

Dietary Patterns and Sample-Relative Nutritional Vulnerability among University Students and Lecturers in Bayelsa State, Nigeria: K-Means, Random Forest and Shapley Analysis

Michael Ige Ediabai Edaba

University of Africa, Toru-Orua, Nigeria, Department of Agricultural Economics and Extension

ABSTRACT

Background: This study examined dietary patterns and a constructed, sample-relative index of dietary vulnerability among 216 university students and lecturers surveyed in Bayelsa State, Nigeria, in 2026. A study-specific ten-food-group Dietary Diversity Score (DDS), total food-frequency intensity (TFFI) and nutrient-dense food-frequency score (NDFS) were combined using an equal-weighted, min-max-normalised formula. K-means explored dietary profiles; Random Forest was used only to assess concordance with the index-defined upper-tertile NVI category; and mean absolute SHAP values described feature importance within that fitted classifier. From the results, Mean DDS was 9.63/10 (SD = 0.66), indicating a marked ceiling effect. The three-cluster solution had weak separation (silhouette = 0.133) and is interpreted as exploratory. Random Forest achieved 95.4% hold-out accuracy and ROC-AUC of 0.988, but these values reflect reproduction of a rule derived from the same dietary variables, not independent prediction of clinical nutritional status. Dairy had the highest mean absolute SHAP importance; direction cannot be inferred from mean absolute values alone. The analysis illustrates exploratory profiling of this institutional sample.

KEYWORDS

Dietary patterns;
Dietary diversity;
Sample-relative index;
Random Forest;
SHapley Additive exPlanations (SHAP);
Bayelsa State

ARTICLE HISTORY

Received 22 February 2026;

Accepted 13 March 2026; Published 23 March 2026

Introduction

Adequate and varied diets are central to health, learning capacity, productivity and long-term socioeconomic wellbeing. Dietary diversity indicators are widely used as practical proxies for food access and diet quality because they summarise the number of food groups consumed over a specified reference period. Greater diversity is generally associated with improved micronutrient adequacy, although a high food-group count does not by itself establish adequate portion sizes, nutrient density or clinical nutritional status [1,2]. Nigeria continues to face substantial food-security and diet-quality challenges, with important differences across locations and population groups. Evidence from south-eastern Nigeria shows that households may report broad consumption of staple groups while still experiencing restricted food quality, inadequate access and low consumption of animal-source foods [3]. These findings support analytical approaches that move beyond a single average dietary-score and examine both frequency and composition.

Artificial intelligence (AI) and machine learning provide tools for examining complex dietary data. Unsupervised algorithms can group respondents with similar food-consumption profiles, while supervised algorithms can classify internally defined risk categories. Reviews of AI in nutrition research show growing use of machine learning for dietary assessment, food recognition, nutrient estimation and intake prediction, but they also highlight persistent concerns about validation, representativeness, interpretability and professional oversight [4-6]. K-means clustering is useful for identifying latent dietary patterns, whereas Random Forest can model nonlinearities and interactions without imposing a single parametric form [7,8]. Explainable AI methods such as SHapley Additive exPlanations (SHAP) can then identify the variables that contribute most strongly to model classifications [9]. These methods are especially valuable when the objective is to translate multidimensional dietary records into interpretable profiles for research and policy.

Few studies in Bayelsa State have combined dietary-pattern

clustering with transparent machine-learning interpretation. The present study contributes an exploratory workflow for a defined institutional sample, while explicitly separating three analytical tasks: description of recorded food-group frequencies, exploratory clustering, and internal rule-emulation by a classifier. It does not claim that machine learning independently diagnoses malnutrition or that the constructed index has clinical validity.

Objectives of the study are to:

1. Describe the respondent profile, food-frequency distribution and ten-food-group Dietary Diversity Score.
2. Construct a transparent, equal-weighted, sample-relative Nutritional Vulnerability Index from dietary diversity, total food-frequency intensity and nutrient-dense food-frequency indicators.
3. Identify latent dietary-consumption patterns using K-means clustering.
4. Classify upper-tertile NVI category using a Random Forest model and evaluate its internal performance.
5. Explain the relative contribution of food groups using SHAP and derive cautious, context-sensitive policy implications.

Conceptual and Theoretical Framework

Conceptual framework

The conceptual framework links the recorded frequencies for ten food groups to three transparent summary indicators and a constructed, sample-relative Nutritional Vulnerability Index (NVI). K-means is used to explore overlapping dietary profiles. Random Forest evaluates how closely a nonlinear classifier reproduces the upper-tertile NVI category from variables in the same dietary domain and is therefore a rule-emulation exercise rather than independent prediction. SHAP values explain feature contributions within the fitted classifier but do

*Correspondence: Dr. Michael Ige Ediabai Edaba, Department of Agricultural Economics and Extension, University of Africa, Toru-Orua, Nigeria. e-mail: michael.edaba@uat.edu.ng

not establish causal effects. All interpretations are restricted to the sampled university students and lecturers.

Conceptual framework for dietary-pattern profiling and sample-relative vulnerability analysis

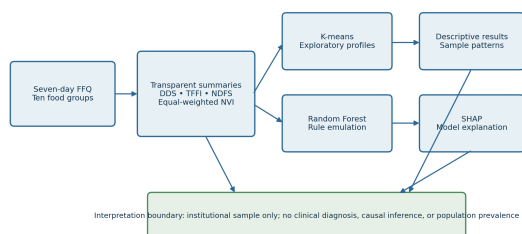


Figure 1. Conceptual framework for dietary-pattern profiling and sample-relative nutritional vulnerability analysis.

Theoretical framework

The first theoretical foundation is the food-security and dietary-diversity perspective. It holds that access to a wider range of food groups generally increases the likelihood of obtaining diverse nutrients. In field studies where detailed nutrient computation is unavailable, food-group counts can provide rapid indicators of dietary variety and food access. However, such indicators require careful alignment with the target population and survey instrument, and they should not be interpreted as direct measures of clinical nutritional adequacy [1,2].

The second foundation is the food-utilisation perspective. Food availability alone does not ensure nutritional wellbeing; individuals must also be able to access, prepare and consume foods in adequate quantity and quality. Consequently, respondents with similar diversity counts may differ in consumption frequency, nutrient density, affordability and resilience to food-price or supply shocks. The NVI is designed to represent this multidimensional dietary-risk logic within the limitations of the available variables.

The third foundation is statistical learning theory. K-means partitions observations according to similarity in multivariate food-frequency profiles, while Random Forest aggregates multiple decision trees to classify an outcome and reduce the instability associated with a single tree [7,8]. Because clustering is exploratory and dependent on preprocessing, distance metrics and the selected number of clusters, validation measures such as the silhouette coefficient are needed to judge whether the groups are well separated [10].

The fourth foundation is explainable AI. High predictive performance is insufficient for health-related decision support when the model cannot be interpreted. SHAP provides local and global explanations by estimating how each feature contributes to a model output. Nevertheless, feature contribution is not the same as causal effect, and AI systems in health and nutrition should remain transparent, accountable, inclusive and subject to human oversight [9,11].

Methodology

Study area

The study was conducted in Bayelsa State, Nigeria, located in the core Niger Delta region between latitudes 4°15'N and 5°23'N and longitudes 5°22'E and 6°45'E. The state is bounded by Delta State to the north, Rivers State to the east, and the Atlantic Ocean to the south and west. Bayelsa State comprises eight Local Government Areas (LGAs), namely Brass, Ekeremor, Kolokuma/Opokuma, Nembe, Ogbia, Sagbama, Southern Ijaw and Yenagoa. The state is characterized by an intricate network of rivers, creeks, mangrove forests and freshwater swamps, making it one of Nigeria's predominantly riverine environments.

Bayelsa State provides a relevant setting for examining institutional dietary patterns because of its riverine food environment and unequal access to some nutrient-dense foods. The study, however, did not sample the general state population; inference is restricted to the participating university students and lecturers.

Research design

The study adopted a cross-sectional exploratory survey design. Dietary information was collected once and analysed descriptively and through K-means clustering, Random Forest rule-emulation and SHAP feature-importance analysis. The design supports pattern description and internal analytical comparison, not causal inference, clinical diagnosis or population-wide prevalence estimation.

Population of the study

The target population for this analysis comprised eligible students and lecturers at the participating tertiary institutions in Bayelsa State. The analysed sample included 162 students and 54 lecturers ($n = 216$). Although recruitment was geographically distributed across the eight LGAs, the sampling frame was institutional and does not represent all adult residents of Bayelsa State.

Sample size determination

The study used a fixed, equal-allocation sample of 216 respondents. This was a feasibility-based design and was not a population-proportionate allocation. Equal numbers were assigned to political/geographical units to obtain balanced coverage; consequently, state-level prevalence estimates and design-based population inference are not warranted. The study employed institutional stratified quota sampling design with 72 students and 20 lecturers from Niger Delta University, 30 students and 10 lecturers from Bayelsa Medical University, 40 students and 14 lecturers from the University of Africa, and 20 students and 10 lecturers from the Federal University of Agriculture, Bassambiri. Because the retained dataset identifies respondents as students or lecturers, the units and recruitment sites were reported as institutional sampling locations rather than as a representative household or resident sample.

Sampling technique

A multistage, institution-based sampling technique was adopted for the study. In the first stage, four tertiary institutions in Bayelsa State were selected as recruitment sites. These were Niger Delta University, Amasoma, Bayelsa Medical University, Yenagoa, University of Africa, Toru-Orua, and Federal University of Agriculture, Bassambiri. In the second stage, eligible students and lecturers were identified within the selected institutions. The distribution comprised 72 students and 20 lecturers from Niger Delta University, 30 students and 10 lecturers from Bayelsa Medical University, 40 students and 14 lecturers from the University of Africa, and 20 students and 10 lecturers from the Federal University of Agriculture, Bassambiri. Participants were required to fall within the age range of 18 to 60 years, be registered students or serving lecturers in the selected institutions, and provide informed consent. Within each institution, respondents were selected using stratified quota sampling. Of the 240 individuals approached, 216 completed the questionnaire, representing a response rate of 90%.

Method of data collection

Primary data were collected through face-to-face interviews in 2026 using a structured Food Frequency Questionnaire (FFQ) designed to obtain information on respondents' consumption of ten food groups: cereals, roots and tubers, pulses and legumes, vegetables, fruits, meat products, fish and seafood, dairy products, fats and oils, and beverages, alongside selected socioeconomic and demographic characteristics. Responses were coded according to the number of days each food item was consumed during the seven-day reference period: not consumed = 0, consumed on one day = 1, two days = 2, three days = 3, four days = 4, five days = 5, six days = 6, and every day

= 7. Item-level scores were aggregated by summing the scores of all food items belonging to each of the ten food groups to obtain the respective food-group frequency scores. A food group was classified as consumed and assigned a value of 1 for the Dietary Diversity Score when at least one item within that group had a frequency score greater than zero; otherwise, it was coded 0. Higher aggregated scores therefore represented more frequently reported consumption during the reference period but did not indicate portion sizes, quantities consumed, energy intake or nutrient adequacy."

The FFQ was adapted from the Food and Agriculture Organization's dietary-diversity guidelines and the Household Dietary Diversity Score (HDDS) instrument developed by Swindale and Bilinsky [2,12]. The original food-group structure was reviewed and modified to reflect commonly consumed foods in Bayelsa State, including locally available cereals, roots and tubers, legumes, vegetables, fruits, meat, fish and seafood, dairy products, fats and oils, and beverages. Adaptation involved replacing unfamiliar food examples with locally recognized foods while retaining the underlying food-group classifications. The questionnaire was initially prepared in English and administered in English because the participants were university students and lecturers. Where clarification was necessary, trained interviewers explained questions in Nigerian Pidgin or the respondent's preferred local language without altering their intended meaning. The adapted instrument was reviewed for content relevance, clarity and cultural suitability before field administration. The adapted questionnaire was pretested among 20 students and lecturers from Michael Okpara University of Agriculture, Umudike, an institution outside the final sample to assess its clarity, cultural appropriateness, response consistency and administration time, after which ambiguous questions and food examples were revised before the main survey.

Method of data analysis

The analysis was conducted using Python version 3.12.13 and scikit-learn version 1.8.0 [13]. Signed feature contributions were calculated using an exact model-agnostic Shapley-value implementation. Data preparation and numerical computations were performed using pandas version 2.2.3 and NumPy version 2.3.5.

DDS, TFFI and NDFS are overlapping transformations of the same ten food-frequency variables. Internal-consistency, KMO, Bartlett and PCA statistics therefore describe dependence among derived indicators; they do not establish external construct or clinical validity. The equal-weighted index is retained for transparency, and the PCA-weight sensitivity analysis is interpreted only as evidence that respondent rankings are insensitive to the two examined weighting schemes. External validation requires anthropometric, biochemical, clinical, nutrient-intake or established food-security outcomes.

Dietary diversity score

A study-specific DDS was calculated across the ten recorded food groups. A group received 1 when its recorded frequency exceeded zero and 0 otherwise, during the seven-day reference period; a "greater than zero" means consumption at least once during that period.

$$DDS_i = \int_{j=1}^{10} I(F_{ij} > 0) \backslash \text{hfil} (1)$$

The DDS ranged from 0 to 10 and should not be interpreted as either the standard 12-food-group HDDS or the Minimum Dietary Diversity for Women indicator. The high mean score of 9.63, together with the finding that 71.8% of respondents consumed all ten food groups, indicates a pronounced ceiling effect and limited variation across respondents. Accordingly, the DDS was interpreted solely as an indicator of broad exposure to the measured food groups during the seven-day reference period and was not compared directly with established short-reference dietary-diversity thresholds.

Total food-frequency intensity

Total food-frequency intensity (TFFI) was computed as the sum of the reported frequency scores across the 10 food groups. The indicator

captures the overall intensity of reported consumption but does not represent portion size, energy intake or nutrient quantity. The Total Food-Frequency Intensity is expressed as:

$$TFFI_i = \int_{j=1}^{10} F_{ij} \backslash \text{hfil} (2)$$

Nutrient-Dense Food-Frequency Score (NDFS)

A NDFS was constructed from pulses/legumes, vegetables, fruits, meat products, fish/seafood and dairy. These groups were selected to complement staple-frequency information with protein-rich and protective-food dimensions. The score is an analytical proxy and not a substitute for nutrient composition analysis. Hence, the NDFS is expressed as:

$$NDFS_i = \int_{j \in D} F_{ij} \backslash \text{hfil} (3)$$

where D = {pulses, vegetables, fruits, meat, fish, dairy}

Constructed sample-relative nutritional vulnerability index

DDS, TFFI and NDFS were min-max normalised to 0-1 and combined with equal weights so that lower diversity and frequency produced higher index values. The resulting NVI is a transparent, sample-relative dietary index, not an AI-generated index, validated clinical scale or diagnosis. Equal weights were used because no independent clinical outcome was available for outcome-derived weighting.

$$NVI_i = 1 - \frac{DDS_i + TFFI_i + NDFS_i}{3} \backslash \text{hfil} (4)$$

Respondents in the upper-tertile NVI category were coded 1, while all other respondents were coded 0. This operational definition is internally consistent and sample-relative; it is not a medical diagnosis of malnutrition.

K-Means clustering

$$Z_{ij} = (X_{ij} - \hat{X}_j) / S_j k = 2k = 5n_{init} = 5010^{-}$$

K-means clustering was used to identify similarities among respondents from their consumption-frequency scores for the ten food groups. Before clustering, each variable was standardised using $z_{ij} = (x_{ij} - \bar{x}_j) / s_j$ (Equation 5a), where x_{ij} is respondent i 's score for food group j , $\bar{x}_j$ is the sample mean and s_j is the sample standard deviation. Candidate solutions from $k = 2$ to $k = 5$ were estimated using K-means++ initialisation, 50 independent initialisations, a maximum of 500 iterations, a convergence tolerance of 10^{-4} and random seed 42. The number of clusters was compared using average silhouette coefficients. Stability of the three-cluster solution was assessed over 100 bootstrap resamples using the adjusted Rand index. Standardised centroids characterised each profile; positive and negative values denote consumption above and below the sample mean, respectively. Given the low silhouette coefficient, the clusters were treated as overlapping, exploratory profiles rather than clinically meaningful groups.

$$J = \int_{i=1}^k \int_{x \in C_i} \|x - \mu_i\|^2 \backslash \text{hfil} (5)$$

Where:

J = within-cluster sum of squares (objective function);

k = number of clusters;

C_i = observations belonging to cluster i ;

x = observation vector;

μ_i = centroid of cluster i .

Random forest classification

Random Forest classified the binary upper-tertile NVI category from the ten food-frequency variables. A 70:30 stratified hold-out split and five-fold cross-validation were reported. The test confusion matrix contained 43 lower-vulnerability cases and 22 high-vulnerability cases ($n = 65$).

$$n = 151n = 65$$

Random Forest was used to classify respondents into lower category and upper-tertile NVI category using the ten food-frequency variables. The high-vulnerability outcome was defined as membership in the upper-tertile NVI category. The complete dataset contained 144 lower-vulnerability cases and 72 high-vulnerability cases. A stratified 70:30 hold-out procedure allocated 101 lower-vulnerability and 50 high-vulnerability cases to the training set (), and 43 lower-vulnerability and 22 high-vulnerability cases to the test set (). Five-fold stratified cross-validation was subsequently conducted on the complete dataset while preserving the class distribution within each fold. Before modelling, duplicate and incomplete records were examined; observations with missing food-frequency values were removed. Model was estimated with 500 decision trees, $\text{max_features} = \text{"sqrt"}$, unrestricted maximum tree depth, a minimum of two observations per terminal leaf, balanced class weights, and a random seed of 42. Hyperparameters were selected through five-fold stratified grid-search cross-validation using ROC-AUC as the optimization criterion. Model performance was assessed using accuracy, precision, recall, F1 score, ROC-AUC and the confusion matrix. The analysis was conducted using Python version 3.12.13 and scikit-learn version 1.8.0. Signed feature contributions were calculated using an exact model-agnostic Shapley-value implementation. Data preparation and numerical computations were performed using pandas version 2.2.3 and NumPy version 2.3.5.

Uncertainty around the hold-out performance estimates was assessed using 10,000 paired stratified bootstrap resamples with a random seed of 42. Within each bootstrap iteration, lower-category and upper-tertile cases were sampled separately with replacement, thereby preserving the original test-set class distribution of 43 and 22 cases, respectively. The 2.5th and 97.5th percentiles of the resulting bootstrap distributions were used to construct 95% confidence intervals for accuracy, F1 score and ROC-AUC.

$$p_i = \frac{1}{B} \int_{b=1}^B I(h_b(x_i) = 1) \, db \quad (6)$$

where

p_i = predicted probability of high nutritional vulnerability for respondent i ;

B = total number of trees in the forest;

$h_b(x_i)$ = prediction of the b^{th} decision tree;

$I(\cdot)$ = indicator function, equal to 1 when the condition is true and 0 otherwise.

This is an internal rule-emulation analysis because the outcome

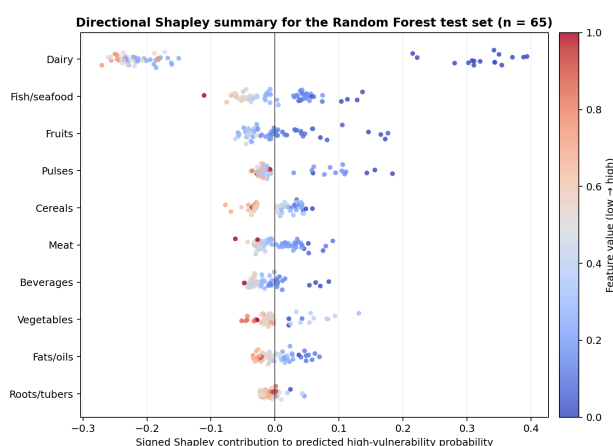


Figure 2. Directional Shapley contributions of the ten food groups to the Random Forest-predicted probability of membership in the upper-tertile NVI category.

was algebraically derived from the same dietary domain as the predictors. Direct application of the NVI threshold is the transparent reference rule and, by definition, reproduces the outcome perfectly. Random Forest cannot improve upon that rule; its performance quantifies only approximate concordance after train-test splitting.

Figure 2 presents the directional Shapley contributions of the ten food groups to the Random Forest's predicted probability of membership in the upper-tertile NVI category. Dairy was the most influential feature, displaying the widest distribution of Shapley values, followed by fish/seafood, fruits, pulses and cereals. Low recorded dairy scores, particularly zero consumption, produced large positive Shapley values and therefore increased the model's predicted probability of membership in the upper-tertile NVI category, whereas higher dairy scores generally generated negative contributions. A similar but weaker pattern was observed for fish/seafood, fruits, cereals, vegetables and meat, where lower frequency scores tended to increase predicted vulnerability and higher scores reduced it. Nevertheless, SHAP values explain how features influence the predictions of the fitted model and should not be interpreted as causal effects, particularly because the vulnerability category was constructed from dietary indicators derived from the same food-frequency variables [9].

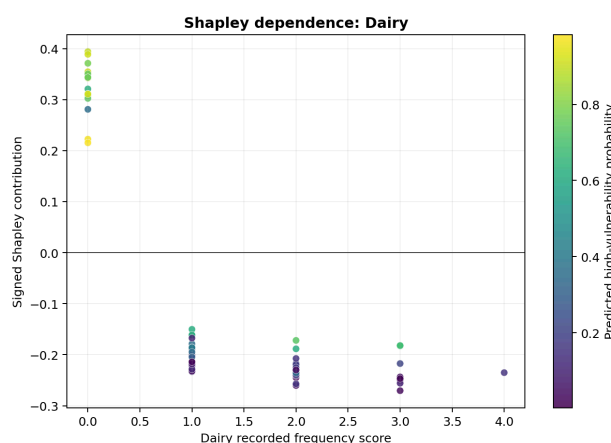


Figure 3. Dependence of dairy-frequency scores on Shapley contributions to the Random Forest-predicted probability of membership in the upper-tertile NVI category.

Figure 3 further illustrates the relationship between recorded dairy-frequency scores and their contributions to the Random Forest predictions. Respondents with a dairy score of zero recorded strongly positive Shapley values, indicating substantial upward shifts in the predicted probability of high vulnerability relative to the background prediction. In contrast, dairy-frequency scores ranging from one to four were associated with consistently negative Shapley values, suggesting that reported dairy consumption reduced the model's estimated probability of belonging to the high-vulnerability category. The magnitude of the negative contribution generally increased with higher dairy-frequency scores, although some variability remained because of interactions with other food groups. Dependence plots reveal how feature values relate to model contributions but do not establish that dairy consumption independently causes lower nutritional vulnerability [14].

Results

Respondent and digital-readiness profile

The cleaned dataset contained 216 respondents. Students accounted for 75.0%, and lecturers and 25.0% respectively; females represented 54.6% and males 45.4%. Most respondents had regular internet access, while slightly more than half reported previous AI training.

The contextual variables in Table 1 describe the sample but were not included as food-group predictors in the Random Forest model.

Table 1. Respondents' Institutional, Demographic and Digital-Readiness Characteristics (N = 216).

Variable	Category	Frequency (N)	Percentage (%)
Respondent status	Student	162	75.0
	Lecturer	54	25.0
Institution	Niger Delta University, Amassoma	92	42.6
	Bayelsa Medical University, Yenagoa	40	18.5
	University of Africa, Toru-Orua	54	25.0
	Federal University of Agriculture, Bassambiri	30	13.9
Gender	Female	118	54.6
	Male	98	45.4
Education	Tertiary completed	122	56.5
	Secondary completed	94	43.5
Regular internet access	Yes	192	88.9
	No	24	11.1
Previous training	AI Yes	118	54.6
	No	98	45.4
Academic performance	Excellent	44	20.4
	Good	142	65.7
	Fair	24	11.1
	Poor	6	2.8
Support needed most	Training	108	50.0
	Infrastructure	44	20.4
	Credit	38	17.6
	AI tools	26	12.0

Source: author's computation

Note. Percentages are based on 216 valid responses. Institutional quotas were: Niger Delta University- 72 students and 20 lecturers; Bayelsa Medical University-30 students and 10 lecturers; University of Africa-40 students and 14 lecturers; and Federal University of Agriculture, Bassambiri-20 students and 10 lecturers. Of 240 eligible individuals approached, 216 completed the questionnaire, giving a response rate of 90.0%.

Source: field survey, 2026

Food-frequency and dietary diversity profile

The mean Dietary Diversity Score (DDS) was 9.63 out of 10 food groups (SD = 0.66; range = 7–10), while 155 respondents (71.8%) reported consuming all ten food groups during the seven-day reference period. This concentration at the upper boundary indicates a pronounced ceiling effect and limits the ability of the DDS to discriminate adequately among respondents. Although most food groups recorded high consumption prevalence, considerable differences were observed in their reported frequency. Vegetables and roots and tubers recorded the highest mean frequency scores, whereas dairy had the lowest mean and consumption prevalence, followed by beverages and fish/seafood.

Table 2. Food-frequency profile by food group.

Food group	Mean	Standard dev.	Minimum	Maximum	Consumed (%)
Cereals	6.26	2.30	1	15	100.0
Roots and tubers	12.90	3.21	2	21	100.0
Pulses/legume	3.71	1.62	1	12	100.0
Vegetables	13.09	3.39	1	24	100.0
Fruits	5.62	3.00	0	23	99.5
Meat products	8.02	3.86	2	41	100.0
Fish/seafood	4.30	2.26	0	14	98.6
Dairy	1.27	1.01	0	4	73.1
Fats and oils	5.12	2.11	1	11	100.0
Beverages	2.15	1.31	0	7	91.7

Note: Values represent aggregated food-frequency scores during the seven-day reference period. Consumed (%) represents the percentage of respondents with a score greater than zero. The scores do not measure serving sizes, quantities consumed, energy intake, nutrient adequacy or clinical nutritional status

Source: field survey, 2026

Latent dietary patterns from K-means

The average silhouette coefficients were 0.1339 for k = 2, 0.1344 for k = 3, 0.1233 for k = 4, and 0.1183 for k = 5. Although the three-cluster solution recorded the highest coefficient, its advantage over the two-cluster solution was very small. Moreover, the low coefficient of

Table 3A. Silhouette coefficients for candidate cluster solutions.

Number of clusters	Silhouette coefficient
2	0.1339
3	0.1344
4	0.1233
5	0.1183

0.1344 indicates substantial overlap and weak separation among the dietary profiles. Across 100 bootstrap resamples, the mean adjusted Rand index was 0.579 (SD = 0.203; median = 0.573; 95% bootstrap interval: 0.195–0.911), indicating moderate but variable cluster stability. Accordingly, the clusters were interpreted as exploratory dietary profiles rather than stable or sharply differentiated population groups.

Source: field survey, 2026

Table 3B presents the dietary and vulnerability characteristics of the three-cluster solution. Cluster 1 contained 15 respondents (6.9%) and recorded the lowest mean DDS of 8.27, the lowest mean total food-frequency score of 38.40 and the highest mean NVI of 0.74. All respondents in this cluster were classified in the upper-tertile NVI category. Cluster 2 accounted for 79 respondents (36.6%) and displayed intermediate values, including a mean DDS of 9.52, total frequency of 58.04 and mean NVI of 0.46. Cluster 3 comprised 122 respondents (56.5%) and recorded the highest mean DDS and total frequency but the lowest mean NVI.

Table 3B. K-means cluster profile and nutritional vulnerability.

Cluster	Frequency	Percentage	Mean DDS	Mean total freq.	Mean NVI	Upper-tertile NVI category (%)
Cluster 1	15	6.9	8.27	38.40	0.74	100.0
Cluster 2	79	36.6	9.52	58.04	0.46	60.8
Cluster 3	122	56.5	9.87	68.25	0.34	7.4

Note. Cluster 1 = low-frequency/high-vulnerability pattern; Cluster 2 = tuber- and vegetable-dominant/moderate-vulnerability pattern; Cluster 3 = comparatively diversified/lower-vulnerability pattern.

The standardized centroids in Table 3C provide further evidence of the food groups distinguishing the clusters. Cluster 1 recorded below-average values for all ten food groups, with particularly large negative deviations for vegetables, roots and tubers, beverages, dairy, and fats and oils. Cluster 2 recorded above-average consumption of roots and tubers and vegetables but below-average values for fruits, meat, fish/seafood, dairy and pulses. Cluster 3 exhibited above-average

Table 3C. Standardized food-group centroids for the three-cluster solution.

Food group	Cluster 1	Cluster 2	Cluster 3
Cereals	-0.494	-0.569	0.429
Roots and tubers	-1.779	0.509	-0.111
Pulses/legumes	-0.807	-0.309	0.299
Vegetables	-2.111	0.364	0.024
Fruits	-0.185	-0.669	0.456
Meat products	-0.523	-0.424	0.339
Fish/seafood	-0.783	-0.576	0.469
Dairy	-0.940	-0.383	0.364
Fats and oils	-0.912	-0.051	0.145
Beverages	-1.256	-0.037	0.178

values for cereals, pulses, fruits, meat, fish/seafood and dairy, supporting its interpretation as the comparatively diversified profile.

Note. Cluster 1 represents the low-frequency/high-vulnerability profile (n = 15); Cluster 2 represents the root-, tuber- and vegetable-oriented profile (n = 79); and Cluster 3 represents the comparatively diversified profile (n = 122). Positive centroid values indicate consumption above the sample mean, whereas negative values indicate consumption below the sample mean.

Source: field survey, 2026

Internal random forest classification performance

The Random Forest reproduced the upper-tertile NVI rule with 95.4% hold-out accuracy. The test confusion matrix contained 42 true lower-vulnerability classifications, 20 true upper-tertile classifications, one false positive and two false negatives (43 lower-category and 22 upper-tertile cases). Five-fold cross-validation estimates are reported separately in Table 4. Paired stratified bootstrap estimates based on 10,000 resamples gave 95% confidence intervals of 0.892–1.000 for accuracy, 0.837–1.000 for the F1 score and 0.959–1.000 for ROC-AUC, demonstrating uncertainty expected from a hold-out sample of 65 observations.

Table 4. Random Forest performance for classification of upper-tertile NVI category.

Metric	Value
Hold-out accuracy	0.954
Precision	0.952
Recall	0.909
F1 score	0.930
ROC-AUC	0.988
5-fold CV accuracy	0.940 (SD = 0.031)
5-fold CV F1	0.909
5-fold CV ROC-AUC	0.986

Note. The high-vulnerability outcome was defined as membership in the upper-tertile NVI category. Accuracy = 0.954 (95% bootstrap CI: 0.892–1.000); F1 = 0.930 (0.837–1.000); ROC-AUC = 0.988 (0.959–1.000). These statistics quantify internal rule concordance and do not establish external or clinical validity.

Source: field survey, 2026

Explainable AI drivers of classification

Dairy had the largest mean absolute SHAP value, followed by fish/seafood, fruits, pulses/legumes and cereals.

Table 5. SHAP and Random Forest feature importance.

Rank	Food group	Mean absolute value
1	Dairy	0.243
2	Fish/seafood	0.046
3	Fruits	0.042
4	Pulses/legumes	0.035
5	Cereals	0.029

Table 5 (continued)

Rank	Food group	Mean absolute value
6	Meat products	0.027
7	Beverages	0.026
8	Vegetables	0.024
9	Fats and oils	0.022
10	Roots and tubers	0.013

Note. Mean absolute Shapley values represent the average magnitude of each food group's contribution to the Random Forest-predicted probability of membership in the upper-tertile NVI category across the 65 hold-out observations. Larger values indicate greater average influence on predictions but neither reveal direction nor establish causality. Directional contributions are presented in Figures 2 and 3.

Source: Authors' computation from field-survey data, 2026.

Dairy was highly discriminative within the fitted classifier, plausibly because it had the lowest prevalence and mean score and was also included in the NDFS used to construct the outcome. This built-in relationship may inflate its importance. SHAP explains the fitted model, not biological mechanisms, causal effects or the consequences of changing consumption [9,14].

Note. DDS = Dietary Diversity Score; TFFI = Total Food Frequency Intensity; NDFS = Nutrient-Dense Food Frequency Score; NVI = Nutritional Vulnerability Index; KMO = Kaiser-Meyer-Olkin Measure of Sampling Adequacy; PCA = Principal Component Analysis.

Discussion

Table 1 describes an institution-based sample rather than the general Bayelsa population. High reported internet access may support digitally assisted data collection in these participating settings, but it should not be extrapolated to all residents. Previous AI training and perceived support needs are contextual variables and were not used to validate the dietary index.

Table 2 shows broad recorded exposure to the ten food groups and a pronounced DDS ceiling effect. Because the FFQ reference period and frequency coding require fuller documentation, the DDS cannot be compared directly with short-reference dietary-diversity indicators. The results describe variety within the survey coding scheme, not micronutrient adequacy.

The high consumption prevalence across most food groups suggests broad dietary exposure within the reference period. However, the substantial differences in mean frequency demonstrate that high diversity did not necessarily imply balanced or equally frequent consumption. For instance, dairy was reportedly consumed by only 73.1% of respondents and recorded a mean score of 1.27, compared with mean scores of 13.09 for vegetables and 12.90 for roots and tubers. Thus, respondents could record high DDS values while consuming some nutrient-dense food groups relatively infrequently. Dietary-diversity scores should consequently be interpreted as indicators of dietary variety rather than direct measures of nutrient intake or nutritional adequacy [2].

The pronounced ceiling effect may partly reflect the seven-day reference period and the rule that a food group was counted once any corresponding food item was consumed at least once. This relatively inclusive criterion increases the likelihood that respondents will report exposure to most food groups, even when consumption is infrequent. Accordingly, the DDS was complemented with total food-frequency intensity, the NDFS and food-group-specific distributions. This combined interpretation provides greater analytical

discrimination than reliance on the food-group counts alone.

The comparison of candidate solutions shows that the three-cluster model produced the highest silhouette coefficient (0.1344), but its advantage over the two-cluster solution (0.1339) was negligible. Moreover, both coefficients are close to zero, indicating that many respondents were located near cluster boundaries and that the dietary profiles were not sharply separated. This interpretation is reinforced by the bootstrap adjusted Rand index, which averaged 0.579 but varied considerably across resamples (SD = 0.203; 95% bootstrap interval: 0.195–0.911). The solution therefore demonstrates moderate average reproducibility but substantial sampling sensitivity. Silhouette coefficients near zero generally indicate overlapping clusters, while variable adjusted Rand values suggest that membership assignments may change when the sample is perturbed []. Consequently, the three clusters should be treated as exploratory summaries of dietary heterogeneity rather than fixed, naturally occurring or clinically distinct nutritional groups.

The standardized centroids nevertheless reveal substantively interpretable differences in consumption patterns. Cluster 1 recorded below-average scores for all ten food groups, with particularly large negative deviations for vegetables, roots and tubers, beverages, dairy, and fats and oils, supporting its description as a low-frequency dietary profile. Cluster 2 showed above-average consumption of roots and tubers and vegetables but below-average scores for fruits, fish/seafood, meat, dairy and pulses, indicating a staple- and vegetable-oriented pattern with comparatively limited animal-source and protective-food frequencies. Conversely, Cluster 3 displayed above-average scores for cereals, pulses, fruits, meat, fish/seafood and dairy, and was therefore the most comparatively diversified profile. These labels describe relative positions within the present sample and should not be interpreted as evidence of nutritional adequacy, causal dietary effects or stable long-term eating behaviour, because cluster-derived dietary patterns may vary with the selected variables, scaling procedure, sample composition and observation period [15,16].

Table 5 shows that dairy was the most influential food-group variable in the Random Forest classifier, recording a mean absolute Shapley value of 0.243, substantially higher than fish/seafood (0.046), fruits (0.042), pulses/legumes (0.035) and cereals (0.029). This result indicates that variation in dairy-frequency scores contributed most strongly to distinguishing respondents classified within the upper-tertile NVI category. The comparatively lower values recorded for meat products, beverages, vegetables, fats and oils, and roots and tubers suggest that these variables exerted less influence on the fitted model's classifications. However, mean absolute Shapley values indicate only the magnitude of a variable's contribution and do not reveal its direction or establish causality; directional effects must therefore be interpreted alongside the beeswarm and dependence plots. Furthermore, dairy's prominence may partly reflect its inclusion in the nutrient-dense food-frequency component used to construct the NVI outcome [9].

Cronbach alpha, KMO, Bartlett and PCA results indicate correlation among DDS, TFFI and NDFS, which is expected because all three are transformations of the same food-frequency variables. These statistics do not establish external construct validity or clinical validity. The PCA-weight sensitivity analysis (Spearman rho = 0.997) shows only that rankings were nearly unchanged under the two weighting schemes examined.

Conclusion

This study examined food-group consumption patterns and sample-relative nutritional vulnerability among 216 university students and lecturers in Bayelsa State. Although respondents recorded a high mean Dietary Diversity Score, the pronounced ceiling effect demonstrated that food-group counts alone provided limited discrimination among dietary profiles. K-means analysis identified three exploratory dietary patterns; however, the low silhouette coefficient and moderate but variable bootstrap stability indicated substantial overlap among the clusters. The Random Forest classifier

demonstrated strong internal concordance with the upper-tertile NVI category, but its performance should not be interpreted as independent prediction because the outcome and predictors were derived from the same food-frequency variables.

The NVI should therefore be regarded as a transparent, equal-weighted and sample-relative analytical index rather than an AI-generated or clinically validated measure of nutritional status. Shapley analysis identified dairy as the most influential food-group variable, while the directional plots showed how different frequency scores shifted the model's predictions. Nevertheless, these contributions explain the behaviour of the fitted classifier and do not establish causal dietary effects. Consequently, the findings should not be generalised beyond the participating institutional population or applied to public-health decision-making without replication in a larger probability sample and external validation against anthropometric, biochemical, clinical, nutrient-intake or established food-security outcomes.

Recommendations

i. Strengthen analytical reproducibility: Future research should archive de-identified datasets, analytical code, preprocessing procedures, model specifications, hyperparameter settings, random seeds, uncertainty estimates, directional Shapley outputs and quantitative calibration statistics, subject to applicable ethical and confidentiality requirements.

ii. Improve sample representativeness: Subsequent studies should include respondents from broader demographic, occupational and geographical groups to determine whether the identified dietary patterns and model performance are reproducible beyond university students and lecturers.

Conflict of Interest

The authors declare that they have no known financial, personal, professional or institutional conflicts of interest that could have influenced the conduct, analysis, interpretation or reporting of this study.

Ethical Approval

The author submitted an application for ethical approval of the research proposal in February 2026, and formal approval was granted in May 2026. A copy of the ethical approval document is attached. Written informed consent was obtained from every participant before data collection. Recruitment was undertaken by trained field personnel who had no role in participants' teaching, assessment or academic progression. Potential participants were informed that participation was voluntary, refusal would have no academic or employment consequences, and they could withdraw at any stage without penalty. No lecturer recruited or administered the questionnaire to a student whom they directly taught or assessed. Responses were anonymised, treated confidentially and stored with access restricted to the research team.

Funding

The authors received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. The study was entirely self-funded by the authors.

Data Availability Statement

The datasets generated and analyzed during the current study are not publicly available because they contain respondent-level survey data. However, anonymized data and supporting analytical materials are available from the corresponding author upon reasonable request for academic and non-commercial research purposes, subject to applicable ethical and confidentiality considerations.

Acknowledgement

The authors sincerely appreciate all the respondents who voluntarily participated in this study and acknowledged the support of the field assistants whose contributions made the successful collection of the data possible.

Limitations

This study employed a cross-sectional design and an institution-based sample of 216 university students and lecturers; consequently, the findings cannot establish temporal or causal relationships and should not be generalized to the entire population of Bayelsa State. Furthermore, the study-specific DDS measured reported exposure to food groups during the seven-day reference period rather than portion sizes, quantities consumed, energy intake, nutrient adequacy or clinical nutritional status. Its pronounced ceiling effect also limited its ability to discriminate effectively among respondents with broadly diverse diets.

References

1. Arimond M, Wiesmann D, Becquey E, Carriquiry A, Daniels MC, Deitchler M, et al. Simple food group diversity indicators predict micronutrient adequacy of women's diets in 5 diverse, resource-poor settings. *J Nutr*. 2010;140(11):2059S-69S. <https://doi.org/10.3945/jn.110.123414>
2. Kennedy G, Ballard T, Dop M. Guidelines for measuring household and individual dietary diversity. FAO; 2011.
3. Ukonu IC, Wallace CA, Lowe NM. Household food security and dietary diversity in south-eastern Nigeria. *Matern Child Nutr*. 2024;20(3):e13599. <https://doi.org/10.1111/mcn.13599>
4. Chotwanvirat P, Prachansuwan A, Sridonpai P, Kriengsinyos W. Advancements in using AI for dietary assessment based on food images: scoping review. *J Med Internet Res*. 2024;26:e51432. <https://doi.org/10.2196/51432>
5. Sosa-Holwerda A, Park OH, Albracht-Schulte K, Niraula S, Thompson L, Oldewage-Theron W. The role of artificial intelligence in nutrition research: A scoping review. *Nutrients*. 2024;16(13):2066. <https://doi.org/10.3390/nu16132066>
6. Zheng J, Wang J, Shen J, An R. Artificial intelligence applications to measure food and nutrient intakes: scoping review. *J Med Internet Res*. 2024;26:e54557. <https://doi.org/10.2196/54557>
7. Breiman L. Random forests. *Machine learning*. 2001;45(1):5-32. <https://doi.org/10.1023/A:1010933404324>
8. Hartigan JA, Wong MA. Algorithm AS 136: A k-means clustering algorithm. *J R Stat Soc. series c (applied statistics)*. 1979;28(1):100-8. <https://doi.org/10.2307/2346830>
9. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, Katz R, Himmelfarb J, Bansal N, Lee SI. From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell*. 2020;2(1):56-67. <https://doi.org/10.1038/s42256-019-0138-9>
10. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math*. 1987;20:53-65. [https://doi.org/10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7)
11. Guidance WH. Ethics and governance of artificial intelligence for health. World Health Organization. 2021;1.
12. Swindale A, Bilinsky P. Household dietary diversity score (HDDS) for measurement of household food access: indicator guide. Washington, DC: Food and Nutrition Technical Assistance Project, Academy for Educational Development. 2006;220.
13. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. *the J Mach Learn Res*. 2011;12:2825-30.
14. Molnar C. Interpretable machine learning. 2020.
15. Sauvageot N, Schritz A, Leite S, Alkerwi AA, Stranges S, Zannad F, Streel S, Hoge A, Donneau AF, Albert A, Guillaume M. Stability-based validation of dietary patterns obtained by cluster analysis. *Nutr J*. 2017;16(1):4. <https://doi.org/10.1186/s12937-016-0223-9>
16. Walthouwer MJ, Oenema A, Soetens K, Lechner L, de Vries H. Are clusters of dietary patterns and cluster membership stable over time? Results of a

longitudinal cluster analysis study. *Appetite*. 2014;82:154-159.
<https://doi.org/10.1016/j.appet.2014.07.019>