**BMJ Health &
Care Informatics**

Cluster analysis of dietary patterns associated with colorectal cancer derived from a Moroccan case–control study

Noura Qarmiche ![ORCID],[1] Khaoula El Kinany,[2] Nada Otmani,[3] Karima El Rhazi,[2] Nour El Houda Chaoui[1]

![Check for updates]

[1]Laboratory of Artificial Intelligence, Data Science and Emerging Systems, National School of Applied Sciences, Sidi Mohamed Ben Abdellah University, Fes, Morocco
[2]Department of Epidemiology, Clinical Research and Community Health, Sidi Mohamed Ben Abdellah University, Fes, Morocco
[3]Health Informatics and Statistics Unit, Department of Epidemiology, Clinical Research and Community Health, Sidi Mohamed Ben Abdellah University, Fes, Morocco

Correspondence to
Noura Qarmiche;
noura.qarmiche@usmba.ac.ma

ABSTRACT

Introduction Colorectal cancer (CRC) is a global public health problem. There is strong indication that nutrition could be an important component of primary prevention. Dietary patterns are a powerful technique for understanding the relationship between diet and cancer varying across populations.

Objective We used an unsupervised machine learning approach to cluster Moroccan dietary patterns associated with CRC.

Methods The study was conducted based on the reported nutrition of CRC matched cases and controls including 1483 pairs. Baseline dietary intake was measured using a validated food-frequency questionnaire adapted to the Moroccan context. Food items were consolidated into 30 food groups reduced on 6 dimensions by principal component analysis (PCA).

Results K-means method, applied in the PCA-subspace, identified two patterns: 'prudent pattern' (moderate consumption of almost all foods with a slight increase in fruits and vegetables) and a 'dangerous pattern' (vegetable oil, cake, chocolate, cheese, red meat, sugar and butter) with small variation between components and clusters. The student test showed a significant relationship between clusters and all food consumption except poultry. The simple logistic regression test showed that people who belong to the 'dangerous pattern' have a higher risk to develop CRC with an OR 1.59, 95% CI (1.37 to 1.38).

Conclusion The proposed algorithm applied to the CCR Nutrition database identified two dietary profiles associated with CRC: the 'dangerous pattern' and the 'prudent pattern'. The results of this study could contribute to recommendations for CRC preventive diet in the Moroccan population.

INTRODUCTION

Colorectal cancer (CRC) is one of the most malignant cancers and the third-leading cause of cancer death in the word[1] accounting for approximately 700 000 annual deaths worldwide.[2]

Diet and lifestyle are likely to play an important role in the development of CRC, but the complexity of this effect is still unclear. Previous studies have focused on the effects of a single food or nutrient and overlooked the interaction or synergy of foods.[3]

WHAT IS ALREADY KNOWN ON THIS TOPIC
⇒ Diet and lifestyle are believed to play a significant role in the onset of colorectal cancer (CRC).

WHAT THIS STUDY ADDS
⇒ This study investigates this relationship by analysing dietary patterns in Morocco through the use of K-means clustering in a principal component analysis subspace.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY
⇒ The results provide a clearer understanding of the link between dietary habits and CRC in Morocco, enabling the creation of tailored recommendations.

Dietary patterns analyses are a broader picture of food and nutrient intake. This is an alternative and complementary approach to exploring the relationship between diet and CRC risk. Thus, in recent years, there has been increasing interest in identifying dietary patterns as consumed by populations.[4] Knowledge of population specific dietary patterns is important to identify groups at risk for underconsumption or overconsumption of particular nutrients and to create dietary pattern-based guidelines, which may be easier to translate into diets for the public for CRC prevention.

Clustering is an unsupervised machine learning approach. It aims to identify a cluster structure characterised by the maximum data similarity inside a cluster and the maximum data dissimilarity between different clusters.[5] The oldest and most popular clustering method is K-means, which is a vector quantisation algorithm that attempts to partition n observations into k non-overlapping clusters represented by their centroids. The centroid of a cluster is usually the average of the points in that cluster. The K-means method was ranked second among the 10 best data mining algorithms and has become a reference for all

Table 1 Percentage of missing data for each variable

Variables	% of missing data
q1, q11, q17, q31, q32	0.03
q6	0.17
q16	0.2
q15	0.4
q24	0.74
q21p1, q22p1	21.58
q23p1, q23p2	21.61

new proposed methods.[6] It has the advantage of being very simple, robust and efficient. It can be used for a wide variety of data types.[7] Principal component analysis (PCA) is a widely used dimension reduction method. It transforms high-dimensional data into lower-dimensional data. Where coherent patterns can be detected more clearly.[8] PCA is the continuous solution of the cluster membership indicators in the K-means clustering method. Indeed, PCA selects the dimensions with the largest variances to find the best low-rank approximation (in L2 norm) of the data through the singular value decomposition.[8]

Primary objective
The main objective of this study was to identify Moroccan dietary patterns associated with CRC using CRC Nutrition dataset, which is a Moroccan multicentre case–control study. For this, we applied k-means clustering method in a reduced subspace defined by the PCA dimension reduction method.

Related works
Several studies have been conducted on dietary patterns and potential CRC risk in different populations. In Portugal, three dietary patterns were identified: 'healthy', 'low milk and dietary fibre intake' and 'Western' using PCA and Ward's method. This study confirmed the higher risk of CRC in subjects with a 'Western' diet and a 'low intake of milk and dietary fibre'.[9] In a Korean population, a PCA was used to identify three dietary patterns (traditional, Western and conservative). Traditional and conservative patterns were inversely associated with CRC risk.[10]

Among middle-aged Americans, PCA identified three main dietary patterns: a fruit and vegetable pattern, a diet food pattern, and a red meat and potato pattern. Dietary patterns characterised by low frequency of meat and potato consumption and frequent consumption of fruits and vegetables and low-fat foods were consistent with a decreased risk of CRC.[11] Three dietary patterns were defined by PCA labelled 'meat-based', 'plant-based' and 'carbohydrate-based' patterns in Uruguay. The highest risk was positively associated with the meat-based model, whereas the plant-based model was strongly protective. The carbohydrate model was only positively associated with colon cancer risk.[12] Among a Japanese population, three dietary patterns were derived from the PCA: 'conservative', 'western' and 'traditional'. The conservative model showed a reduced association of CRC. The Western model showed a significant positive linear trend for colon. There was no apparent association of the traditional Japanese dietary pattern on overall or site-specific risk of CRC.[13] A Canadian population-based study identified three main dietary patterns using factor analysis, namely a meat-based diet pattern, a plant-based diet pattern and a sugar-based diet pattern. The results suggest that the meat-based diet and the sugar-based diet increase the risk of CRC. In contrast, the plant-based diet decreased the risk of CRC.[14]

For most of these studies, data were obtained by case–control surveys and dietary intakes were assessed using the food-frequency questionnaire (FFQ).

MATERIALS AND METHODS
Study design
This was a Moroccan, national, retrospective, non-interventional and multicentre study in patients wityh CRC.

Setting
This study was conducted in five major University Hospital centres in Morocco, namely Hassan II UHC of Fez, Avicenna UHC of Rabat, Mohammed VI UHC of Oujda, Averroes UHC of Casablanca and Mohammed VI UHC of Marrakech between September 2009 and February 2017. Participating centres were distributed across the country to ensure geographical representation.
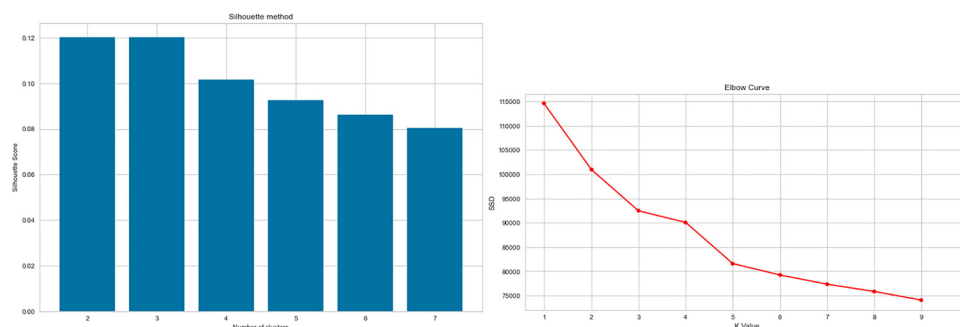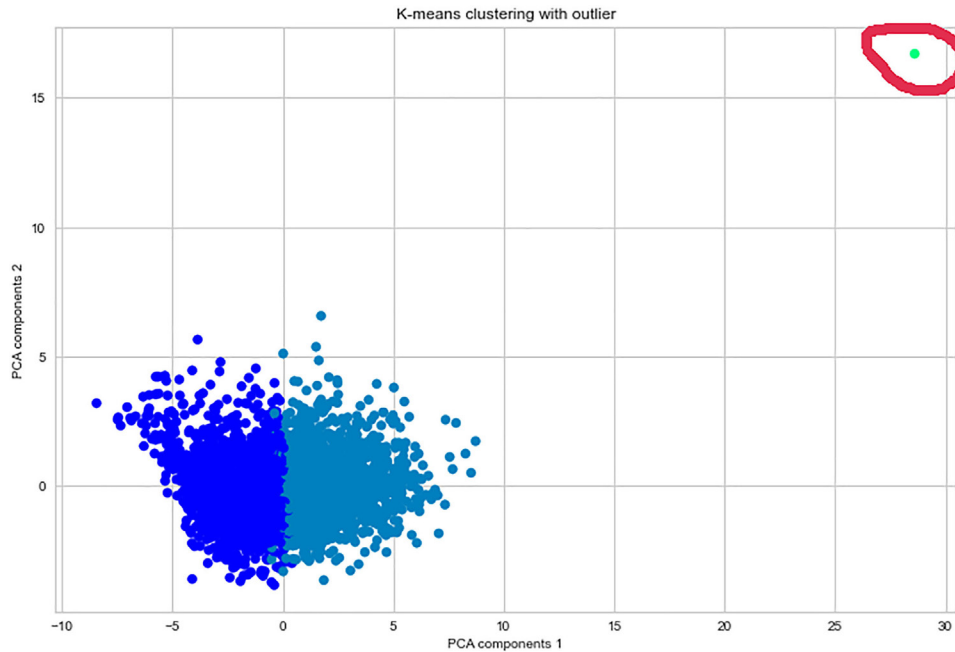

Figure 1 Elbow curve and silhouette histogram.

Figure 2 K-means clustering for outlier detection. PCA, principal component analysis.

Participants

Cases and controls were individually matched on age (±5 years), sex and centre (ratio 1:1). Cases were defined as patients who had recently confirmed CRC diagnosis by histopathology and who did not start any treatment protocol (chemotherapy, radiotherapy, hormonal therapy or surgery) at the time of inclusion. Other eligibility criteria were 18 years of age or older, no history of diabetes mellitus, ability to give consent and ability to communicate and conduct the interview. Controls were selected from the same local population and hospitals as the cases, among healthy subjects accompanying other patients or visitors. Cases and controls both met the same eligibility requirements, with the exception of the criterion that did not have a personal history of CRC or any other type of cancer.[10 15]

Data collection

Data were collected in face-to-face interviews conducted by trained interviewers. All participants were invited to answer questions on the following topics: sociodemographic information (age, sex, centre, residency, profession, marital status, education level, income level and type of habitat), clinical data, substances use, physical activity levels, anthropometric measurements, genetic data and dietary data. Dietary information was obtained via a validated semiquantitative FFQ. This questionnaire was based on the GA2LEN FFQ and was adapted to the Moroccan context.[16] To objectively assess the frequency of food consumption, a detailed frequency scale has been established, including the following options: rarely/never, once to three times per month, once/week, twice to four/week, five to six times/week, once/day, twice to three times/day and equal or more than four times/day.[17]

The 255 FFQ items were initially combined into 30 different food and beverage groups, as follows: bread, breakfast with grains, couscous, pasta, cake, rice, sugar, sweets without chocolate, chocolate, vegetable oil, margarine and vegetable fat, butter and animals fat,

Table 2 Total variance explained by the principal components

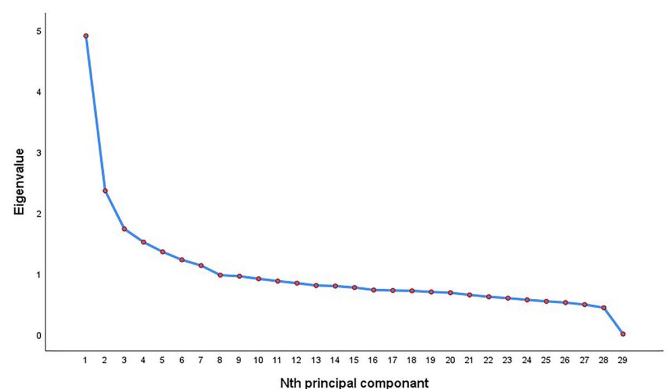| Principal component	Eigenvalue	% variance	% cumulative
1	4901	16899	16899
2	2354	8119	25017
3	1729	5961	30978
4	1512	5213	36191
5	1351	4659	40850
6	1220	4207	45057

Figure 3 PCA scree plot. PCA, principal component analysis.

Table 3 Principal component loadings (correlations between features and principal components (r-value))

CP1	CP2	CP3	CP4	CP5	CP6
Vegetable oil (0.85) Cake (0.65) Chocolate (0,58) Miscellaneous_ foods (0.54) Milk (0.53) Nuts(0.5) Juice(0.49) Rice(0.44) Sugar(0.44) Other dairy products (0.43) Cheese (0.42) Non-alcoholic_ beverages (0.41) pasta (0.41)	Vegetables (0.47) Red meat (−0.45) Breakfast with_ grains (0.42) Offal (−0.42)	Sweets except chocolate (−0.52)	Fruits (−0.52) Fish (−0.5)	Poultry (0.55) Potatoes (0.41)	Bread (−0.45)

nuts, legumes, vegetables, potatoes, fruits, juice, non-alcoholic beverages, coffee/tea, meat, dried meat, poultry, offal, fish, milk of cow/milk of soya, cheese, other dairy products, miscellaneous foods and alcohol. The details of the components of each group are detailed here.[16]

Bias

This non-interventional study is subject to various biases and structural limitations inherent in observational studies. Participants recorded their usual food intake over a longer period (1 year), which could lead to errors in the results. This information bias was addressed at the time of recruitment by trained investigators who collected the data with maximum accuracy. To account for potential confounders in this study, a large amount of data that could affect exposure and outcomes (such as physical activity, body mass index (BMI), alcohol and tobacco use) were collected, and the data were fairly complete for the outcomes.

Study size

The sample size for the study was determined by taking into account the prevalence of red meat consumption as a key exposure of interest. Data from the National Survey of Dietary Habits in Morocco revealed that 62.7% of Moroccan adults eat red meat at least twice a week. The following formula specific for individual-matched case–control studies, the sample size was calculated with 5% type I error, a 90% statistical power and a minimum difference in risk of 43% as reported by the WCRF/AICR report.
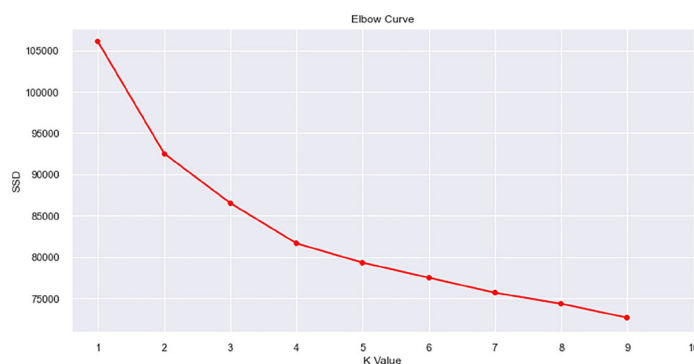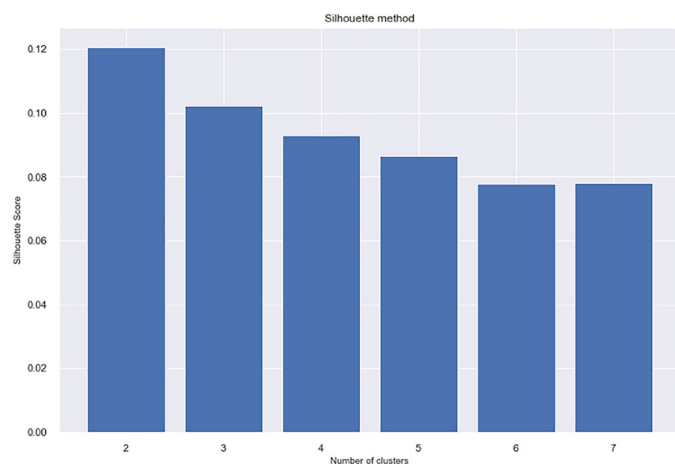
$$n_c = \frac{\left(Z_\alpha \left(\psi+1\right)+2Z_\beta\sqrt{\psi}\right)}{\left(\psi-1\right)^2\left(\psi+1\right)P_{01}}$$

Where n_c= sample size for case–control pairs.
ψ= OR.
P_0= The probability of obtaining a matched pair in which the case is unexposed and the control is exposed.

The number of pairs needed for the study was 1496 rounded to 1500.

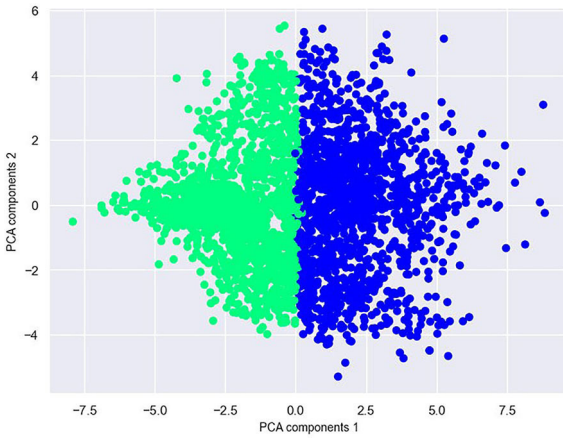**Figure 4** Elbow curve and silhouette histogram after outlier removal. SSD, sum of squares of distances

Figure 5 K-means clustering after outlier removal. PCA, principal component analysis.

Statistical analyses
Data cleaning and handling

In total, 3032 participants were recruited for the study, 1516 cases and 1516 controls. However, 7 participants with unspecified primary cancer, 6 cases with old biopsies, 10 participants with missing dietary data, 2 duplicate records and 8 unmatched records were excluded.

The participation rate in this study was 97% (1516/1555) for cases and 76% (1516/2000) for controls. The final sample included in this study was 1483 cases and 1483 controls.

Data preprocessing

Missing values for each variable were replaced by its mean if the percentage of missing data for that variable is less than 20%, otherwise the variable will be removed from the study.[18] SimpleImputer, which is a sklearn class, was used as imputation method.

All FFQ values are on the same scale and are between 2 and 9, so there was no need to normalise them.

K-means method has been used to detect outliers, which are extreme values, abnormally different from the variable distribution.[19] In clustering analyses, they are in the form of too small groups that must be removed.[20] Detecting outliers allows improving the quality of clustering.[21]

Unsupervised learning algorithms
Principal component analysis

PCA, a dimensionality reduction algorithm, was used to reduce the number of food groups by mapping each instance of a given data set to a k-dimensional subspace called principal components, where k<d. The scree plot was used to identify the number of principal components to retain, which shows the proportion of variance explained by each component. The first component covers most of the model and covers the maximum variance, while each subsequent component covers a lesser value of the variance.[22]

Table 4 Characteristics of consumption across the two dietary patterns

Cluster	Prudent pattern (mean±SD)	Dangerous pattern (mean±SD)	P value
CP1			
Oil	2.78±0.42	3.61±0.69	<0.001
Cake	2.49±0.9	4.04±1.52	<0.001
Chocolate	2.28±0.78	3.27±1.5	<0.001
Miscellaneous foods	2.41±0.3	2.62±0.46	<0.001
Milk	2.76±0.44	3.11±0.46	<0.001
Nuts	2.51±0.94	3.3±1.27	<0.001
Juice	2.38±0.61	2.74±0.84	<0.001
Rice	3.12±1.02	3.78±0.96	<0.001
Sugar	3.83±0.77	4.32±0.84	<0.001
Other dairy products	2.04±0.21	2.13±0.4	<0.001
Cheese	3.37±1.74	5.39±1.57	<0.001
Non-alcoholic_ beverages	2.58±0.73	2.84±0.91	<0.001
Pasta	2.89±1.04	3.8±1.08	<0.001
CP2			
Vegetables except potatoes	6.3±1.23	6.2±1.04	<0.001
Red meat	4.21±1.22	4.44±1.15	<0.001
Breakfast with grains	2.79±1.22	3.24±1.39	<0.001
Offal	2.15±0.37	2.24±0.47	<0.001
CP3			
Sweets Except chocolate	2.17±0.72	2.62±1.24	<0.001
CP4			
Fruits	5.33±1.49	5.15±1.28	<0.001
Fish	3.74±0.98	4.19±0.96	<0.001
CP5			
Poultry*	4.48±0.98	4.49±0.95	0.586
Potatoes	5.19±1.41	5.46±1.07	<0.001
CP6			
Bread	8.03±0.97	8.32±0.71	<0.001

K-means clustering

K-means clustering aims to divide M points in N dimensions into a set C of K clusters Cj with cluster mean cj to reduce the sum of squared errors.[23 24] This is described as follows:

$$E = \sum_{j=1}^{k} \sum_{x_i \in c_j} \| c_j - x_i \|^2 \qquad (1)$$

Where, E is sum of the square error of objects with cluster means for K cluster and distance metric between a data point and a cluster mean. The Euclidean distance is defined as:

$$\| x - y \| = \sqrt{\sum_{i=1}^{v} |x_i - y_i|^2} \qquad (2)$$

Table 5 Distributions of sociodemographic characteristics of the study population by the two clusters

	Prudent pattern	Dangerous pattern	P value
Age			
[18–30[	1.96	2.29	0.753
[30–45[	9.68	9.04	
[45–60[	20.17	18.58	
[60–75[	16.12	14.94	
>75	3.61	3.61	
Sex			
Female	25.94%	24.38%	0.994
Mal	25.60%	24.08%	
Marital status			
Single	5.16%	4.55%	0.086
Married	38.48%	37.91%	
Divorced	1.92%	1.65%	
Widowed	5.97%	4.35%	
Residence			
Urban	35.35%	36.73%	<0.001
Rural	16.19%	11.74%	
Level of education			
Illiterate	31.10%	25.67%	
Primary	9.75%	9.04%	0.001
Secondary	7.05%	8.09%	
Higher	3.64%	5.67%	
Profession			
Unemployed	7.82%	5.60%	0.001
Housewife	20.13%	17.17%	
Student	0.30%	0.51%	
Working	19.43%	20.57%	
Retired	3.84%	4.62%	
Smoking status			
Non-smoker	42.12%	40.92%	0.95
Smoker	5.40%	4.75%	
BMI (kg/m^2)			
(16–18.5)	1.00	1.07	
(18.5–25)	21.73	20.87	
(25–30)	21.73	21.25	0.1
≥30	7.13	5.23	
Physical activity			
Yes	10.56%	11.32%	0.061
No	40.98%	37.13%	
Monthly household income (DHMAD)			
≤2000	42.80%	34.00%	0.001
(2000–5000)	6.64%	10.42%	
(5000–10 000)	2.09%	4.05%	

BMI, body mass index.

Following vector defines the average of a cluster by:

$$c_j = \frac{1}{|c_j|} \sum_{i \in c_j} x_i \qquad (3)$$

Choice of the optimal number of clusters K

In order to determine the optimal number of clusters, we used the Elbow method complemented by silhouette analysis, which calculates the separation distance between the resulting clusters and provides a way to visually assess their number.[25–27]

Proposed method

The K-means method has been applied in the PCA-subspace, as strongly advised by several studies.[8 28 29] Indeed, the continuous solution of the cluster indicators is given by the PCA principal components and the optimal solution of the K-means clustering is inside the PCA-subspace .

Association test

To test the association between clusters and CRC status, the simple logistic regression test was used. Result was presented by OR value and its CI.

Student's t-test was used to assess the relationship between the clusters and food consumption. P values less than 0.05 were considered statistically significant.

The algorithm proposed in this study is presented in online supplemental figure 1.

RESULTS
Data preprocessing
Managing missing data

The number of missing values was calculated by the isnull().sum() function of Pandas. The results obtained are presented in table 1 (only the variables that contained missing data have been reported).

Missing data for variables q1, q6, q11, q15, q16, q17, q24, q31, q32 were replaced by the mean, using Sklearn's simple imput function.

The variables q21p1, q22p1, q23p1, q23p2 that corresponds to alcohol consumption were removed from the study because they contained more than 20% of missing data.

Detection of outliers

The Elbow and Silhouette methods (figure 1) indicate that the appropriate number of clusters k is 3.

K-means identified three distinct groups in our population study (figure 2.). However, it is very evident that one of the groups is simply an outlier since it contains only one point. After checking the database, we verified the existence of an outlier (q15=99) and deleted the record corresponding to this value before running our algorithm again with the new database.

Dimensionality reduction

According to the scree plot figure 3, we have retained six principal components, which were defined by PCA.

From table 2, we notice that the first principal component constitutes 16.89% of the variance. The composition of the first and second axis constitutes 25.01% of the total variance. While the cumulative variance of the 6 principal components represents 45.05% of the total.

The correlation of each principal component with its constituents is presented in table 3 (only correlations >0.4 are reported).

K-means clustering

The results of the Elbow and Silhouette methods (figure 4) indicate that the appropriate number of clusters k is 2.

K-means clustering identified two distinct groups in this population (figure 5). A total of 1433 participants (48.33%) were in cluster 0 while 1531 (51.67%) were in cluster 1. 55.95% of individuals in cluster 0 were controls while 44.04% were cases. Cluster 1 is composed of 44.41% controls and 55.59% cases.

Mean and SD consumption of food groups in each cluster are shown in table 4. The p value between groups was significant (<0.001) for most food groups, with the exception of poultry (p=0.586).

We describe cluster 1 as a 'dangerous pattern' because it showed high loadings of vegetable oil, cake, chocolate, cheese, red meat, sugar and butter. Cluster 0 was termed the 'prudent diet' cluster due to moderate consumption of almost all foods with a slight increase in fruits and vegetables (online supplemental figure 2).

The student test showed a significant relationship between CRC and cluster (p<0.001). Indeed, people who belong to the 'dangerous pattern' have a higher risk to develop CRC with an OR 1.59 (95% CI 1.375 to 1.383).

The distributions of sociodemographic characteristics by cluster are presented in table 5 . No significant differences between dietary patterns were found by age, sex, BMI, marital status, physical activity or smoking status with p values equal to 0.753, 0.994, 0.1, 0.086, 0.061 and 0.95, respectively.

The proportions of the unemployed and housewives were greater in the conservative profile, while the proportions of working and retired people were higher in the dangerous cluster. We also note that the number of people in the dangerous cluster increases proportionally with income and educational level.

DISCUSSION

The proposed algorithm applied to the CCR Nutrition database, which is a multicente case–control study conducted in a population of 1496 pairs of Moroccan subjects with and without CRC, identified 2 dietary profiles associated with CRC: the 'dangerous pattern' and the 'prudent profile'. The 'dangerous pattern' was characterised by a high consumption of vegetable oil, cakes, chocolate, cheese, red meat, sugar and butter. While the 'prudent pattern' was characterised by a moderate consumption of almost all foods with a slight increase in fruits and vegetables. The frequency of cases was higher in the 'dangerous' group than in the 'prudent' group.

This study proposes a new methodological approach that combined two unsupervised machine-learning techniques: PCA and K-means. The K-means method has been applied in the PCA-subspace. Several studies have shown the advantages of this approach.[8 18 28] Indeed, the continuous solution of the cluster indicators is given by the principal components of the PCA and the optimal solution of the K-means clustering is in the PCA subspace. Moreover, the performance of clustering is better at reduced cost and noise. A recent statistical methods review for dietary pattern analysis reported the advantages and the disadvantages of PCA and k-means clustering algorithm. Compared with traditional statistical methods, classification via machine learning techniques reduces misclassification rate, increases generalisability, allows grading of movement quality, and simplifies experimental design.

Other strengths of our research should be mentioned; first, it is the first study on the clustering of dietary profiles related to CRC in Morocco by an unsupervised machine learning approach, according to the literature search. On the other hand, in our case–control study, we included recent diagnosed CRC cases to avoid diet changes. In addition, trained interviewers ensured FFQ questionnaires fulfilment in order to maintain the responses objectivity.[15]

Two limitations of our study must be highlighted; the first one, our clustering was based on food groups containing foods known to be protective against CRC and others known to be risk factors. Thus, clustering of these foods may neutralise their effects and make discrimination difficult. The second one, food consumption was based on frequencies without considering the daily quantities which can influence the clustering.

A recent study used Global Dietary database (Canada, India, Italy, South Korea, Mexico, Sweden and the USA) found that CRC could be predicted based on a list of important dietary data using supervised and unsupervised machine learning approaches. This study identified the following two patterns, total fat, mono unsaturated fats, linoleic acid, cholesterol, omega-6 as moderate to high correlated dietary features to positive CRC, and fibre and carbohydrates as negative correlation with CRC cases. A systematic review of 17 years of evidence (2010–2016) revealed two distinct global dietary patterns related to CRC risk: a 'healthy' pattern, characterised by high intake of fruits and vegetables, higher intakes of one or more of the following foods; whole grains, nuts and legumes, fish and other seafood, milk and other dairy products, and an 'unhealthy' dietary pattern characterised by high intakes of red and processed meat, sugar-sweetened beverages, refined grains and desserts and potatoes.

Several studies in American, European and Asian populations have found three dietary patterns related

to CRC[9 11 13 14 30]: 'Western or meat-based diet' which is related with higher risk of CRC, 'healthy or conservative or prudent' which is related with low risk of CRC and 'low milk and dietary fibre intake or traditional' which is relatively related with higher risk of CRC. We could not obtain a very clear group due to diverse nature of nutrition landscape in the Moroccan population, although there were higher intakes of some harmful foods in the cases compared with the controls (meat, sugar and chocolate). The difference in poultry consumption was non-significant between the two clusters, which was similarly reported in a previous study.[31]

The perspectives of this work are as follows: first to repeat the clustering process, but this time with single foods to overcome the limitation of grouping protective and risk foods in the same group, and neutralise their effect. Second, to develop an easy and user-friendly web application that allows the simple user to identify him/herself in a dietary pattern and evaluate whether he/she is following a healthy diet or not, which is the best approach to make a personal prevention as recommended by the latest WHO guidelines.[32]

CONCLUSION

The combination of the two unsupervised learning methods PCA and K-means identified two clusters describing two main dietary patterns related to CRC in the Moroccan population, labelled: 'prudent' and 'dangerous'. The number of cases was relatively higher in the 'dangerous' group than in the 'prudent' group. The unsupervised learning approach proposed in this paper was effective and confirmed the results of the literature but in a more discriminant manner.

Data availability statement No data are available.

ORCID iD
Noura Qarmiche http://orcid.org/0000-0002-1786-5049

REFERENCES

1. 900-world-fact-sheets.pdf. 2021. Available: https://gco.iarc.fr/today/data/factsheets/populations/900-world-fact-sheets.pdf
2. Torre LA, Bray F, Siegel RL, *et al*. Global cancer statistics, 2012. *CA Cancer J Clin* 2015;65:87–108.
3. Clinton SK, Giovannucci EL, Hursting SD. The world cancer research fund/american institute for cancer research third expert report on diet, nutrition, physical activity, and cancer: impact and future directions. *J Nutr* 2020;150:663–71.
4. Schwerin HS, Stanton JL, Smith JL, *et al*. Food, eating habits, and health: a further examination of the relationship between food eating patterns and nutritional health. *Am J Clin Nutr* 1982;35(5 Suppl):1319–25.
5. Sinaga KP, Yang MS. Unsupervised K-means clustering algorithm. *IEEE Access* 2020;8:80716–27.
6. Wu X, Kumar V, Ross Quinlan J, *et al*. Top 10 algorithms in data mining. *Knowl Inf Syst* 2008;14:1–37.
7. Wu J. Advances in k-means clustering. In: *Advances in K-means Clustering: A Data Mining Thinking*. Berlin, Heidelberg: Springer Science & Business Media, 2012.
8. Ding C, He X. *K*-means clustering via principal component analysis. Twenty-first international conference; Banff, Alberta, Canada.New York, New York, USA, 2004:29
9. Magalhães B, Bastos J, Lunet N. Dietary patterns and colorectal cancer: a case-control study from Portugal. *Eur J Cancer Prev* 2011;20:389–95.
10. Park Y, Lee J, Oh JH, *et al*. Dietary patterns and colorectal cancer risk in a Korean population: a case-control study. *Medicine (Baltimore)* 2016;95:e3759.
11. Flood A, Rastogi T, Wirfält E, *et al*. Dietary patterns as identified by factor analysis and colorectal cancer among middle-aged Americans. *Am J Clin Nutr* 2008;88:176–84.
12. De Stefani E, Ronco AL, Boffetta P, *et al*. Nutrient-derived dietary patterns and risk of colorectal cancer: a factor analysis in Uruguay. *Asian Pac J Cancer Prev* 2012;13:231–5.
13. Shin S, Saito E, Sawada N, *et al*. Dietary patterns and colorectal cancer risk in middle-aged adults: a large population-based prospective cohort study. *Clin Nutr* 2018;37:1019–26.
14. Chen Z, Wang PP, Woodrow J, *et al*. Dietary patterns and colorectal cancer: results from a Canadian population-based study. *Nutr J* 2015;14:8.
15. Mint Sidi Ould Deoula M, Huybrechts I, El Kinany K, *et al*. Behavioral, nutritional, and genetic risk factors of colorectal cancers in Morocco: protocol for a multicenter case-control study. *JMIR Res Protoc* 2020;9:e13998.
16. El Kinany K, Garcia-Larsen V, Khalis M, *et al*. Adaptation and validation of a food frequency questionnaire (FFQ) to assess dietary intake in Moroccan adults. *Nutr J* 2018;17:61.
17. El Kinany K, Mint Sidi Deoula M, Hatime Z, *et al*. Consumption of modern and traditional Moroccan dairy products and colorectal cancer risk: a large case control study. *Eur J Nutr* 2020;59:953–63.
18. Cismondi F, Fialho AS, Vieira SM, *et al*. Missing data in medical databases: impute, delete or classify? *Artif Intell Med* 2013;58:63–72.
19. Benzaki Y. Tout savoir sur les valeurs aberrantes (outliers). mr. mint: apprendre le machine learning de A à Z. 2017. Available: https://mrmint.fr/outliers-machine-learning

20 DataTechNotes. Anomaly detection example with K-means in python. 2022. Available: https://www.datatechnotes.com/2020/05/anomaly-detection-with-kmeans-in-python.html

21 Dino L. Outlier detection using K-means clustering in python medium. 2022. Available: https://towardsdev.com/outlier-detection-using-k-means-clustering-in-python-214188fc90e8

22 Keerthi Vasan K, Surendiran B. Dimensionality reduction using principal component analysis for network intrusion detection. *Perspectives in Science* 2016;8:510–2.

23 Farhang Y. n.d. Face extraction from image based on k-means clustering algorithms. *Ijacsa*;8.

24 Hartigan JA, Wong MA. Algorithm as 136: a k-means clustering algorithm. *Applied Statistics* 1979;28:100.

25 Umargono E, Suseno JE, Vincensius Gunawan SK. K-means clustering optimization using the elbow method and early centroid determination based on mean and median formula. The 2nd International Seminar on Science and Technology (ISSTEC 2019); Yogyakarta, Indonesia.Paris, France, November 3, 2020:121–9

26 Tout ce que vous voulez savoir sur l'algorithme K-Means. Mr. mint: apprendre le machine learning de A à Z. 2018. Available: https://mrmint.fr/algorithme-k-means

27 Zhou HB, Gao JT. Automatic method for determining cluster number based on silhouette coefficient. *AMR* 2014;951:227–30.

28 Xu Q, Ding C, Liu J, *et al*. PCA-guided search for k-means. *Pattern Recognition Letters* 2015;54:50–5.

29 Refining initial points for K-means clustering | bibsonomy. 2022. Available: https://www.bibsonomy.org/bibtex/29433d748d0d60d70afdeb54f9418baad/ans

30 Garcia-Larsen V, Morton V, Norat T, *et al*. Dietary patterns derived from principal component analysis (PCA) and risk of colorectal cancer: a systematic review and meta-analysis. *Eur J Clin Nutr* 2019;73:366–86.

31 The VARCLUS procedure. In: 43. n.d:

32 Ward JH. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association* 1963;58:236–44.