Received: 25 July 2021 | Accepted: 11 January 2022

DOI: 10.1111/jhn.12992

ORIGINAL ARTICLE

BDA The Association of UK Dietitians THE OFFICIAL JOURNAL OF THE BRITISH DIETETIC ASSOCIATION Journal of Human Nutrition and Dietetics

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)

Vanderlei Carneiro Silva[1,2] | Bartira Gorgulho[3] | Dirce Maria Marchioni[4] | Tânia Aparecida de Araujo[1] | Itamar de Souza Santos[2] | Paulo Andrade Lotufo[2] | Isabela Martins Benseñor[2]

[1]Department of Epidemiology, School of Public Health, University of São Paulo, São Paulo, Brazil

[2]Center of Clinical and Epidemiological Research, University Hospital, University of São Paulo, São Paulo, Brazil

[3]Department of Food and Nutrition, School of Nutrition, Federal University of Mato Grosso, Cuiaba, Brazil

[4]Department of Nutrition, School of Public Health, University of São Paulo, São Paulo, Brazil

Correspondence
Vanderlei Carneiro Silva, Center of Clinical and Epidemiological Research, University Hospital, University of São Paulo, Av. Lineu Prestes 2565, 3rd Floor, São Paulo 05508 000, Brazil.
Email:vnd.cs@hotmail.com

Abstract

Background: Machine learning investigates how computers can automatically learn. The present study aimed to predict dietary patterns and compare algorithm performance in making predictions of dietary patterns.

Methods: We analysed the data of public employees ($n = 12{,}667$) participating in the Brazilian Longitudinal Study of Adult Health (ELSA-Brasil). The *K*-means clustering algorithm and six other classifiers (support vector machines, naïve Bayes, *K*-nearest neighbours, decision tree, random forest and xgboost) were used to predict the dietary patterns.

Results: *K*-means clustering identified two dietary patterns. Cluster 1, labelled the Western pattern, was characterised by a higher energy intake and consumption of refined cereals, beans and other legumes, tubers, pasta, processed and red meats, high-fat milk and dairy products, and sugary beverages; Cluster 2, labelled the Prudent pattern, was characterised by higher intakes of fruit, vegetables, whole cereals, white meats, and milk and reduced-fat milk derivatives. The most important predictors were age, sex, per capita income, education level and physical activity. The accuracy of the models varied from moderate to good (69%–72%).

Conclusions: The performance of the algorithms in dietary pattern prediction was similar, and the models presented may provide support in screener tasks and guide health professionals in the analysis of dietary data.

KEYWORDS

classification algorithms, clustering analysis, dietary patterns, machine learning

Key points

- Machine learning (ML) investigates how computers can automatically learn. The present study aimed to predict dietary patterns and to compare the performance of various ML algorithms for making the predictions of dietary patterns.
- *K*-means clustering identified two major dietary patterns. The models presented may provide support in screener tasks.

INTRODUCTION

Diet has been shown to play a fundamental role in the prevention of chronic diseases, and a healthy lifestyle can help reduce the risk of diseases.[1] Many dietary assessment methods attempt to estimate total food and nutrient intake, although this detailed estimation may lead to the collection of a large amount of data, which can be complex and difficult to summarise into dietary patterns.[2] In addition, dietary intake data are not always available, and the assessment methods require structure and preparation from interviewers, especially in large samples.[3,4]

Machine learning (ML) is a fast-growing discipline that investigates how computers can automatically learn to recognise complex patterns and make intelligent data-based decisions.[5] This area comprises numerous different types of algorithms that can process large amounts of data, such as nutrition information, and ultimately transform data into knowledge.[6] The approach to dietary patterns allows us to evaluate diets overall instead of food or nutrients alone.[7] Cluster analysis is a method applied to characterise dietary patterns.[8,9] Cluster analysis aggregates individuals with similar dietary patterns into mutually exclusive categories according to the means of food intake variables, such as the frequency of food consumed, the percentage of energy contributed by each food or the average grams of food intake.[10]

Socio-demographic and clinical data are basic information always collected in healthcare and epidemiological surveys. Associations between dietary patterns and socioeconomic and demographic characteristics have been previously observed in low- and middle-income countries and high-income countries.[11–13] It is widely understood that these characteristics are associated with dietary patterns and can influence food choices.[14,15] ML could be used to analyse these data as a complementary method for screening and guiding healthcare professionals when there is a need to identify dietary patterns and their associated factors. Few studies have examined the predictive accuracy of alternative ML methodologies in predicting dietary patterns based on self-reported dietary intake or even considered socio-demographic and health data as associated factors.[6,16–18]

Dietary assessment can be used to guide public health policies and population interventions.[19] Moreover, when resources and time for data collection are limited and a less detailed assessment fails to meet the research objectives, faster and more objective tools can be useful.[20] The present study aimed to predict dietary patterns and compare the performance of various ML algorithms in making predictions of dietary patterns. We test the hypothesis that the patterns derived by a clustering analysis can be predicted from the socio-demographic and clinical data sets of the public employees participating in the Brazilian Longitudinal Study of Adult Health (ELSA-Brasil).

METHODS

Participant recruitment

The ELSA-Brasil is a multicentre cohort study involving 15,105 active and retired civil servants aged 35–74 years from teaching and research institutes in the following six Brazilian cities: Salvador, Belo Horizonte, Vitória, Rio de Janeiro, São Paulo and Porto Alegre**.** The ELSA-Brasil study was designed to investigate the incidence of cardiovascular diseases and diabetes and their biological and social determinants. The present study included cross-sectional data from the baseline examination, which was carried out between August 2008 and December 2010.[21,22] Active or retired employees aged 35–74 years who answered the Food Frequency Questionnaire (FFQ**)** were eligible to participate. The exclusion criteria were: intention to leave the institution, current or recent pregnancy within the prior 4 months, severe cognitive or communication difficulty and, if retired, residence outside the research centre metropolitan region. After recruitment, the participants were interviewed at the work facility (Phase 1) and scheduled a date to visit the research centre to undergo several examinations, such as anthropometrics, blood pressure (BP) measurement, electrocardiogram (Phase 2).

Dietary assessment

A semiquantitative FFQ developed for the ELSA-Brasil study was used. The FFQ presents a list of 114 food items and was based on a previously validated questionnaire.[23] The subjects were asked to estimate how often, on average, they consumed a standardised portion of a given food item in the 12-month period preceding the interview. Nutrition Data System for Research (NDSR) software (University of Minnesota, 2010) was used to determine the nutritional composition of the foods and preparations and daily energy intake in kilocalories.[24] The daily intake was quantified by the number of servings consumed per day × weight (standard portion in grams) × frequency of consumption × nutritional composition of the food serving. The details of the elaboration[23] and validation of the questionnaire[25] are as described in the previous publications. Of the total sample, we excluded $n = 2438$ (16%) participants with an implausible energy intake of less than 500 or greater than 4000 kcal day^{-1}.[26] The final sample comprised 12,667 public employees, including 5217 (41%) men and 7450 (59%) women.

Socio-demographic and clinical predictors

The protocol of the ELSA-Brasil study included clinical tests and interviews, which required volunteers to visit a clinical research centre.[22] The following sociodemographic and clinical data were collected and included in the analysis: sex, age (years), education level (elementary [or less], high school, or college), retirement status (no vs. yes), self-reported race/ethnicity (White, Brown, Black or other [Asian or indigenous]), per capita income in US$ (categorised in terciles [using US$ 1.00 = R$ 2.00 as the approximate baseline examination exchange rate]), living alone or with another person (with another person vs. alone), marital status (not married vs. married), smoking habit (never, ex-smoker or current smoker), physical activity (based on the leisure time section of the long version of the International Physical Activity Questionnaire), health self-assessment (good, regular or bad) and location of the research centre (Salvador, Belo Horizonte, Vitória, Rio de Janeiro, São Paulo or Porto Alegre).

The weight and height measurements were performed with the participants wearing light clothes and no shoes. We measured body weight to the nearest 0.1 kg with a calibrated scale (Toledo 2096PP) and height with a vertical audiometer (Seca-SE-216) to the nearest 0.1 cm. The body mass index (BMI) was calculated by dividing weight in kilograms by height in metres squared (kg/m^2). The waist circumference was measured with a tape measure to the nearest 0.1 cm around the midpoint between the inferior costal border and the iliac crest, whereas the hip circumference was measured at the point of greatest circumference in the gluteal region. The waist-to-hip ratio (WHR) was calculated by dividing the waist circumference by the hip circumference in centimetres.

BP was measured using a validated Omron HEM 705CPINT oscillometer device. Three measurements were performed at 1-min intervals, and the mean of the two latter BP measurements was the value used to define hypertension, which was defined as systolic BP ≥ 140 mmHg, diastolic BP ≥ 90 mmHg or verified treatment with antihypertensive drugs during the previous 2 weeks. Dyslipidaemia was defined as low-density lipoprotein cholesterol ≥ 130 mg/dL or the use of medication to treat dyslipidaemia. Diabetes was defined as a previous diagnosis of diabetes, the use of medication to treat diabetes, fasting plasma glucose ≥ 126 mg/dL, 2-h plasma glucose ≥200 mg/dL or HbA1c ≥ 6.5%. Cardiovascular disease was defined as self-reported prior myocardial infarction, stroke or revascularisation.

Statistical analysis

The continuous variables are presented as medians and interquartile ranges, and the categorical variables are presented as frequencies. All analyses were performed in R, version 4.0.2 (R Foundation for Statistica Computing). The associations between the categorical variables were tested usinf chi-squared tests. The comparisons of the values of the continuous variables by dietary pattern (i.e., Western or Prudent) were performed using a Mann–Whitney test.

ML algorithms

The statistical packages of R software were employed to implement the models. *K*-means clustering algorithm was used to identify the dietary patterns. Then, six different classifier algorithms (support vector machine [SVM], naïve Bayes [NB], K-nearest neighbours [KNN], decision tree [DT], random forest [RF] and xgboost) were used to predict the dietary patterns of each participant. The predictors used by the classifier algorithms were ranked according to the importance of their influence on the prediction of the dietary patterns. The order of importance was based on the varImp function (caret package), importance function of a RF model (randomForest package) and xgboost model (EIX package); all analyses were performed using the R software. We calculated the agreement between the patterns predicted and the classification scheme provided by the cluster algorithm.[27,28] The accuracy was determined using confusionMatrix (e1071 package); in new individuals for whom the dietary pattern label was not known by the algorithm, only information regarding the sociodemographic and clinical data was used. In general, values equal to 0.5 correspond to the performance of a random classifier, values less than 0.6 (and greater than 0.5) indicate moderate predictive performance and values greater than 0.7 indicate good predictive performance.[16,28–30] Furthermore, other evaluation metrics, such as the sensitivity, specificity, and positive and negative predictive values, were calculated. In the following, the functionality of each ML algorithm is briefly described.

Clustering analysis

The *K*-means clustering algorithm was used to divide the participants into groups based on their dietary intake data. The following packages were installed and executed: cluster and factoextra of R, version 4.0.2. *K*-means clustering is one of the most popular algorithms.[31–33] This method of clustering partitions the data; thus, each instance is placed in a cluster, and there is no hierarchical relationship among the *K* clusters.[8,34] The frequencies of food intake were converted to z-scores and input into the algorithm. The clustering distance measurements were carried out using Euclidean distances. We retained two dietary patterns considering homogeneity in the derived groups, the balance between the classes and the range of groups previously found in the literature.[8,9] Factor interpretability was examined to confirm the final number of dietary patterns and whether a group was sufficiently large for an adequate statistical power, that is, at least 10% of the total sample.[10] The class imbalance problem is closely related to cost-sensitive learning.[5]

Then, the original database (n = 12,667) was randomly divided into two subsets using the R function sample. The first subset was used for training, and the second subset was used for testing, with sample sizes of 8866 (70%) and 3801 (30%), respectively. The training stage of the classifier algorithms was based on the following sociodemographic and clinical data: sex, age, education level, retirement status, race/ethnicity, marital status, per capita income, living alone, location of the research centre, smoking habit, physical activity, health self-assessment, BMI, WHR, dyslipidaemia, hypertension, diabetes and cardiovascular disease. During the training stage, we used 10-fold cross-validation, and this method was used to adjust the hyperparameters of the models. Cross-validation (*K*-fold cross validation) is a sampling method used to analyse the performance of ML algorithms. Cross-validation consists of randomly dividing the data into mutually exclusive *K* folds of equal sizes.[35]

SVMs

SVM classifiers operate by separating classes using linear decision boundaries called hyperplanes.[36] The model is a representation of mapped points such that the examples of each category are divided by a clear and well-defined space.[37] The SVM classifier is based on the theory of statistical learning and kernel methods and is a linear, binary, non-probabilistic classifier. Despite this characteristic, this classifier is also applied to regression problems and non-linearly separable data.[38] To apply this approach to non-linearly separable instances, a transformation function is typically employed to map the data to a space that enables linear decisions (usually of a higher dimensionality).[29] The following parameters determine the performance of this algorithm: radial kernel; cost = 10; and gamma = 0.01. The following packages were used: e1071 and kernlab.

NB

The NB algorithm (e1071 package) is extremely simple but powerful. NB is based on Bayes' theorem and aims to calculate the probability that an unknown sample belongs to a certain class; thus, this algorithm predicts the most likely class.[39] In addition, the classification is called naïve because it considers that the effect of an attribute on the occurrence of a class is independent of the presence or absence of any other attribute. This algorithm is useful in large databases and can yield results superior to other more sophisticated techniques.[5]

KNN

The KNN algorithm (caret package) predicts an unknown entry based on the label of the training examples of the closest neighbours in the characteristics space. The determination of the label of an unknown example is usually based on the majority vote of the *K* closest neighbours. For its use, a training data set is necessary to define the distance calculation metric between the data and the number of *K* closest neighbours that will be used by the algorithm at the time of classification.[39] The distance from the data point to its neighbours in the training data set was calculated using the Euclidean distance measure. During the classification stage, accuracy was used to select the optimal model using the largest value, and the final value used for the model was *k* neighbours = 39.

DT

The DT algorithm has a flowchart-like tree structure, where each internal node (non-leaf node) denotes a test of an attribute, each branch represents an outcome of the test and each leaf node (or terminal node) holds a class label.[5] The structure refers to an upside-down tree, with the roots at the top and the leaves at the bottom. The DT algorithm is used as a decision support tool that can naturally induce rules and, for each rule, a decision needs to be made. By contrast to the other predictive algorithms, the results allow the identification of a set of well-defined rules. The construction of a DT is appropriate for exploratory knowledge discovery, and the algorithm can address multidimensional data.[40] In general, DT classifiers achieve good accuracy. In this analysis, the rpart algorithm and caret package were used. The use of this algorithm (e1071 package) was performed by the value of the following two parameters: (1) the class method (needed to predict the classes of the dietary patterns) and (2) the minimum number of observations that must exist in a node (we used the value = 20).

Ensemble algorithms

Ensemble methods are known to perform better than other algorithms in numerous ML applications.[41] The RF algorithm is based on the ensemble strategy. This algorithm generates several decision trees, and each tree is trained with a random distribution. A major advantage of the RF is the ease of measuring the relative importance of each attribute for the prediction by analysing how many nodes in the trees use a given attribute to reduce the overall impurity of the forest.[35] The RF model was built through the randomForest function, which is a part of the package with the same name. The analysis was performed using the following established basic parameters: number of trees (ntrees) = 500, minimum size of terminal nodes (nodesize) = 5 and number of variables used in each tree (mtry) = 6. The importance of the variables is also determined by the algorithm as a result of the average of the reduction in the accuracy of the prediction as one variable is removed from the model and the other variables are included.[42] We also used the eXtreme Gradient Boosting package (XGBoost), which is an efficient and popular implementation of gradient boosted decision trees. XGBoost strictly prioritises computational speed and model performance with good accuracy using most data sets.[43] To identify the interactions among the variables, the Explain Interactions in the XGBoost (EIX) package was used. We created a ranking of the interactions using the function importance with the parameter option 'interactions'. Plots were created using the package ggplot2. Package EIX uses a table, which was generated by the xgb model based on information concerning their trees, their nodes and leaves. EIX considers pairs of variables in the model, where the variable on the bottom (child) has a higher gain than the variable on the top (parent).

RESULTS

From the original data set ($n = 12{,}667$), the following two major dietary patterns were derived: one pattern with 7157 (57%) participants and another pattern with 5510 participants (43%). Subdivisions with three or more patterns did not substantially improve intragroup homogeneity.

Table 1 shows the characteristics of the study sample by dietary pattern. The Prudent dietary pattern group had a higher proportion of women and older people (median of 54 years vs. 50 years) and was characterised by a higher level of education, white individuals, single people, living alone, higher income, retirees, nonsmokers and physically active people than the Western dietary pattern group ($p < 0.001$). Among the health characteristics, the group following the Prudent dietary pattern included a higher proportion of people with chronic diseases (dyslipidaemia, diabetes and cardiovascular diseases) who self-perceived their health as good.

Table 2 shows the dietary intakes by pattern. The participants in the Prudent pattern group presented a higher consumption of fruit, vegetables, whole cereals, white meats, lower-fat dairy products and milk. The subjects in the Western pattern group presented a higher energy intake and mean consumption of refined cereals, beans and other legumes, tubers, pasta, processed and red meats, eggs, high-fat dairy products and milk, salted snacks, and sugary beverages.

Table 3 shows the performance of the models. Using only the socio-demographic and clinical data to predict the dietary pattern in the test set, the accuracy of the SVM, NB, KNN and DT classifiers was 0.71, 0.70, 0.69 and 0.70, respectively. The accuracies of the ensemble-type models (RF and XGBoost) were slightly higher than that of the previously mentioned models. Furthermore, the sensitivity, specificity, and positive and negative predictive values are presented; each metric measures the classification ability related to one of the two dietary patterns.

Table 4 shows the predictors ranked by each algorithm. All variables were used during the initial training stage; however, the variables are presented in descending order by the level of importance. The most common selected features are sex, age, education level, per capita income and physical activity. By contrast, BMI, diabetes, hypertension, cardiovascular disease and location of the research centre were the least important. In our complementary analysis, the importance function was used to extract the predictors in the order of their importance in the ensemble models. The Supporting information (Figures S1 and S2) confirms that the demographic data (age, income, sex and education) are among the most associated with the dietary patterns.

The Supporting information (Figure S3) provides a plot that considers pairs of variables in the model, where the variable on the bottom (child) has a higher gain than the variable on the top (parent). This figure is a matrix plot with the colour of the square at the intersection of two variables indicating the value of the *sumGain* measure (sum of the gain value in all nodes in which a given variable occurs). Among the analysed predictors, there is an emphasis on the interaction between BMI and WHR in addition to the research centres São Paulo and Belo Horizonte.

The Supporting information (Figure S4) also provides the performance of the XGBoost model based on the number of decision trees, indicating a stability of approximately 200 trees. Log loss is the most important classification metric based on probabilities. For any given problem, a lower log-loss value implies better predictions.

DISCUSSION

Two major dietary patterns were identified in our sample, and ML algorithms were trained to predict these patterns using only the socio-demographic and clinical characteristics of the sample data. The Prudent pattern was characterised by higher intakes of whole cereals, fruit, vegetables, white meats, milk and reduced-fat milk derivatives. The Western pattern was characterised by higher intakes of refined cereals, beans, processed and red meats, eggs, high-fat dairy products and milk, salted snacks, and sugary beverages. Confirming our hypothesis, after the training stage, all algorithms were able to classify individuals into a dietary pattern based on their socio-demographic and clinical characteristics with moderate-to-good accuracy (69%–72%).

Our results, which are presented in Table 1, confirm that differences exist between the two identified patterns and that the algorithms used can classify individuals based on the present features. However, although people receiving treatment for diet-related chronic diseases are

TABLE 1 Characteristics of the study population, Longitudinal Study of Adult Health (ELSA-Brasil), 2008–2010

	General		Western		Prudent	
Variable	***n***	**%**	***n***	**%**	***n***	**%**
Study population	12,667	100.0	7157	100.0	5510	100.0
Data set						
Training	8866	70.0	5042	70.5	3824	69.4
Test	3801	30.0	2115	29.5	1686	30.6
Sex						
Male	5217	41.2	3543	49.5**	1674	30.4
Female	7450	58.8	3614	50.5	3836	69.6
Age (years)[a]	52	45–59	50	44–56**	54	35–61
Education level						
Elementary (or less)	1423	11.2	1128	15.8**	295	5.4
High school	4072	32.2	2876	40.2	1196	21.7
College	7172	56.6	3153	44.0	4019	72.9
Retirement status						
No	10,046	79.3	6064	84.7**	3982	72.3
Yes	2621	20.7	1093	15.3	1528	27.7
Race/ethnicity						
White	6994	55.2	3424	47.8**	3570	64.8
Mixed	3379	26.7	2240	31.3	1139	20.7
Black	1831	14.4	1245	17.4	586	10.6
Others[b]	463	3.7	248	3.5	215	3.9
Marital status						
Single	4486	35.4	2286	31.9**	2200	39.9
Married	8181	64.6	4871	68.1	3310	60.1
Per capita income (US$)						
1° tercile	4225	33.4	3281	45.8**	944	17.1
2° tercile	4492	35.5	2441	34.1	2051	37.2
3° tercile	3950	31.1	1435	20.1	2515	45.7
Living alone						
With another person	11,043	87.2	6471	90.4**	4572	83.0
Alone	1624	12.8	686	9.6	938	17.0
Smoking habit						
Never	7306	57.7	3946	55.1**	3360	61.0
Ex-smoker	3780	29.8	2049	28.6	1731	31.4
Current smoker	1581	12.5	1162	16.3	419	7.6
Physical activity[c]						
Sedentary	5798	45.8	3955	55.3**	1843	33.5
Insufficiently active	3354	26.5	1776	24.8	1578	28.6
Active	3515	27.7	1426	19.9	2089	37.9

TABLE 1 (Continued)

	General		Western		Prudent	
Variable	*n*	%	*n*	%	*n*	%
Self-assessment health						
Good	10,266	81.1	5650	78.9**	4616	83.8
Regular	2167	17.1	1367	19.1	800	14.5
Bad	234	1.8	140	2.0	94	1.7
BMI (kg/m^2)[a]	26.3	23.7-29.5	26.4	23.7-29.6	26.2	23.6-29.4
Waist-to-hip ratio[a]	0.9	0.8-1.0	0.9	0.8-1.0**	0.9	0.8-0.9
Dyslipidaemia[d]						
No	5237	41.3	3206	44.8**	2031	36.9
Yes	7430	58.7	3951	55.2	3479	63.1
Hypertension[e]						
No	8159	64.4	4584	64.1	3575	64.9
Yes	4508	35.6	2573	35.9	1935	35.1
Diabetes[f]						
No	10,634	83.9	6082	85.0**	4552	82.6
Yes	2033	16.1	1075	15.0	958	17.4
Cardiovascular disease[g]						
No	12,188	96.2	6922	96.7*	5266	95.6
Yes	479	3.8	235	3.3	244	4.4

Note: *p* values are derived from Mann–Whitney-tests or chi-squared tests.

Abbreviation: BMI, body mass index.

[a] Median and interquartile range.

[b] Others = Asian + indigenous.

[c] Sedentary: does not perform physical activity; insufficiently active: < 150 min/week[1] or exercise less than 3 days a week; active: 150 min/week[1] at least 3 days a week.

[d] LDL ≥ 130 mg/dL[1] or the use of cholesterol reducers.

[e] Systolic blood pressure ≥ 140 mmHg, diastolic blood pressure ≥ 90 mmHg or verified treatment with antihypertensive drugs during the previous 2 weeks.

[f] Defined as an account of a previous diagnosis of diabetes, the use of medication for diabetes or meeting the diagnostic value of diabetes.

[g] Defined as a report of a heart attack, stroke or revascularisation.

*$p < 0.05$; **$p < 0.001$.

more likely to be exposed to dietary advice and health messages, health and clinical data, such as health self-assessment, dyslipidaemia, hypertension and cardiovascular disease, had no significant influence on the performance of the algorithms. In this analysis, the socio-demographic data were more related to the dietary pattern than the profile of comorbidities in the sample.

Comparisons of our findings with those of previous studies are limited by the scarcity of studies predicting dietary patterns using ML algorithms. Panaretos et al.[6] used two techniques (the KNN and RF algorithms) to evaluate participants' health based on the dietary information of 3042 men and women (45 ± 14 years old) who were enrolled in the ATTICA study. In that study, the ML techniques were superior to a linear regression in the correct classification of the individuals according to the health score (accuracy of approximately 38% vs. 6%, respectively). Pencina et al.[16] evaluated the performance of Fisher's discriminant functions in identifying the dietary patterns of women (*n* = 1828) and men (*n* = 1666) aged 18–76 years, who were participants of the Framingham Nutrition Studies. The model correctly classified approximately 80% of the participants. Our results were less accurate, although the predictions were made without any information regarding diet. Among the potential predictors, sex, age, education level, per capita income and physical activity were the most important features in the final models.

Our study fills the following important gap in the literature: the prediction of dietary patterns based on demographic, social and health factors; such predictions open the door for interventions based on not only the components of food consumption but also the

TABLE 2 Food consumption by dietary pattern, Brazilian Longitudinal Study of Adult Health (ELSA-Brasil), 2008–2010

	Western pattern		Prudent pattern	
Food groups (g or mL day^{-1})	Median	IQR	Median	IQR
Energy[a]	2601.6**	2081.2–3128.8	2452.1	1989.9–2985.7
Refined cereals	150.0**	100.0–250.0	50.0	5.0–100.0
Whole cereals	6.7**	1.1–24.6	48.0	16.2–106.7
Beans and other legumes	140.0**	70.0–350.0	78.4	33.6–151.2
Fruit	321.7**	181.1–537.2	559.8	367.0–816.3
Vegetables	160.2**	104.1–231.5	277.9	185.6–404.0
Tubers	48.6**	27.1–85.3	38.4	20.4–71.6
Red meats	71.0**	46.3–112.0	51.9	25.3–88.9
White meats	88.3**	48.6–143.8	116.5	72.0–198.0
Eggs	7.0**	3.5–13.7	6.7	3.2–10.0
Processed meats	15.6**	6.9–28.1	11.9	3.4–25.4
Pasta	22.0**	11.0–33.0	17.0	11.0–29.1
Salted snacks	25.9**	14.0–43.4	21.8	10.5–37.7
High-fat dairy products and milk	81.0**	14.7–252.0	20.1	2.52–96.0
Lower-fat dairy products and milk	12.0**	2.1–94.5	185.3	48.0–366.4
Sugary beverages	120.0**	33.6–264.0	60.0	16.8–208.8

Note: *p* values are derived from a Mann–Whitney test.

Abbreviation: IQR, interquartile range.

[a]kcal day^{-1}.

**$p < 0.001$.

TABLE 3 Performance measures of the machine learning algorithms

	SVM	NB	KNN	DT	RF	XGBoost
Accuracy	0.71	0.70	0.69	0.70	0.72	0.72
95% CI	(0.69–0.72)	(0.69–0.72)	(0.68–0.71)	(0.69–0.72)	(0.71–0.74)	(0.71–0.73)
Sensibility	0.62	0.66	0.74	0.63	0.65	0.68
Specificity	0.77	0.74	0.64	0.76	0.78	0.75
Positive predictive value	0.69	0.67	0.72	0.68	0.70	0.67
Negative predictive value	0.72	0.73	0.65	0.72	0.74	0.76

Abbreviations: CI, confidence interval; DT, decision trees; KNN, *K*-nearest neighbours; NB, naïve Bayes; RF, random forest; SVM, support vector machine; XGBoost, eXtreme gradient boosting.

determinants in the population related to diet. The algorithms used can classify individuals into subgroups by predicting the most common dietary patterns and present features associated with specific patterns in the population. Four predictive algorithms (SVM, NB, DT and KNN) were explored, and the results obtained were similar. Among the ensemble-type models (RF and xgboost), the accuracy was slightly higher than that of the models previously mentioned. This finding demonstrates that all classifiers used can be helpful in predicting dietary patterns and identifying the main determinants of diet.

ML and other tools supported by technology do not replace existing dietary assessment methods but can be used as complementary approaches.[44,45] The proposed algorithms can be used as a screening tool in the field of nutritional epidemiology. However, similar to all other methods, they have certain limitations: they rely on expert technology teams to deploy the models, the presence of computational structures in health systems and the

TABLE 4 Feature selection by the machine learning algorithms

	Algorithm							
Order	SVM		DT		KNN		NB	
1	Per capita income	0.69	Sex	1.00	Per capita income	0.68	Per capita income	0.68
2	Education level	0.66	Per capita income	1.00	Education level	0.65	Education level	0.65
3	Physical activity	0.63	Physical activity	0.75	Physical activity	0.63	Physical activity	0.63
4	Age	0.61	Education level	0.72	Age	0.61	Sex	0.62
5	Sex	0.59	Age	0.62	Sex	0.60	Age	0.60
6	Race/ethnicity	0.58	Smoking habit	0.25	Race/ethnicity	0.58	Race/ethnicity	0.58
7	Waist-to-hip ratio	0.57	Location of the research centre	0.23	Waist-to-hip ratio	0.57	Waist-to-hip ratio	0.57
8	Active worker	0.56	Diabetes	0.20	Active worker	0.56	Active worker	0.56
9	Smoking habit	0.54	Race/ethnicity	0.18	Smoking habit	0.55	Smoking habit	0.54
10	Marital status	0.54	Active worker	0.12	Marital status	0.54	Marital status	0.54
11	Living alone	0.54	Cardiovascular disease	0.09	Dyslipidaemia	0.54	Dyslipidaemia	0.54
12	Dyslipidaemia	0.53	Living alone	0.09	Living alone	0.54	Live alone	0.54
13	Health self-assessment	0.52	Health self-assessment	0.07	Health self-assessment	0.53	Health self-assessment	0.52
14	Location of the research centre	0.51	Hypertension	0.03	Hypertension	0.51	Diabetes	0.51
15	Diabetes	0.51	Waist-to-hip ratio	0.02	Location of the research centre	0.51	Location of the research centre	0.51
16	Cardiovascular disease	0.51	Marital status	0.02	Diabetes	0.51	Body mass index	0.51
17	Hypertension	0.50	Body mass index	0.02	Body mass index	0.51	Cardiovascular disease	0.51
18	Body mass index	0.50	Dyslipidaemia	0.01	Cardiovascular disease	0.50	Hypertension	0.50

Abbreviations: DT, decision trees; KNN, *K*-nearest neighbours; NB, naïve Bayes; SVM, support vector machines.

training of healthcare professionals regarding how to correctly interpret the results of the models. Therefore, the use of combined tools is recommended.[3]

Some limitations should also be addressed. The present study adopted a cross-sectional analysis and, although the features used can predict dietary patterns, we did not assess causality. The collected dietary data were self-reported and are subject to the interviewees' memory biases. Several subjective decisions were made in the process of the analysis, such as the definition and collapse of some food groups, the retention of the number of groups and the labelling of the identified dietary patterns.

This study has some strengths. The data analysed were obtained from a large and multicentre sample of adult and elderly individuals. Although the sample consisted of only civil servants, it aggregated an admixed, multiethnic population recruited from six major centres and captured non-isolated eating practices. In the present study, foods associated with a Western diet represented 57% of the analysed sample. Therefore, some generalisability of these results to the general population living in the metropolitan areas of the country is possible. The FFQ used for the data collection was developed and validated in the study population. The participants were invited to attend a clinical research centre for examinations and clinical evaluations, which guaranteed a high standard of quality control in the predictors used in the study.

ACKNOWLEDGEMENTS

We thank all participants of the ELSA-Brasil study. The ELSA-Brasil study was supported and received funding from the Ministry of Health of Brazil, Ministry of Science, Technology and Innovation and National Development Council Scientific and Technological Advice (grants 01 06 0010.00 RS; 01 06 0212.00 BA; 01 06 0300.00 ES; 01 06 0278.00 MG; 01 06 0115.00 SP and 01 06 0071.00 RJ). However, this subproject did not receive funding. The funding received at the study baseline had no influence on the design, analysis, drafting, interpretation or decision of the version submitted for publication.

CONFLICT OF INTERESTS

The authors declare that there are no conflict of interests.

ETHICAL STATEMENT

This study was performed according to the guidelines suggested by the Declaration of Helsinki, and the study protocol was reviewed and approved by the Ethics Committee of the School of Public Health of the University of São Paulo under number 2.566.286. After explaining the purpose of the survey to the participants, informed consent was obtained from the study participants willing to participate in the study.

AUTHOR CONTRIBUTIONS

All authors participated in the conception and design of the study and the analysis and interpretation of the data. Vanderlei Carneiro Silva and Isabela Martins Benseñor contributed to the design, acquisition, analysis and interpretation of the data, as well as the drafting and revision of the text. Bartira Gorgulho, Dirce Maria Marchioni and Tânia Aparecida de Araujo contributed to the design, analysis, and revision of the text. Itamar de Souza Santos and Paulo Andrade Lotufo contributed to the acquisition, analysis and revision of the text. All authors critically reviewed, edited and approved the final manuscript submitted for publication.

TRANSPARENCY DECLARATION

The lead author affirms that this manuscript is an honest, accurate and transparent account of the study being reported. The reporting of this work is compliant with the STROBE guidelines. The lead author affirms that no important aspects of the study were omitted and that any discrepancies from the study as planned were explained.

ORCID

Vanderlei Carneiro Silva http://orcid.org/0000-0003-1595-1880

Bartira Gorgulho http://orcid.org/0000-0002-1714-3548

Dirce Maria Marchioni http://orcid.org/0000-0002-6810-5779

Tânia Aparecida de Araujo http://orcid.org/0000-0001-5894-8695

Itamar de Souza Santos http://orcid.org/0000-0003-3212-8466

Paulo Andrade Lotufo http://orcid.org/0000-0002-4856-8450

Isabela Martins Benseñor http://orcid.org/0000-0002-6723-5678

PEER REVIEW

The peer history review for this article is available at https://publons.com/publon/10.1111/jhn.12992

REFERENCES

1. World Health Organization (WHO). Sustainable healthy diets. Guiding Principles. Rome: Food and Agriculture Organization of the United Nations (FAO) and WHO; 2019. p. 37.
2. Giabbanelli PJ, Adams J. Identifying small groups of foods that can predict achievement of key dietary recommendations: data mining of the UK National Diet and Nutrition Survey, 2008-12. Public Health Nutr. 2016;19:1543–51.
3. Shim J-S, Oh K, Kim HC. Dietary assessment methods in epidemiologic studies. Epidemiol Health. 2014;36:e2014009.
4. Neuhouser ML, Patterson RE, Kristal AR, Eldridge AL, Vizenor NC. A brief dietary assessment instrument for assessing target foods, nutrients and eating patterns. Public Health Nutr. 2001;4:73–8.
5. Han J, Kamber M, Pei J. Data mining: concepts and techniques. 3rd ed. Burlington, MA: Morgan Kaufmann; 2011. p. 703.
6. Panaretos D, Koloverou E, Dimopoulos AC, Kouli GM, Vamvakari M, Tzavelas G, et al. A comparison of statistical and machine-learning techniques in evaluating the association between dietary patterns and 10-year cardiometabolic risk (2002-2012): the ATTICA study. Br J Nutr. 2018;120:326–34.
7. Muga MA, Owili PO, Hsu C-Y, Rau H-H, Chao JC-J. Association between dietary patterns and cardiovascular risk factors among middle-aged and elderly adults in Taiwan: a population-based study from 2003 to 2012. PLoS One.2016;11:1–18.
8. Newby PK, Tucker KL. Empirically derived eating patterns using factor or cluster analysis: a review. Nutr Rev. 2004;62:177–203.
9. Devlin UM, Mcnulty BA, Nugent AP, Gibney MJ. The use of cluster analysis to derive dietary patterns: methodological considerations, reproducibility, validity and the effect of energy misreporting. Proc Nutr Soc. 2012;71:599–609.
10. Thorpe MG, Milte CM, Crawford D, McNaughton SA. A comparison of the dietary patterns derived by principal component analysis and cluster analysis in older Australians. Int J Behav Nutr Phys Act. 2016;13:13.
11. Mayén A-L, Marques-Vidal P, Paccaud F, Bovet P, Stringhini S. Socioeconomic determinants of dietary patterns in low- and middle-income countries: a systematic review. Am J Clin Nutr. 2014;100:1520–31.
12. Arruda SPM, Silva AAM, da,Kac G, Goldani MZ, Bettiol H, Barbieri MA. Socioeconomic and demographic factors are associated with dietary patterns in a cohort of young Brazilian adults. BMC Public Health. 2014;14:1–13.
13. Mayén A-L, Bovet P, Marti-Soler H, Viswanathan B, Gedeon J, Paccaud F, et al. Socioeconomic differences in dietary patterns in an east African country: evidence from the Republic of Seychelles. PLoS One. 2016;11:e0155617.
14. Krieger JP, Pestoni G, Cabaset S, Brombach C, Sych J, Schader C, et al. Dietary patterns and their sociodemographic and lifestyle determinants in Switzerland: results from the national nutrition survey menuCH. Nutrients. 2019;11:1–16.
15. Baraldi LG, Martinez Steele E, Canella DS, Monteiro CA. Consumption of ultra-processed foods and associated sociodemographic factors in the USA between 2007 and 2012: evidence from a nationally representative cross-sectional study. BMJ Open. 2018;8:e020574.
16. Pencina MJ, Millen BE, Hayes LJ, Agostino RBD. Performance of a method for identifying the unique dietary patterns of adult women and men: the Framingham nutrition studies. J Am Diet Assoc. 2009;108:1453–60.
17. Hearty AP, Gibney MJ. Analysis of meal patterns with the use of supervised data mining techniques—artificial neural networks and decision trees. Am J Clin Nutr. 2008;88:1632–42.
18. Hoffmann K, Schulze MB, Schienkiewitz A, Nöthlings U, Boeing H. Application of a new statistical method to derive dietary patterns in nutritional epidemiology. Am J Epidemiol. 2004;159:935–44.
19. Subar AF, Freedman LS, Tooze JA, Kirkpatrick SI, Boushey C, Neuhouser ML, et al. Addressing current criticism regarding the value of self-report dietary data. J Nutr. 2015;145:2639–45.
20. Pérez Rodrigo C, Morán Fagúndez LJ, Riobó Serván P, Aranceta Bartrina J. Screeners and brief assessment methods. Nutr Hosp. 2015;31(Suppl 3):91–8.

21. Aquino EM, Barreto SM, Bensenor IM, Carvalho MS, Chor D, Duncan BB, et al. Brazilian Longitudinal Study of Adult Health (ELSA-Brasil): objectives and design. Am J Epidemiol. 2012;175: 315–24.
22. Bensenor IM, Griep RH, Pinto KA, Faria CP, Felisbino-Mendes M, Caetano EI, et al. Routines of organization of clinical tests and interviews in the ELSA-Brasil investigation center. Rev Saúde Pública. 2013;47:37–47.
23. Molina MCB, Faria CP, Cardoso LO, Drehmer M, Velasquez-Meléndez JG, Gomes ALC, et al. Diet assessment in the Brazilian Longitudinal Study of Adult Health (ELSA-Brasil): development of a food frequency questionnaire. Rev Nutr. 2013;26:167–76.
24. National Cancer Center (NCC). Nutrition data system for research software. Minneapolis; 2010.
25. Molina MCB, Benseñor IM, Cardoso LO, Velasquez-Melendez G, Drehmer M, Pereira TSS, et al. Reprodutibilidade e validade relativa do Questionário de Frequência Alimentar do ELSA-Brasil. Cad Saude Publica. 2013;29:379–89.
26. Willet W. Correction for the effects of measurement error. In: Willet W editor. Nutritional epidemiology. 2nd ed. New York, NY: Oxford University Press; 1998. p. 74–147.
27. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas. 1960;20:37–46.
28. Landis J, GG K. The measurement of observer agreement for categorical data. Biometrics. 1977;33:159–74.
29. Kastorini CM, Papadakis G, Milionis HJ, Kalantzi K, Puddu PE, Nikolaou V, et al. Comparative analysis of a-priori and a-posteriori dietary patterns using state-of-the-art classification algorithms: a case/case-control study. Artif Intell Med. 2013;59: 175–83.
30. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982; 143:29–36.
31. Hearty ÁP, Gibney MJ. Comparison of cluster and principal component analysis techniques to derive dietary patterns in Irish adults. Br J Nutr. 2008;101:598–608.
32. Winkvist A, Hörnell A, Hallmans G, Lindahl B, Weinehall L, Johansson I. More distinct food intake patterns among women than men in northern Sweden: a population-based survey. Nutr J. 2009;8:12.
33. Villegas R, Salim A, Collins M, Flynn A, Perry I. Dietary patterns in middle-aged Irish men and women defined by cluster analysis. Public Health Nutr. 2004;7:1017–24.
34. Reddy CK, Vinzamuri B. A survey of partitional and hierarchical clustering algorithms. In: Aggarwal CC, Reddy CK editors. Data clustering algorithms and applications. 1st ed. CRC Press Taylor & Francis Group; 2014. p. 8–106.
35. Lima TPF, Sena GR, Neves CS, Vidal SA, Lima JTO, Mello MJG, et al. Death risk and the importance of clinical features in elderly people with COVID-19 using the Random Forest Algorithm. Rev Bras Saúde Matern Infant. 2021;21:445–51.
36. Sidey-Gibbons JAM, Sidey-Gibbons CJ. Machine learning in medicine: a practical introduction. BMC Med Res Methodol. 2019;19:64.
37. Abe S. Support Vector Machines for pattern classification. 1st ed. New York: Springer London; 2010. p. 417.
38. Olson DL, Delen D. Advanced data mining techniques. Berlin, Heidelberg: Springer Berlin Heidelberg; 2008. p. 182.
39. Hand DJ, Mannila H, Smith P. Principles of data mining. 1st ed. Cambridge, MA: MIT Press; 2001. p. 546.
40. Witten IH, Frank E, Hall MA. Data mining practical machine learning tools and techniques. 3rd ed. Elsevier; 2011. p. 629.
41. Xiao Y, Wu J, Lin Z, Zhao X. A deep learning-based multi-model ensemble method for cancer prediction. Comput Methods Programs Biomed. 2018;153:1–9.
42. Silva SHG, Teixeira AF, dos S, Menezes MD, de,Guilherme LRG, Moreira FM, et al. Multiple linear regression and random forest to predict and map soil properties using data from portable X-ray fluorescence spectrometer (pXRF). Ciência e Agrotecnologia. 2017;41:648–64.
43. Cahoolessur D, Rajkumarsingh B. Fall detection system using XGBoost and IoT. R&D J. 2020;36:8–18.
44. Illner A-K, Freisling H, Boeing H, Huybrechts I, Crispim S, Slimani N. Review and evaluation of innovative technologies for measuring diet in nutritional epidemiology. Int J Epidemiol. 2012; 41:1187–203.
45. Boland M, Bronlund J. eNutrition—the next dimension for eHealth? Trends Food Sci Technol. 2019;91:634–9.

AUTHOR BIOGRAPHIES

Vanderlei Carneiro Silva is a nutritionist and a PhD in Epidemiology. Researcher on dietary intake assessment, nutritional epidemiology and machine learning. He is interested in topics such as: aging, diet and the use of technology in health and nutrition. Works with data analysis and visualization in the Brazilian Longitudinal Study of Adult Health (ELSA-Brasil).

Bartira Gorgulho is a nutritionist and a Professor of Nutrition.

Dirce Maria Marchioni is a nutritionist and a Professor of Nutrition. She has experience in the field of Nutrition, with emphasis on assessment of dietary intake, working mainly on the following topics: food consumption, diet and dietary recommendations.

Tânia Aparecida de Araujo is a nutritionist and a PhD in Public Health. She works with research related to health management, vulnerable populations and food and nutrition security. Interest in topics such as epidemiology, aging, social inequities and health.

Itamar de Souza Santos is a researcher in the Center of Clinical and Epidemiological Research and a Professor of Medicine. Graduated in applied and computational mathematics, with qualification in Public Health. The main line of research is in clinical epidemiology.

Paulo Andrade Lotufo is a researcher and a Professor of Medicine. Organizer of epidemiological studies such as the Brazilian Longitudinal Study of Adult Health (ELSA-Brasil). A full-time professor, he works at the University Hospital and directs the Center for Clinical and Epidemiological Research at the University of São Paulo.

Isabela Martins Benseñor is a researcher and a Professor of Medicine. She works in two lines of research: epidemiology of chronic diseases and symptoms. Vice-coordinator of the Brazilian Longitudinal Study of Adult Health (ELSA-Brasil).

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Silva VC, Gorgulho B, Marchioni DM, Araujo TAd, Santos IdS, Lotufo PA, et al. 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). J Hum Nutr Diet. 2022;1–12. https://doi.org/10.1111/jhn.12992