Food Nutrient Classifier — Text Mining

Alexander Ivanov
me@sasheto.net

Andrei Marculesteanu
andreimarculesteanu28@gmail.com

Michael McLean
michael.mclean@student.auc.nl

Divyaansh Sethia
sethiadivyansh@gmail.com

Abstract

We attempt to classify foods from the USDA Foundation Foods and Branded Foods databases based on their nutrient composition using Random Forest and kNN models. With a limited set of 7 food categories, we find that both models perform better with a hand-picked subset of 23 nutrients than with all nutrients. The classifiers were most accurate at predicting Milk and Dairy Products and Meats and Meat Products. The most common errors across all models were misclassification of Nuts and Seed Products and classification of Fruits as Vegetables.

1 Introduction

Despite the current explosion of machine learning algorithms into the mainstream, most of the attention has been focused on chatbots. In such a climate, other spheres which may also benefit from their inclusion tend to be undermined. However, recent advances in machine learning have also enabled the development of classification algorithms tailored to nutritional applications, allowing researchers to categorize foods based on their nutrient profiles rather than relying solely on manual or rule-based approaches. Supervised methods such as **k-Nearest Neighbors (kNN)**, **Random Forests**, and **Support Vector Machines (SVMs)** have been successfully applied to large nutrient databases to uncover complex, nonlinear patterns among macronutrients and micronutrients, facilitating tasks such as predicting processing levels or identifying foods with similar nutritional impacts (Kirk et al., 2022). By leveraging these algorithms, nutrition scientists can perform precision nutrition analyses at scale, which yields insights that can inform public health recommendations and personalized dietary planning (Zhaoping et al., 2023).

Our motivation for employing machine learning in food classification stems from two key observations in our datasets. First, although each item may list over 50 nutrient attributes, many of these are frequently null or uninformative. Other papers have discussed the impact of dropping sparsely populated nutrients and have shown that markedly improved model performance (Nalin et al., 2024). Second, the original category labels (28 in the **Foundation Foods** and 350 in the **Branded Foods**) often contain very small classes that are difficult for classifiers to learn. By remapping to seven broader, more distinct categories, we aim to reduce class imbalance and enhance overall accuracy.

Accordingly, our research question evolved from "Machine learning algorithms to recommend alternative foods with similar nutritional values" to "**Which nutrients are most informative for classifying foods?**", reflecting a shift from a recommendation task to a feature-selection and classification task better aligned with the data's structure and our capacity to extract meaningful nutrient predictors. The report will proceed as follows: in the **Research Context** section we will discuss our projects relation to the field of text mining and its broader contexts. In the **Methodology** section we will outline the transformations performed on the dataset, as well as the algorithms used and the motivation behind their usage. In the **Results** section we will discuss the results of the classifiers and in the **Discussion** section we will discuss these results more in-depth, as well as our limitations. Finally, in the **Conclusion** section we will give a brief conclusion of our findings and discuss further avenues for exploration. The source code for our project is available at github.com/mfmcl/tm-project.

2 Research Context

Text mining techniques are becoming increasingly prevalent in the food and nutrition domain, especially in the field of genetics. Other relevant uses can be found in tasks such as recipe recommendation, food safety monitoring, and nutrient-based classification. Similarly, text mining methods based on **bag-of-words** have been applied to nutrition literature to quantify the prevalence of statistical methods and software in dietetics research, demonstrating the technique's power to process large corpora of scientific articles and derive practice-relevant trends (Alison et al., 2021).

Our research question aims at identifying the most informative nutrients for food classification, which is intrinsically connected to core text-mining paradigms such as feature selection and dimensionality reduction. In text mining, selecting a subset of terms (features) that best distinguish document categories is vital. Similarly, we seek to discover which nutrient attributes serve as the "keywords" that most effectively differentiate food groups. By adapting these principles to structured nutrient data, we extend text-mining methodologies into the realm of nutritional epidemiology and food informatics.

3 Methodology

3.1 Data Acquisition

The datasets used in our project are from the USDA FoodData Central website which contains several datasets with different characteristics. For our purposes we chose the **Foundation Foods (FF)** and **Branded Foods (BF)** datasets, since our original idea was to train a classifier on the **FF** dataset and use the **BF** dataset as a test set for classifying the category and possibly the specific food item. The **FF** dataset comprises 340 minimally processed food items across 28 original categories, while the **BF** dataset includes approximately 409 000 branded products classified into about 350 categories.

3.2 Data Transformation and Feature Preparation

The initial datasets came with a lot of metadata and unnecessary information about the food products, such as ingredients which were not needed for our investigation. Thus, we filtered the data set to 4 main properties: *foodType*, *foodCategory*, *foodName* and an array containing the *nutrients* which included their *name*, *unit* of measurement and *amount*. We removed food items with portions measured in *ml* and nutrients measured in units not directly convertible to *mg* (namely *kcal*, *kJ* and *sp gr*) Subsequently, we scaled the remaining nutrients to *g*. Nutrient values were already standardized to amount per 100g in the original dataset. Next, we looked at the items of the individual categories and decided to drop the *Sweets* category since it only had 1 member: *sugar*. We looked at the rest of the food items in the **FF** data set and decided to do the following category remapping:

- We dropped the following categories, since they had low membership: *Baked Products*, *Spices and Herbs*, *Fats and Oils*, *Soups, Sauces, and Gravies*, *Restaurant Foods*, *Sweets* and *Beverages*.
- We remapped all categories pertaining to animal meat products under the common category: *Meats and Meat Products*

For mapping the 350 categories of **BF** into the 7 categories we had left, we used the **ChatGPT LLM** in the following exchange: **ChatGPT Conversation Link**. We looked over its mapping and made any desired changes manually.

Finally, for our features we took the nutrients of each food item and replaced all null values with 0. We also dropped all nutrients that weren't in the intersection of nutrients between **BF** and **FF** before training. This left for a total of 68 nutrients to act as features.

3.3 Algorithms and Libraries

For our algorithms we first chose to use a **Random Forest (RF)**. This technique was chosen due its robustness to high-dimensional data, capacity to model nonlinear relationships, and built-in measures of feature importance. The characteristics make it ideal for uncovering which nutrients drive classification accuracy.

After implementing the first algorithm we decided to also build a **k-Nearest Neighbors (kNN)** classifier, so that we have a comparison. This method was selected for its simplicity and effectiveness in clustering items with similar nutrient vectors, particularly advantageous once the feature space and categories were reduced and balanced.

The implementation of the algorithms was done using the classifiers provided by the **scikit-learn**

python library. Other libraries used, include: **numpy** and **pandas** for data manipulation and preprocessing and **matplotlib** and **seaborn** for exploratory data analysis and visualization.

4 Results

When trained to classify on all of the 68 nutrients and tested on the entirety of the **BF** data set: the **RF** model got an accuracy of **0.4830** and the **kNN** model got an accuracy of **0.3203**. The confusion matrix for **RF** can be seen in (Figure 1) and the confusion matrix for **kNN** can be seen in (Figure 4).

Arsenault et al. discuss the best subset of nutrients for classifying the nutritional value of a food. We tried to see if the best subset for classifying nutritional value is also good at classifying the food category. Our nutrient subset based on Arsenault et al. is this:

- Protein
- Fiber, total dietary
- Fiber, soluble
- Fiber, insoluble
- Calcium, Ca
- Fatty acids, total monounsaturated
- Fatty acids, total polyunsaturated
- Vitamin C, total ascorbic acid
- Fatty acids, total saturated
- Sodium, Na
- Total Sugars
- Sugars, Total

When trained to classify on the optimal nutrients as discussed in Arsenault et al. and tested on the entirety of the **BF** data set: the **RF** model got an accuracy of **0.5198** and the **kNN** model got an accuracy of **0.4739**. The confusion matrix for **RF** can be seen in (Figure 2) and the confusion matrix for **kNN** can be seen in (Figure 5).

We also hand selected a subset of nutrients that extend the Arsenault et al. nutrients:

- Protein
- Fiber, total dietary
- Fiber, soluble
- Fiber, insoluble
- Calcium, Ca
- Fatty acids, total monounsaturated
- Fatty acids, total polyunsaturated
- Vitamin C, total ascorbic acid
- Fatty acids, total saturated
- Sodium, Na
- Total Sugars
- Sugars, Total
- Carbohydrate, by difference
- Magnesium, Mg
- Vitamin D (D2 + D3)
- Vitamin D3 (cholecalciferol)
- Potassium, K
- Glucose
- Riboflavin
- Vitamin B-6
- Biotin
- Fructose
- Cholesterol

When trained to classify on the hand selected nutrient subset and tested on the entirety of the **BF** data set: the **RF** model got an accuracy of **0.5688** and the **kNN** model got an accuracy of **0.5287**. The confusion matrix for **RF** can be seen in (Figure 3) and the confusion matrix for **kNN** can be seen in (Figure 6).

5 Discussion

When training on the entire set of 68 nutrients, the **RF** model was especially good at classifying Dairy and Egg Products. Both models were also very accurate at predicting Meats and Meat Products, which matched our expectation. Interestingly, the classifiers struggled with Fruits/Vegetables. This is in line with our own

confusion as to which category certain foods belong in (in a social context, despite the biological classification being clear). Both models also struggled with classifying Dairy and Egg products/Nut and Seed products/Fruits and Fruit Juices. Interesting to note is that **kNN** when trained on all nutrients gives a higher probability to the categories Legumes and Legume Products and Nut and Seed Products irregardless of the actual food category.

Using the subset of nutrients provided by Arsenault et al., the accuracy of the **RF** improved up to 51.98%, and the **kNN** accuracy increased to 48%. From the confusion matrices, we can see that Dairy and Egg Products are still predicted with the highest accuracy. This is likely due to the distribution of products in each category being skewed.

Interestingly, our hand-picked subset of nutrients improved the accuracy of both classifiers by roughly 5% compared to the accuracy of the classifiers trained on the Arsenault et al. subset. This is potentially due to the extra discriminatory information encoded in the added nutrients, for example a food high in cholesterol has a higher chance of being an egg due to eggs having notoriously high cholesterol.

All models were poor at classifying Fruits and Fruit Juices. All models except the **kNN** trained on all nutrients confused Fruits and Fruit Juices for Vegetables and Vegetable Products. The **kNN** trained on all nutrients instead confused Fruits and Fruit Juices for Legumes and Legume Products. This could be due to Fruits, Vegetables and Legumes all originating from plants, and so the relative distance between these three categories is smaller than between the three categories and other categories. The classifiers that confused Fruits and Fruit Juices for Vegetables and Vegetable Products coincidentally also perform well at classifying Vegetables and Vegetable Products correctly. The confusion of fruits for vegetables could thus arise from overfitting to classify vegetables correctly.

5.1 Limitations

We cannot conclude whether we have found the best subset of nutrients for food category classification as we only had time to check 3 subsets. Our best performing subset of hand-selected nutrients that extend those selected by Arsenault et al. could be extended further to encode more discriminatory information. We know however that adding more nutrients as features can make the accuracy lower, as we reported the lowest accuracy when we used the full set of 68 nutrients. As such future research could look into the optimal extension on our hand selected nutrient list.

Another limitation could be low training data, as the classifiers are trained on the **FF** data set, which after category filtering has a total number of 316 entries. We initially thought that we could only train on **FF** data due to the format of both datasets and how they presented categories. After some adjusting of the data we realized that grouping food items into 7 categories makes it possible for **BF** to be used to both train and test the classifiers, but by that point the scope of the project had been set with no time to change it. As such, future work could utilize the full scope of **BF**'s more than 400 000 items for both training and testing the classifiers.

6 Conclusion

Our experiments up to this point have shown that careful consideration of the nutrients used in the classification process can have a visible impact on the final results. Our current accuracy would be similar to a coin flip, however we have successfully shown that the choice of nutrients has a measurable impact on the final results. Also, the two different algorithms seem to present results which were expected, with the **RF** classifier being more accurate than the **kNN** model. Overall, we tend to see that a balanced number of nutrient features leads to a better classifier. Too little nutrients leads to losing helpful discriminatory information while too many nutrients leads to overfitting.

In terms of future improvements, one way would be to test on more subsets of nutrients as the ones we used may not be fully representative. Further improvements can be made to the dataset used as the categories are sometimes very weighted towards a few categories, especially in the **FF** dataset. Another interesting avenue would be to test more advanced machine learning algorithms such as neural networks. For the purpose of the project, we restricted ourselves to low-computational algorithms, such as **Random Forests**, however more advanced models would likely also yield better results.

References

[Kirk et al. 2022] Kirk, D., Kok, E., Tufano, M., Tekinerdogan, B., Feskens, E. J. M. and Camps, G. 2022. *Machine Learning in Nutrition Research* Advances in nutrition (Bethesda, Md.)

[Zhaoping et al. 2023] Zhaoping L., Shavawn F., Emily J., David H. 2023. *Perspective: A Comprehensive Evaluation of Data Quality in Nutrient Databases* Advances in nutrition (Bethesda, Md.)

[Nalin et al. 2024] Nalin A., Sumit B., Riya D., Ganesh B. 2024. *Machine learning and natural language processing models to predict the extent of food processing*

[Alison et al. 2021] Alison C., Marijka J. B., Eleanor J. B. 2021. *Perspective: A Comprehensive Evaluation of Data Quality in Nutrient Databases* Advances in nutrition (Bethesda, Md.)

[Arsenault et al. 2012] Arsenault J. E., Fulgoni V. L., Hersey J. C., Muth M. K. 2012. *A Novel Approach to Selecting and Weighting Nutrients for Nutrient Profiling of Foods and Diets* Journal of the Academy of Nutrition and Dietetics

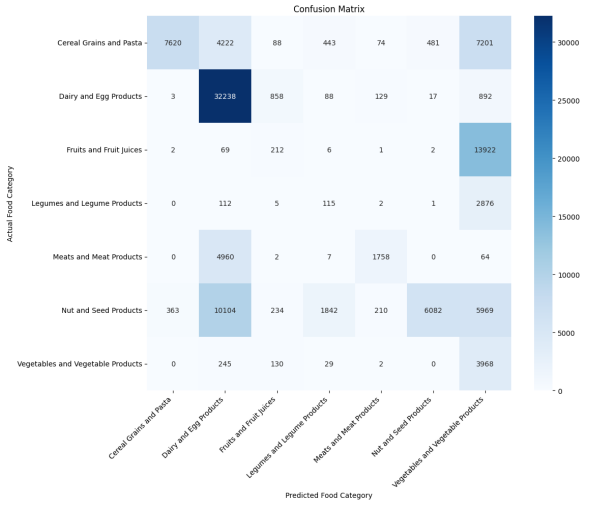

Figure 1: **RF** Confusion Matrix 68 Nutrients

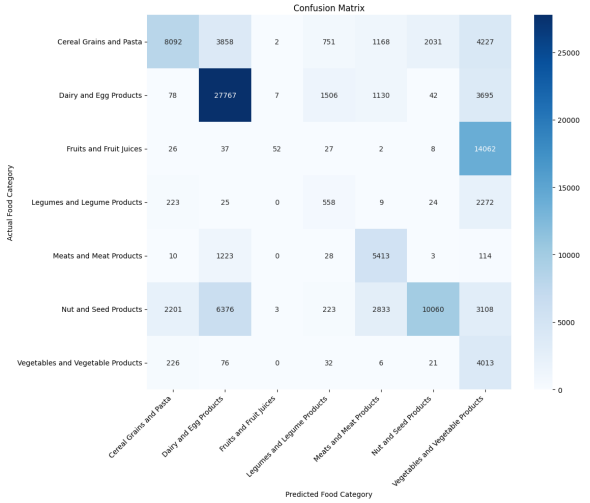

Figure 2: **RF** Confusion Matrix Arsenault et al. Nutrients

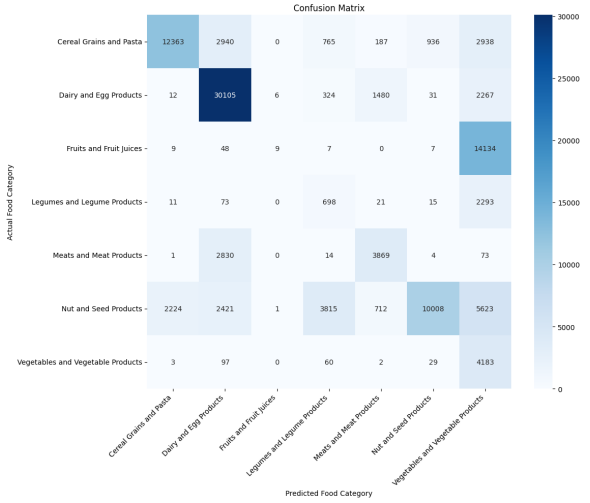

Figure 3: **RF** Confusion Matrix Hand Selected Nutrients

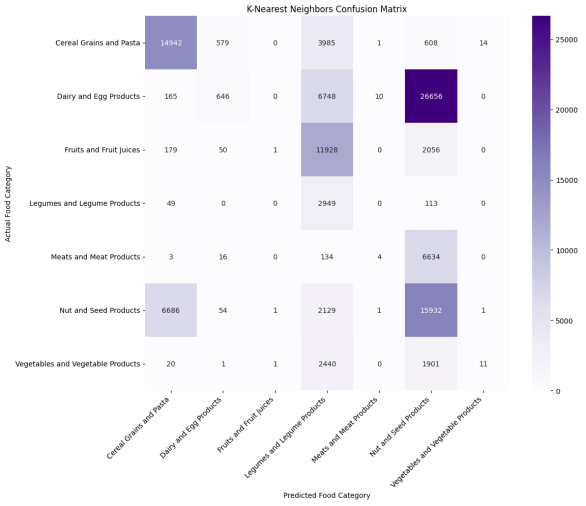

Figure 4: **kNN** Confusion Matrix 68 Nutrients

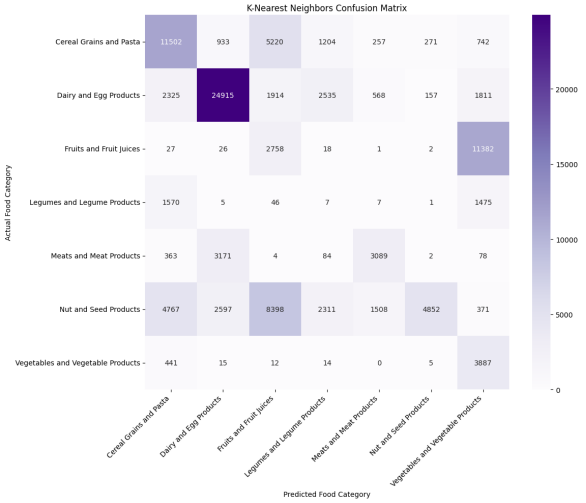

Figure 5: **kNN** Confusion Matrix Arsenault et al. Nutrients

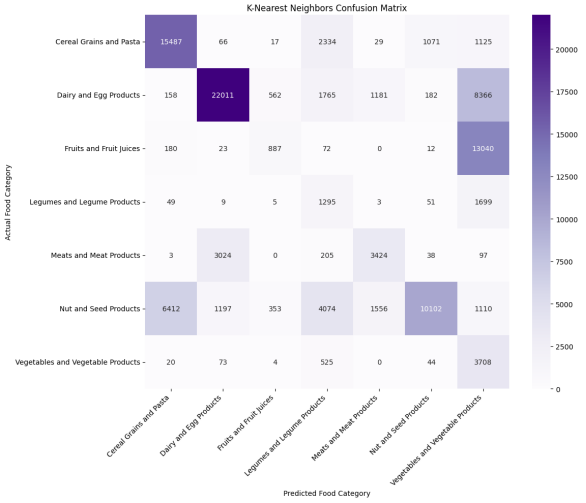

Figure 6: **kNN** Confusion Matrix Hand Selected Nutrients