Identifying whole food allergens using Machine Learning

Yuthika Shekhar and Abhijeet Talaulikar
32 million Americans have food allergies.
1 in 10 Adults
1 in 13 Children
8 MAJOR FOOD ALLERGENS

- MILK
- TREE NUTS
- EGGS
- PEANUTS
- FISH
- WHEAT
- SHELLFISH
- SOYBEANS

Source: https://www.fda.gov/food/food-labeling-nutrition/food-allergies
Objective

- The problem with food labels
- Raw foods vs whole foods
- Impossible to make exhaustive list of allergens
- Absence of hierarchical allergen ontologies

This would assist the FDA in regulating the labels on food products.
Data Sources

Composition of Foods Raw, Processed, Prepared USDA National Nutrient Database for Standard Reference

The U.S. Department of Agriculture (USDA) National Nutrient Database for Standard Reference, contains data on 7,793 food items and up to 150 nutrients.

Open FDA Substance Data

Get Unique Ingredient Identifier (UNII) codes for substances and their synonyms based on substance’s molecular structure. This is generated through a joint effort of FDA and GSRS.

FALCPA food allergen list

This contains data of 8 major food allergens and their derivatives.
Methodology

- Data Collection & Cleaning
- Data Annotation
- Modeling
Data collection & cleaning

- Map constituents to their UNII codes in USDA nutrient database.
- For 150 nutrients, the dataset has 150 one-hot encoded columns for presence of ingredients.
Data Annotation

From the 8 major allergens, find their derivatives that are a part of the USDA food database.

Using fuzzy string match with Levenshtein distance

Label derivatives as Potential Allergens and others as Safe.
Modelling

• With around 8000 observations, 150 features and 1 binary label, we can use a classifier like SVM or gradient boosting to label new foods as potential allergens or safe.

• Find multidimensional association rules with high support and confidence that link nutrients to the label.
Expected Outcomes: Allergen Ontologies

- Thai Peanut Sauce
  - Peanut Butter
    - Peanuts
      - ARA H1 Allergen
    - Salt
    - Sugar
    - Proteins
    - Peanut Oil
      - ARA H2 Allergen
We address the following areas with the solution

- Product Safety Surveillance
- Artificial Intelligence
- Empowering patients and consumers to make better-informed decisions.
References

- https://open.fda.gov/apis/other/substance/
- https://www.fda.gov/food/food-labeling-nutrition/food-allergies
- https://precision.fda.gov/uniisearch