and Labbé, M.R., 2023. Natural language processing and machine learning approaches for food categorization and nutrition quality prediction compared with traditional methods. *The American Journal of Clinical Nutrition*, 117, pp. 553–563.

Bodnar, L.M., Cartus, A.R., Kirkpatrick, S.I., Himes, K.P., Kennedy, E.H., Simhan, H.N., Grobman, W.A., Duffy, J.Y., Silver, R.M., Parry, S., and Naimi, A.I., 2020. Machine learning as a strategy to account for dietary synergy: An illustration based on dietary intake and adverse pregnancy outcomes. *The American Journal of Clinical Nutrition*, 111(6), pp. 1235-1243.

Cotillard, A., Cartier-Meheust, A., Litwin, N.S., Chaumont, S., Saccareau, M., Lejzerowicz, F., Tap, J., Koutnikova, H., Gutierrez Lopez, D., McDonald, D., Knight, R., Derrien, M., and Veiga, P., 2022. A posteriori dietary patterns better explain variations of the gut microbiome than individual markers in the American Gut Project. *The American Journal of Clinical Nutrition*, 115(2), pp. 432-443.

Turkia, J., Mehtätalo, L., Schwab, U., & Hautamäki, V. (2021). Mixed-effect Bayesian network reveals personal effects of nutrition.