Proceedings of Science and Mathematics
Faculty of Science,
Universiti Teknologi Malaysia

Vol. 8, 2022, page 159-169

Statistical Analysis of Food Nutrition by Using K-Means Clustering

Nurul Anis Suraya Rosmahadi, Muhammad Fauzee Hamdan*
Department of Mathematical Sciences, Faculty of Science
Universiti Teknologi Malaysia, 81310 Johor Bahru, Johor, Malaysia
*Corresponding author: mfauzee@utm.my

Abstract

In Malaysia, there are numerous varieties of foods about which individuals are unaware of the nutritional information. Perhaps the compositions of some of the items are comparable. The focus of this research is to use K-means clustering to discover groupings of foods in Malaysia that have similar nutritional information. To put it another way, the objective is to sort them into clusters based on how similar they are. This research examined at the food compositions dataset, which would include calories, carbohydrates, protein, fat, cholesterol, sodium, and sugar. K-means clustering was applied to analyze this data because in clustering, objects that act or are near to each other are grouped in one cluster, while objects that are far apart or unfamiliar are put in another. The fundamental idea of this technique is to define a K centroid for each cluster in order to acquire the result. In general, the K-means clustering procedure starts with K-tentative centroids and performs two phases repeatedly. The first is gathering clusters around centroids, and the second is maintaining the K centroid up to date. The clusters acquired from the data can be used to identify similarities between types of foods that are comparable in their characteristics in each of the compositions. The dataset is analyzed using the SPSS software package, which applies cluster analysis methods. The results consist of initial cluster centers, table of iteration history, final cluster centers, distances between the final clusters and the number of cases in each cluster. A dendrogram is illustrated to show the relationships between similar sets of data.

Keywords: Nutrition; K-means clustering; Cluster Analysis; Non-hierarchical; Centroids

1. Introduction

If there is one thing Malaysians are proud of, it is without a doubt their cuisine. Malaysia is a foodie's paradise, and Malaysians in general appreciate their cosmopolitan society's unique gastronomic heritage. As a result, Malaysia's cuisine is distinctive; it makes greater use of the richness of natural resources and foods available than any other Asian country. Malaysian cuisine reflects the country's diverse nationalities.

Health and wellness are a complex thing and needs a comprehensive approach in a lifestyle to reach and maintain it. Food is a prime pillar for health and fitness. Complexity rises as we move into characteristics of food items to classify what is health and what is unhealthy that too taking the consideration of health and fitness goals of an individual. Promoting the consumption of healthier foods for example, eating more fruits and vegetables is a focus of public health efforts that gives advantages to overall community health.

Nevertheless, people can discover all diverse form of ingredients at anywhere together with nearby meals outlet, stall and road hawker. This results in trouble of people to pick a right meal for them. In this age of globalization, malnutrition in all its forms continues to be one of the greatest challenges faced by our generation. Unhealthy diets Unhealthy diets are a vital cause of malnutrition. Poor eating habits consist of overeating or under-eating, not having enough of the healthy foods people need each day, or eat up and drink up too many types of food and beverages, which are low in fiber or high in fat, salt and sugar. This concerning issue is very important for people to make proper decision on how to choose healthy food and drinks from now on. A statistical system for examining food nutrition based on cluster analysis.

In this research, K-means clustering will be applied for analyzing the food nutrition result. The element of data will contain calories, carbohydrate, sodium, sugar, protein, fat and cholesterol. There are many methods a researcher can used in order to analyze the data. For this research, clustering algorithms will be selected to analyze the food nutrition according to their classes. The result will be achieved at the end of this research.

The major goals of this study are to use K-means clustering to analyze the similarities of meals or dishes in Malaysia. The goals are to identify groups of foods in Malaysia based on their nutritional facts that are similar by using K-means clustering. Next, the objectives are to suggest and compare the result of the nutritional values of foods. The number of clusters will be achieved in the end of results. The software that will be chosen in this study is Statistical Package for the Social Sciences (SPSS). By applying this software, it will generate a various results of K-means clustering.

In this research, there are fifty foods categorized under types of foods which are fruits, noodles, seafood, bread, rice dishes and desserts. They were analyzed using K-means clustering. The nutritional facts that have been analyzed consists of calories, carbohydrates, protein, fat, cholesterol, sodium and sugar. All of the respective foods and their nutritional facts were taken from two website which are *www.nutriotionix.com* and *https://www.myfitnesspal.com/*.

This K-means clustering research could help researchers to better understand the many types of foods available in Malaysia and their nutritional values. Furthermore, this science is essential for comparing the outcomes achieved using K-means clustering. Aside from that, it can assist Malaysians in deciding which meals are appropriate for their lifestyle and preferences.

2. Literature Review

2.1. Clustering Algorithm

Data mining is a technology used in different disciplines to search for significant relationships among variables in large data sets which can discover hidden relationships and patterns [5]. The basic approach in data mining is to summarize the data and to extract reasonable and previously unknown useful information. Data mining is widely used in commercial applications.

Cluster analysis is a technique used in data mining that involves the process of grouping objects with similar characteristics, and each group is referred to as a cluster. Cluster analysis is used in various fields, such as biology, image processing and genetics. An order for clustering, the order in which the sequences are compared can have an effect on the multiple final alignment, hence a good order must be chosen. Type of clustering algorithm can be divided into five categories which are hierarchical clustering algorithm, K-means clustering algorithm, density-based clustering algorithm, self-organization maps (SOM) and expectation-maximization (EM) clustering algorithm.

Cluster analysis is a technique used in data mining that involves the process of grouping objects with similar characteristics, and each group is referred to as a cluster. Cluster analysis is used in various fields, such as biology, image processing and genetics. An order for clustering, the order in which the sequences are compared can have an effect on the multiple final alignment, hence a good order must be chosen. Type of clustering algorithm can be divided into five categories which are hierarchical clustering algorithm, K-means clustering algorithm, density-based clustering algorithm, self-organization maps (SOM) and EM clustering algorithm.

2.1.2. Hierarchical Clustering Research

Dendrograms, which are hierarchical clustering solutions in the form of trees, are of tremendous interest in a variety of application domains. Hierarchical trees allow people to see data at several levels of abstraction. Flats partitions of various granularities can be recovered during data analysis due to the consistency of clustering solutions at many levels of granularity, making them excellent for interactive exploration and display. Furthermore, clusters frequently have subclusters, and hierarchical structures naturally represent the underlying application domain, such as biological taxonomies [17].

The most popular type of cluster analysis in food science and technology is agglomerative hierarchical cluster analysis. The aim is to identify a series of clusters within a nested structure. This technique assumes a hierarchical structure in the data set. It starts with each object as a separate cluster and then merges the two closest clusters ana until only one cluster analysis is left. The basic steps of an agglomerative hierarchical clustering are as below:

a) Calculate a distance matrix includes the distances between all the clusters.
b) Merge two closest clusters.
c) Update the distance matrix to include the distance between the new cluster and the original ones, considering a clustering procedure.
d) Repeat the procedures above until only one cluster is left.

A past study presented that a hierarchical cluster analysis was used to identify subgroups with homogeneous gait patterns which the researchers' purpose of their exploratory study was to investigate whether a large group of inured and healthy runners can be clustered into subgroups of homogeneous gaits patters based on 3D kinematic data [7]. The study involved with a sample of 291 injured and healthy runners were queried from an existing database of running kinematics. It is an example of big dataset.

A hierarchical cluster tree or dendrogram was formed with the linkage-function in MATLAB. The functions were used with the Ward's linkage method and Euclidean distance. The subgroups were formed in an agglomerative manner for example, starting with each observation as their own subgroup and at every step pairing the two closest subgroups together until only one group remains. The final number of subgroups was chosen based on a stopping rule which means that a large percentage decrease in the coefficient followed by a plateau. The number of subgroups was also confirmed by visual inspection of the dendrogram. The results where 5 subgroups were identified, however, runners with similar injuries or no injury did not cluster together. Instead, different types of injuries, and healthy control subjects, were evenly distributed across the 5 subgroups.

2.1.3. Non-Hierarchical Clustering Research

Non-hierarchical clustering possesses as a monotonically increasing making of strengths as cluster themselves progressively become members of larger clusters. These clustering methods do not possess tree-like structures and new clusters are formed in successive clustering either by splitting or merging clusters. One of the non-hierarchical cluster analyses is the partitioning method. Consider a given number of clusters, for instance g, as the objective and the partition of the object to obtain the required g clusters. In contrast to the hierarchical clustering method, this partitioning technique permits objects to change group membership through the cluster formation process. The partitioning method usually begins with an initial solution, after which reallocation occurs according to some optimality criterion.

A past study used non-hierarchical clustering to better identify specific gait patterns in patients with cerebellar ataxia (CA), Hereditary SP (HSP), and Parkinson's disease (PD) compared to each other and to healthy subjects [Serrao]. This study involved patients with degenerative neurological diseases such as cerebellar ataxia, Parkinson's disease and spastic paraplegia often display progressive gait function decline that inexorably impacts their autonomy and quality of life. The aim of this study is to determine whether an entire dataset of gait parameters recorded in patients with degenerative neurological diseases can be clustered into homogeneous groups distinct from each other ad from healthy objects. The dataset of this study is 129 patients.

2.1.4. K-Means Clustering Research

K-means clustering was applied to address the scalability issues associated with traditional recommender systems. Recommender systems have the ability to filter unseen information for predicting whether a particular user would prefer a given item when making a choice. This process has been dependent on robust application of data mining and machine learning techniques, which are known to have scalability issues when being applied for recommender systems. An issue with traditional K-means clustering algorithms is that they choose the initial k-centroid randomly, which leads to inaccurate recommendations and increased cost for offline training of clusters [16].

This past research highlights how centroid selection in K-means based recommender systems can improve performance as well as being cost saving. The proposed centroid selection method has the ability to exploit underlying data correlation structures, which has been proven to exhibit superior accuracy and performance in comparison to the traditional centroid selection strategies, which choose centroids randomly. By completing the research, these experiments proved that the proposed approach provides a better-quality cluster and converges quicker than existing approaches, which in turn improves accuracy.

A limitation of K-means clustering algorithm in this past study is that it highly depends on k, number of clusters and k must be predefined. The researchers suggested that developing some statistical methods to compute the k value, depending on the data distribution for future research.

2.2. Euclidean Distance Measure

The Euclidean distance measure is frequently used as a distance measure, and is easy to use in two dimensional planes. As the number of dimensions increases, the calculability time also increases. The formula defines data objects i and j with a number of dimension equal to p. the distance between the two objects $d(i,j)$ is expressed as given in formula (2.1) below. The formula is rather straightforward as the distance is calculated from the cartesian coordinates of the points using the Pythagorean theorem.

$$d(i,j) = \sqrt{\sum_{k=1}^{n}(x_{ij} - x_{ik})^2} \qquad (1)$$

where,

$d(i,j)$ = root of square distance between object i and j
x_{ij} = the value of the kth variable for ith object
x_{ik} = the value of the kth variable for jth object
n = number of variables

2.3. Ward's Linkage

Ward's linkage is a hierarchical cluster analysis technique. The concept is similar to analysis of variance (ANOVA). The rise in the error sum of squares (ESS) after merging two clusters into a single cluster is used to compute the linkage function, which specifies the distance between two clusters. Ward's linkage aims to choose the subsequent clustering steps in such a way that the rise in ESS at each step is minimized.

3. Methodology

3.1. Cluster Analysis

Cluster analysis provides insight into the data by dividing objects into groups of objects (clusters), making objects in one cluster more similar to each other than objects in other clusters [15]. For example, suppose we collected a set of pebbles from the bank of a stream, observed their size, shape, and color characteristics, then sorted comparable pebbles into the same piles. By doing so, we may perform a cluster analysis physically. A cluster is a collection of similar pebbles [9].

The skeleton of any subject is the part that cannot be eliminated without damaging the subject itself when it is stripped of detail. Details can be added and understood in connection to each other once the skeleton has been seen. As a result, cluster analysis is worthwhile. Its skeleton is made up of six steps which are:

a) Obtain the data matrix.
b) Standardize the data matrix.
c) Compute the resemblance matrix.
d) Execute the clustering method.
e) Rearrange the data and resemblance matrices.
f) Compute the cophenetic correlation coefficient.

Thus, given that no information on group definition is formally evaluated in advance, the major problems of cluster analysis will be discussed as follows:

a) What measure of inter-subject similarity is to be used and how is each variable to be "weighted" in the construction of such a summary measure?
b) After inter-subject similarities are achieved, how are the classes to be formed?
c) After the classes have been formed, what summary measures of each cluster are appropriate in a descriptive sense; that is, how are the clusters to be defined?
d) Assuming the adequate descriptions of the clusters can be acquired, what inferences can be drawn regarding their statistical significance?

3.2. K-Means Clustering

Each data point is assigned to the nearest partition depending on some similarity parameter throughout each pass of the algorithm (such as Euclidean distance measure). A data may switch partitions with each succeeding pass, consequently changing the values of the original partitions [8]. K-means clustering is a type of partitioning-based grouping approach that involves iteratively moving data points between clusters. Based on the features discovered, it is used to partition either the cases or the variables of a dataset into non-overlapping groups, or clusters. Which dimensions of the dataset we wish to reduce the dimensionality of [15] determines whether the procedure is applied to the cases or the variables.

Starting with k initial candidate cluster centroids, the k-means algorithm splits n number of objects into k number of clusters (where k is the number of desired clusters specified by the user). Each object is assigned to the nearest centroid, while a candidate cluster is a collection of objects assigned to a single candidate centroid. For each candidate cluster, the candidate centroids are substituted with computed centroids, and the process is repeated iteratively until no change in cluster membership or centroid placements is seen in the final iteration [8].

3.2.1 Objectives of K-Means Clustering

The objectives of K-means clustering are to produce groups of cases/variables with a high degree of similarity within each group and a low degree of similarity between groups. The K-means clustering technique can also be described as a centroid model as one vector representing the mean is used to describe each cluster. the main use of k-means clustering to be more of a way for researchers to gain qualitative and quantitative insight into large multivariate data sets than a way to find a unique and definitive grouping for the data.

It is very useful in exploratory data analysis and data mining in any field of research, and as the growth in computer power has been followed by a growth in the occurrence of large data sets. Its ease of implementation, computational efficiency and low memory consumption has kept the k-means clustering very popular, even compared to other clustering techniques. The objectives of using K-means clustering in this research is to cluster all the foods according to their similarity.

3.2.2 Methods of K-Means Clustering

In general, the cluster finding process according to K-means starts from K tentative centroids and repeatedly applies two steps:

a) Collecting clusters around centroids.
b) Updating centroids as within cluster means.

Initialization is when user chooses the number K of clusters and puts K hypothetic cluster centroids among the entity points.

a) Cluster update: Given K centroids $c_k\ (k = 1, 2, \dots, K)$, each of the entities $i \epsilon I$ is assigned to one of the centroids according to minimum distance rule: distances between i and each c_k are calculated, and i is assigned to the nearest c_k. For each centroid c_k, the entities assigned to it form cluster $S_k\ (k = 1, 2, \dots, K)$.

b) Centroid update: At each of the given K clusters S_k , its gravity centre is computed and set as the new centroid $c_k{}' \ (k = 1, 2, \dots, K)$.

c) Halting test: new centroids $c_k{}'$ are compared with those from the previous iteration. If $c_k{}' = c_k$ for all $c_k'(k = 1, 2, \dots, K)$, stop and output both $c_k{}'$ and S_k for all $c_k{}' \ (k = 1, 2, \dots, K)$. Otherwise, set c_k' as c_k and go to the step 1.

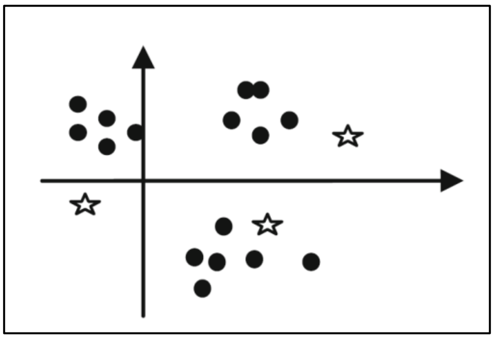

Figure 1 Initialization of centroids

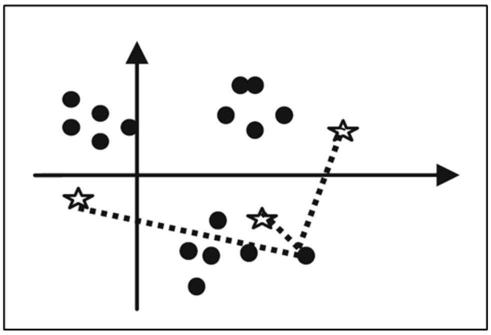

Figure 2 Cluster update using minimum distance rule

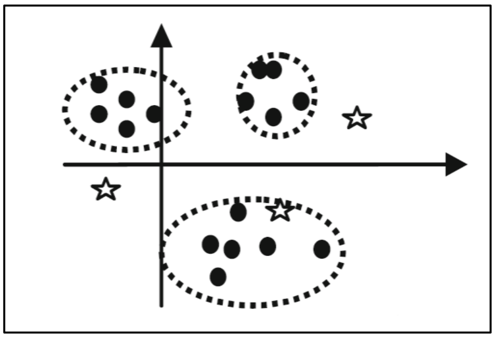

Figure 3 Cluster update completed

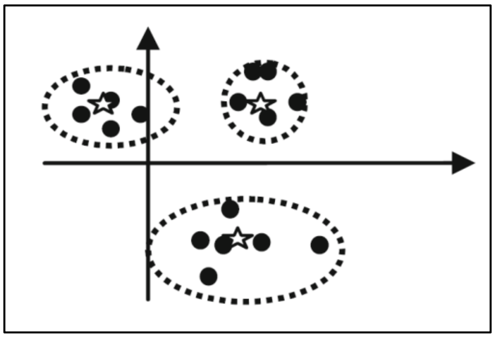

Figure 4 Centroid update completed

The algorithm is appealing in several aspects. Conceptually it may be considered a model for the human process of typology making, with types represented by clusters S_k and centroids c_k. Also, it has nice mathematical properties. This method is computationally easy, fast and memory-efficient. However, researchers and practitioners point to some fewer desirable properties of K-Means. Specifically, they refer to lack of advice with respect to:

a) The initial setting which are the number of clusters K and initial positioning of centroids.

b) Instability of clustering results with respect to the initial setting and data standardization.

c) Insufficient interpretation aids.

These issues can be alleviated, to an extent, as will be explained later in this section. A decoder-based summarization model underlying the method is that the entities are assigned to clusters in such a way that each cluster is represented by its centroid, sometimes referred to as the cluster's standard point or prototype. This point expresses, intentionally, the typical tendencies of the cluster.

3.3. Source of Data

The data used in this study is a food nutritional facts database. The data used in this research was obtained from free nutritional facts guide website. The data was taken from https://www.nutritionix.com/. This website is an interactive nutrition tools and world-renowned nutrition database that helps millions of consumers to understand nutrition every single day. It also provides a track mobile application named 'Nutrionix' that developed by a team of registered dietitians. This application offers many features such as uses state of the art natural language technology to make it quick and easy to track what consumer has eat. Moreover, it can calculate daily calories by estimate how many calories that consumer should consume each day to maintain or lose weight.

Next, the data also obtained from https://www.myfitnesspal.com/. The website is responsible for taking control of the consumer's goals to track calories and break down ingredients. It also has their own application named MyFitnessPal. The method of obtaining the data is look up the nutrition information for virtually any food by using their search engine. After look up to any specific food, we can see many groups of nutritional facts of the particular food. Additionally, we can choose the serving size that we want based on the measure units available there from small portion to extra-big portion.

3.4. Data Description

All of the data is based on types of foods and their nutritional values which are calories, carbohydrates, protein, fat, cholesterol, sodium and sugar. Calorie is a unit of energy, often used as a measurement of the amount of energy that food provides. In term of science, according to Merriam-Webster, calorie is the amount of heat required at a pressure of one atmosphere to raise the temperature of one gram of water one degree Celsius that is equal to about 4.19 joules. Carbohydrates or carbs are sugar molecules. They contain hydrogen and oxygen in the same ratio as water (2:1). Next, protein is a nutrient found in food that is made up of many amino acids joined together, is a necessary part of the diet, and is essential for normal cell structure and function. Fat is any ester of fatty acids, or a mixture of such compounds. Fat is a nutrient that give people energy that are either saturated or unsaturated and most foods with fat have both types.

Cholesterol is a type of fat found in human's blood. The liver is responsible to makes cholesterol for human's body. Sodium is a mineral that occurs naturally in many of the foods. Sodium chloride or salt as common called is the most common type of sodium found in nature. And lastly, sugar is the generic name for sweet-tasting, soluble carbohydrates, many of which are used in food. Simple sugars, also called monosaccharides, include glucose, fructose and galactose.

In this study, 60 various of foods were chosen to run the analysis which are 10 types from fruits, noodles and seafood category. 9 types of foods from breads and 11 types of foods are from rice dishes category. All the foods were coded according to its types. The F code is for fruits, N code for noodles, S code for seafood, B code for breads, R code for rice dishes and D code for desserts.

4. Results and discussion

4.1. Methods in Cluster Analysis

To analyze the dataset, six stages model building methods in cluster analysis will be presented to conduct the analysis. The stages are stated as below:

i. Stage 1- objectives of the cluster analysis.
ii. Stage 2- research design of the cluster analysis.
iii. Stage 3- assumptions in cluster analysis.
iv. Stage 4- deriving clusters and assessing overall fit.
v. Stage 5- interpretations of the clusters.
vi. Stage 6- validating and profiling of the clusters.

4.1.1. Stage 1: Objectives of the cluster analysis

This procedure begins in this study by examining a set of data regarding the similarity of food nutrition, with the goal of classifying all foods into a different number of clusters.

4.1.2. Stage 2: Research design of the cluster analysis

A dissimilarity matrix is a distance between two samples based on the same criterion. These matrices are calculated using Euclidean distances. Only 10x10 matrices are shown as below in Table x due to the large amount of data.

Table 1 Proximity matrix by using Euclidean distance

	1	2	3	4	5	6	7	8	9	10
1	0.000	2.828	7.810	2.000	5.099	4.000	2.000	5.099	3.162	2.000
2	2.828	0.000	5.568	2.000	3.464	3.742	2.000	3.742	2.449	2.000
3	7.810	5.568	0.000	6.856	5.385	5.745	6.708	6.557	6.586	6.708
4	2.000	2.000	6.856	0.000	5.292	3.162	2.828	5.477	3.742	2.828
5	5.099	3.464	5.385	5.292	0.000	6.164	3.162	1.414	2.449	3.162
6	4.000	3.742	5.745	3.162	6.164	0.000	4.472	6.928	5.657	4.472
7	2.000	2.000	6.708	2.828	3.162	4.472	0.000	3.162	1.414	0.000
8	5.099	3.742	6.557	5.477	1.414	6.928	3.162	0.000	2.000	3.162
9	3.162	2.449	6.856	3.742	2.449	5.657	1.414	2.000	0.000	1.414
10	2.000	2.000	6.708	2.828	3.162	4.472	0.000	3.162	1.414	0.000

4.1.3. Stage 3: Assumptions in cluster analysis

A sample of cases is usually acquired, and the cluster is then created to represent the structure of the entire population. The food type sample is considered to be the representative sample in this study.

4.1.4. Stage 4: Deriving clusters and assessing overall fit

The number of clusters will be set for this thesis, and the outcomes will be studied for each number of clusters. Iterative K-means clustering algorithms can be shifted from one cluster to the next until the decreasing inside cluster distances or the maximizing between cluster distances is achieved. Using SPSS software for K-means clustering, researchers can create cluster centers and allocate all of the items to clusters based on their minimal center distance.

4.1.4.1. Non-Hierarchical cluster analysis

- Number of clusters, *K*=2

Table 2 Initial cluster centers for *K*=2

Nutrition type	Cluster	
	1	2
Calorie	6	1
Carbohydrates	2	6
Protein	1	6
Fat	6	1
Cholesterol	6	2
Sodium	6	2
Sugar	1	6

Table 3 Iteration history for *K*=2

Iteration	Change in cluster centers	
	1	2
1	3.895	3.332
2	0.329	0.376
3	0.138	0.138
4	0.000	0.000

Table 4 Final cluster centers for *K*=2

	Cluster	
	1	2
Calorie	5	2
Carbohydrates	3	4
Protein	2	5
Fat	5	2
Cholesterol	5	3
Sodium	5	2
Sugar	4	4

Table 5 Distances between the final cluster centers for *K*=2

Cluster	1	2
1		5.971
2	5.971	

Table 6 Number of cases in each cluster for *K*=2

Cluster	1	31.000
	2	29.000
Valid		60.000
Missing		0

- Number of clusters, *K*=3

Table 7 Initial cluster centers for *K*=3

Nutrition type	Cluster		
	1	2	3
Calorie	5	6	1
Carbohydrates	1	1	6
Protein	4	1	5
Fat	4	6	1
Cholesterol	1	6	6
Sodium	2	6	4
Sugar	3	6	3

Table 8 Iteration history for *K*=3

Iteration	Change in cluster centers		
	1	2	3
1	2.852	2.715	3.346
2	0.511	0.606	0.766
3	0.389	0.339	0.472
4	0.000	0.000	0.000

Table 9 Final cluster centers for *K*=3

	Cluster		
	1	2	3
Calorie	3	5	2
Carbohydrates	3	3	5
Protein	4	2	5
Fat	3	5	2
Cholesterol	2	6	3
Sodium	2	5	2
Sugar	4	4	3

Table 10 Distances between the final cluster centers for *K*=3

Cluster	1	2	3
1		5.510	3.635
2	5.510		6.573
3	3.635	6.573	

Table 11 Number of cases in each cluster for *K*=3

Cluster	1	12.000
	2	28.000
	3	20.000
Valid		60.000
Missing		0

- Number of clusters, *K*=4

Table 12 Initial cluster centers for *K*=4

Nutrition type	Cluster			
	1	2	3	4
Calorie	6	5	1	2
Carbohydrates	1	1	6	5
Protein	1	4	6	3
Fat	6	4	1	1
Cholesterol	6	1	2	6
Sodium	6	2	2	5
Sugar	6	3	6	1

Table 13 Iteration history for *K*=4

Iteration	Change in cluster centers			
	1	2	3	4
1	2.795	2.439	2.740	3.232
2	0.223	0.839	0.409	0.503
3	0.000	0.000	0.000	0.000

Table 14 Final cluster centers for *K*=4

	Cluster			
	1	2	3	4
Calorie	5	4	2	3
Carbohydrates	2	2	5	4
Protein	2	4	5	3
Fat	5	3	2	3
Cholesterol	6	3	2	6
Sodium	5	2	2	3
Sugar	4	4	4	3

Table 10 Distances between the final cluster centers for *K*=3

Cluster	1	2	3	4
1		5.387	8.335	4.549
2	5.387		4.317	4.620
3	8.335	4.317		5.212
4	4.549	4.620	5.212	

Table 11 Number of cases in each cluster for *K*=3

Cluster	1	19.000
	2	10.000
	3	16.000
	4	15.000
Valid		60.000
Missing		0.000

4.1.5. Stage 5: Interpretation of clusters

K=4 is picked as the best solution to explain the number of clusters. It displays the total number of items in each cluster.

4.1.6. Stage 6: Validating and profiling of the clusters

It is critical to use this method to run all feasible tests to confirm the cluster solution's validity and to ensure that the cluster solutions are effective. The profiling stage entails describing each cluster's attributes in order to explain how they differ from one another. These procedures begin as soon as the clusters are detected. The clusters have all been discovered and characterized.

4.2. Dendrogram

A Ward’s linkage has been used in average linkage (between groups) in hierarchical clustering to obtain a dendrogram. Based on their respective dissimilarities distance, the groupings of calories composition are shown in a dendrogram.

Conclusion

Cluster analysis is a technique for grouping similar observations based on the observed values of numerous variables for each individual. Cluster analysis is conceptually related to discriminant analysis. In the latter, the group membership of a sample of observations is known in advance, but in the former, it is unknown for any observation. This study's dataset contains 60 foods, each of which has a successful analysis. It has been discovered that four clusters are a successful result for the number of clusters. Because there are only 3 iterations for *K*=4, this is the case. When *K*=2 and *K*=3, they have four iterations, which is not the ideal number to choose as the best findings to evaluate. This can be accomplished by determining the smallest number of iterations for each cluster.

Acknowledgement

I thank Allah swt for giving strength and patience to complete this. Also, I thank to my final year project report supervisor, Dr Muhammad Fauzee Hamdan offering advice and encouragement.

References

[1] Alin, A. (2010). Minitab. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(6), 723-727.

[2] Corpet, F. (1988). Multiple sequence alignment with hierarchical clustering. *Nucleic acids research*, *16*(22), 10881-10890.

[3] da Silva Torres, E. A. F., Garbelotti, M. L., & Neto, J. M. M. (2006). The application of hierarchical clusters analysis to the study of the composition of foods. *Food chemistry*, *99*(3), 622-629.

[4] de Micheaux, P. L., Drouilhet, R., & Liquet, B. (2013). The R software. Springer.

[5] Erdoğan, Ş. Z., & Timor, M. (2005). A data mining application in a student database. *Journal of aeronautics and space technologies*, *2*(2), 53-57.

[6] Ganguly, P (2020, Jun 26). Goals and Applications of Cluster Analysis. *Data Brio Academy*.

[7] Henry, D. B., Tolan, P. H., & Gorman-Smith, D. (2005). Cluster analysis in family psychology research. *Journal of Family Psychology*, *19*(1), 121.

[7] Jauhiainen, S., Pohl, A. J., Äyrämö, S., Kauppi, J. P., & Ferber, R. (2020). A hierarchical cluster analysis to determine whether injured runners exhibit similar kinematic gait patterns. *Scandinavian Journal of Medicine & Science in Sports*, *30*(4), 732-740.

[8] Kumar, A., Sinha, R., Bhattacherjee, V., Verma, D. S., & Singh, S. (2012, March). Modeling using K-means clustering algorithm. In *2012 1st International Conference on Recent Advances in Information Technology (RAIT)* (pp. 554-558). IEEE.

[9] Mehar, A. M., Matawie, K., & Maeder, A. (2013, December). Determining an optimal value of K in K-means clustering. In *2013 IEEE International Conference on Bioinformatics and Biomedicine* (pp. 51-55). IEEE.

[10] Mirkin, B. (2011). *Core concepts in data analysis: summarization, correlation and visualization*. Springer Science & Business Media.

[11] Mishra, B. K., Rath, A., Nayak, N. R., & Swain, S. (2012, August). Far efficient K-means clustering algorithm. In Proceedings of the International Conference on Advances in Computing, Communications and Informatics (pp. 106-110).

[12] Morissette, L., & Chartier, S. (2013). The k-means clustering technique: General considerations and implementation in Mathematica. Tutorials in Quantitative Methods for Psychology, 9(1), 15-24.

[13] Schonlau, M. (2002). The clustergram: A graph for visualizing hierarchical and nonhierarchical cluster analyses. *The Stata Journal*, *2*(4), 391-402.

[14] Serrao, M., Chini, G., Bergantino, M., Sarnari, D., Casali, C., Conte, C., ... & Marinozzi, F. (2018). Identification of specific gait patterns in patients with cerebellar ataxia, spastic paraplegia, and Parkinson’s disease: A non-hierarchical cluster analysis. *Human movement science*, *57*, 267-279.

[15] Sharif, S. M., Kusin, F. M., Asha’ari, Z. H., & Aris, A. Z. (2015). Characterization of waterquality conditions in the Klang River Basin, Malaysia using self organizing map and K-means algorithm. *Procedia Environmental Sciences*, *30*, 73-78.

[16] Wohwe Sambo, D., Yenke, B. O., Förster, A., & Dayang, P. (2019). Optimized clustering algorithms for large wireless sensor networks: A review. *Sensors*, *19*(2), 322.

[17] Wu, J. (2012). Cluster analysis and K-means clustering: an introduction. In *Advances in K-means Clustering* (pp. 1-16). Springer, Berlin, Heidelberg.

[18] Zahra, S., Ghazanfar, M. A., Khalid, A., Azam, M. A., Naeem, U., & Prugel-Bennett, A. (2015). Novel centroid selection approaches for K-Means-clustering based recommender systems. *Information sciences*, *320*, 156-189.

[19] Zhao, Y., Karypis, G., & Fayyad, U. (2005). Hierarchical clustering algorithms for document datasets. *Data mining and knowledge discovery*, *10*(2),