Agrupamento de Alimentos por Características Nutritivas

Aluno: Semar Augusto da Cunha Mello Martins

Introdução:
Com o crescimento do “mundo fitness” ao redor do mundo, as pessoas estão se tornando mais preocupadas com os alimentos que elas ingerem. Será possível agrupar alimentos de acordo com os macro e micronutrientes deles? Será que, com as características que existem na base de dados é possível colocar “alface, tomate e brócolis” em um grupo, “chocolate, doritos e hamburguer” em outro? Ou será que os alimentos estão distribuídos de forma mais ampla do que nós pensamos usualmente?

Base de Dados:
A base de dados escolhida foi a base de dados “Open Food Facts”, disponível no Kaggle. É uma base de dados com mais de 70 features e cerca de 333 mil entradas. Dentre as features há desde características sobre o “criador” do alimento, onde ele foi produzido, a forma em que ele foi embalado, a loja em que ele é vendido, cidades onde ele é vendido, até as características que queremos avaliar, que são a quantidade de macronutrientes (carboidratos, proteínas e gorduras) em 100g, quantidade de micronutrientes (vitaminas e minerais) em 100g, se há ou não aditivos na comida, se há álcool na comida e o score da comida de acordo com o governo francês e o governo inglês.

Modelagem:
Os dados serão lidos e colocados em uma matriz, em que cada linha é uma entrada dos dados e cada coluna é uma feature.

Planejamento:
Para a implementação do TP serão necessárias 3 grandes etapas: limpeza, agrupamento e avaliação.
Com relação à limpeza dos dados, será necessário retirar as features com que não nos importamos, uma vez que queremos apenas grupos de alimentos com relação a macronutrientes e micronutrientes. Com isso, podemos retirar features como quem é o criador, qual a marca do alimento, entre outras precisaremos fazer a limpeza com relação a dados que estão faltando, pois existem algumas entradas que não possuem informações em algumas features.

Durante a fase de agrupamento, a ideia inicial é usar um algoritmo de k-means.

Para a avaliação usaremos medidas como pureza, entropia e coeficiente de silhueta