Project Proposal

Theveenan Nirmalan (20898397) & Saif Abuosba (20896949)

MSCI 446

Prof. Lukasz Golab

2023-02-14

General Goal of Project:

For our machine learning project, our goal is to use an existing dataset and perform clustering using unsupervised machine learning methods to obtain different insights from the dataset. There are no specific insights we hope to find, rather we plan to use parameter tuning to explore what possible insights can be gathered from the raw data. The dataset we plan to perform unsupervised learning on contains raw data on nutritional metrics of many different foods (including calories, carbohydrates, sugars etc.).

Specific Problem for Study:

The problem we are tackling with this project relates to classifying and determining relationships between different foods based on their nutritional metrics. Specifically, we are hoping to determine relationships and classifications between foods that may break previous misconceptions within the topic of nutrition, or find insights that pose new questions which could potentially incite further research and discussion within the world of nutritional science.

This topic of classifications and relationships of food based on nutritional metrics is extremely important and relevant to all of human life (and also many species of non-human life on earth), since this data we are working with relates to nutrition which is of high significance to their biology and overall well being. The decisions humans make regarding nutritional decisions can have long lasting impacts on human life, so it is critical to be well informed. This is also why many researchers around the world regularly study nutritional sciences for new discoveries.

It is very common to see previous assumptions regarding nutrition disproven, and to see new claims regarding nutrition created, which is all due to the fact that nutritional science is a very complex and layered topic. As machine learning students, we may not be able to make discoveries at the scientific level regarding foods since we are not qualified in that domain - although we are able to use data regarding these foods to identify relationships and groupings that may provide new insights to further create dialogue within the nutritional community.

Dataset Description:

The dataset that will be used for this project is available publicly online and is provided in the format of a csv file (link to dataset: https://corgis-edu.github.io/corgis/csv/food/) . The dataset comes from the ‘CORGIS Datasets Project’ which is described to be ‘The Collection of Really Great, Interesting, Situated Datasets’. The specific data in the dataset was sourced from the ‘United States Department of Agriculture’s Food Composition Database’.

The dataset consists of 7083 records/rows and provides information on a mix of foods and beverages. There are 38 columns/features present in the dataset. The features of the dataset include both numerical and categorical data which are represented in the form of strings for categorical data, and integers and floats for numerical data.

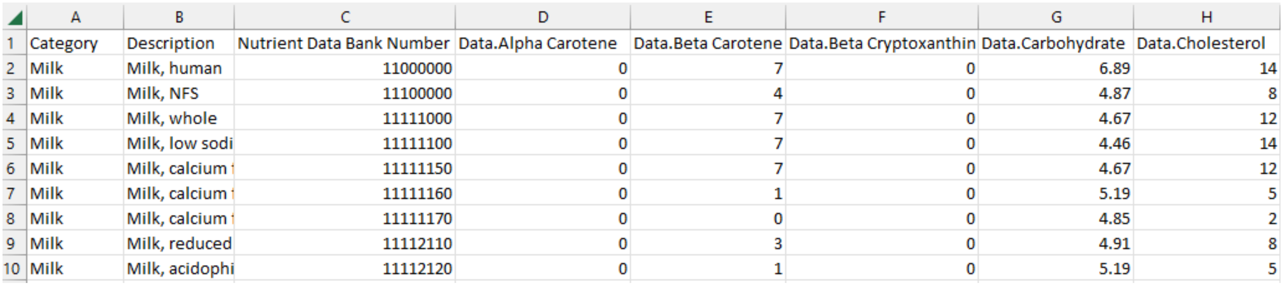
List of features with data type and example value

Feature	Type	Example Value
Category	Categorical - String	"Milk"
Description	Categorical - String	"Milk, human"
Nutrient Data Bank Number	Numerical - Integer	11000000
Data.Alpha Carotene	Numerical - Integer	0
Data.Beta Carotene	Numerical - Integer	7
Data.Beta Cryptoxanthin	Numerical - Integer	0
Data.Carbohydrate	Numerical - Float	6.89
Data.Cholesterol	Numerical - Integer	14
Data.Choline	Numerical - Float	16
Data.Fiber	Numerical - Float	0
Data.Lutein and Zeaxanthin	Numerical - Integer	0
Data.Lycopene	Numerical - Integer	0
Data.Niacin	Numerical - Float	0.177
Data.Protein	Numerical - Float	1.03
Data.Retinol	Numerical - Integer	60
Data.Riboflavin	Numerical - Float	0.036
Data.Selenium	Numerical - Float	1.8
Data.Sugar Total	Numerical - Float	6.89
Data.Thiamin	Numerical - Float	0.014

Data.Water	Numerical - Float	87.5
Data.Fat.Monosaturated Fat	Numerical - Float	1.658
Data.Fat.Polysaturated Fat	Numerical - Float	0.497
Data.Fat.Saturated Fat	Numerical - Float	2.009
Data.Fat.Total Lipid	Numerical - Float	4.38
Data.Major Minerals.Calcium	Numerical - Integer	32
Data.Major Minerals.Copper	Numerical - Float	0.052
Data.Major Minerals.Iron	Numerical - Float	0.03
Data.Major Minerals.Magnesium	Numerical - Integer	3
Data.Major Minerals.Phosphorus	Numerical - Integer	14
Data.Major Minerals.Potassium	Numerical - Integer	51
Data.Major Minerals.Sodium	Numerical - Integer	17
Data.Major Minerals.Zinc	Numerical - Float	0.17
Data.Vitamins.Vitamin A - RAE	Numerical - Integer	61
Data.Vitamins.Vitamin B12	Numerical - Float	0.05
Data.Vitamins.Vitamin B6	Numerical - Float	0.011
Data.Vitamins.Vitamin C	Numerical - Integer	5
Data.Vitamins.Vitamin E	Numerical - Float	0.08
Data.Vitamins.Vitamin K	Numerical - Float	0.3

Example values and data types sourced from: https://corgis-edu.github.io/corgis/csv/food/

Sample rows from data set (not all features present)



	A	B	C	D	E	F	G	H
1	Category	Description	Nutrient Data Bank Number	Data.Alpha Carotene	Data.Beta Carotene	Data.Beta Cryptoxanthin	Data.Carbohydrate	Data.Cholesterol
2	Milk	Milk, human	11000000	0	7	0	6.89	14
3	Milk	Milk, NFS	11100000	0	4	0	4.87	8
4	Milk	Milk, whole	11111000	0	7	0	4.67	12
5	Milk	Milk, low sodi	11111100	0	7	0	4.46	14
6	Milk	Milk, calcium	11111150	0	7	0	4.67	12
7	Milk	Milk, calcium	11111160	0	1	0	5.19	5
8	Milk	Milk, calcium	11111170	0	0	0	4.85	2
9	Milk	Milk, reduced	11112110	0	3	0	4.91	8
10	Milk	Milk, acidophi	11112120	0	1	0	5.19	5

From the first 9 rows of the dataset shown above, it is evident that there are multiple variants for the same food/beverage. For example, there are nutritional facts for:

- Milk, low sodium, whole
- Milk, reduced fat (2%)
- Milk, low fat (1%)

We will present why this may pose a problem for our algorithm in the next section.

Machine Learning Methods:

With our data, the first step we plan to take is to somehow reduce the amount of data points that represent variants of the same food 'category'. **If we don't do this step, this could potentially create a conflict for our algorithm as it might cluster the dataset purely based on the 'category' of each record if we were to use k-means clustering with k ≈ # of distinct categories**. In other words, the reason we want to do this is because when we go to cluster our foods later on, we want foods to be 'equally weighted' during the clustering process. Otherwise we may see situations where all the variants of a single food become their own cluster even though we may want that food to be clustered with other foods. For example, we may want to see oranges clustered with other fruits such as apples, bananas, and pineapples instead of our model unwantedly creating a cluster of all different types of oranges (tangerines, fat free oranges, etc). Although we also plan to perform our operations on the raw dataset (prior to reducing data points) so that we can see how our data reduction has affected the final clustering results.

Next, we plan to first perform principal component analysis (PCA) to reduce the dimensionality of our data since our dataset has many features and some of these features may provide little value to our goals.

Then, the main method we plan to use for our clustering solution is K-means clustering. Since we are doing exploratory data analysis, we will try using different values for K and other parameters to obtain different clustering results. Also, since there are many features available, we will try our K-means clustering using different features to see how clustering for different features may provide different insights. Our features will include all columns except for the ‘Category’, ‘Description’, and ‘Nutrient Data Bank Number’ columns. All of these features are numeric and so we may not need to perform extra tasks which would’ve been required if we were performing K-means on categorical variables. With our K-means clustering, we hope to have all of our data points representing different foods clustered into groups which represent different attributes. For example, with K=2, we may expect to have two clusters of “healthy” and “unhealthy” foods, and for K=4, we may expect to have 4 clusters for “Meats”, “Dairy”, “Fruits/Veggies”, and “Grain”.

This is how we plan to apply unsupervised machine learning methods to our project.

Bibliography:

- Whitcomb, R., Choi, J. M., & Guan, B. (2021, October 15). *Food CSV file*. CORGIS Datasets Project. Retrieved February 6, 2023, from https://corgis-edu.github.io/corgis/csv/food/