Prem Panchal

Final Report

Food Recommender for Health

Problem Statement

Eating food is one of the most important things people do throughout their lives. While it is necessary at a survival level, food has become something much bigger than that; there are millions of recipes and all kinds of food in this world. Despite there being so much food, a large majority of it is unhealthy and overwhelming in calories or other health considerations. One's diet is a major contribution to their health, and it can be very difficult to manage one's diet such that they have a balanced one. Even when people do find some healthy food options that they enjoy, it is very hard to keep up a variety of these kinds of food making it hard to keep the habit going.

By using this model, one can categorize their food in 3 categories for all of carbohydrates, fat, and protein and find multiple dishes with their required dietary needs. Additionally, people can determine what regional cuisines are best met for their needs so more research can be done on that regional food.

Data Wrangling

When obtaining the data the first step I had to take was to confirm how the data was organized. The data came with multiple data files for a variety of diet types, as well as a file called 'all_diets' and I first checked to confirm that the all_diets.csv file contained all the combined data of the other individual files.

After confirming the data set had all the necessary rows I removed the 'extraction_day' column as it contained no relevant data with only 1 date for all 7806 entries; additionally I cleaned the column names to be easier to work with. After that I used the ratio of 4:4:9 for carbs:proteins:fats to calculate the total calories for each row as that could be valuable later in my model. Even from a quick glance of the data, I could see that there was some issues with the data, mostly being that some recipes seem to be based on the whole item while some were based on a serving size. To account for this later in my EDA , I created 4 categories for each nutrition: low, moderate, high, very high.

Exploratory Data Analysis

One of the first things noticed when performing EDA was that there are items with a value of 0 within the nutritional groups of fat,carbs, and proteins. I decided to keep these in the beginning as there are definitely cases where one can make food with 0 content for them but it was a concern at first. There is also an uneven split of cuisine types in the dataset, with American and Mediterranean cuisine having significantly more items in the data than the other 17 cuisines as shown in the figure below.

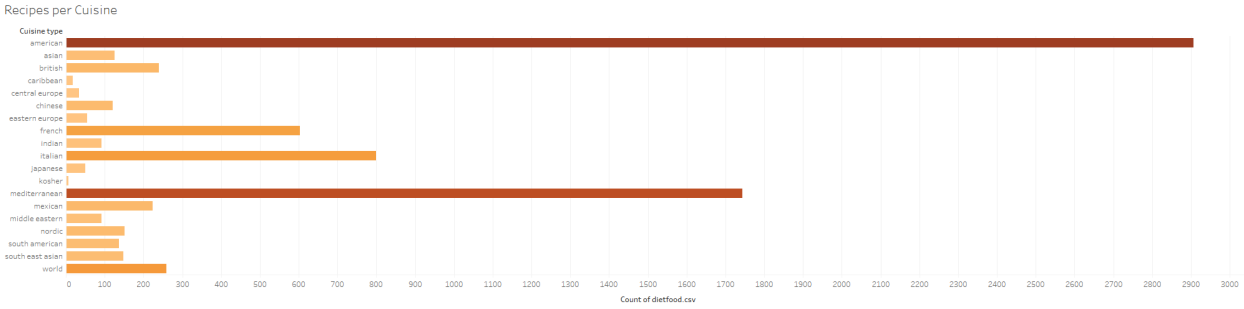

Figure 1

The next step in EDA involved looking at the inconsistencies in the distribution of the data. As seen in figure 2 and 3 below, there is a wide distribution on the right side for all 3 nutritions the comes from a very small amount of data being well into the thousands for these recipes. This large amount of outliers would cause a lot of issues down the line and needed to be dealt with. Since there were so many outliers, the mean could not be used as a method of removing overly large values, and I decided to remove rows by ensuring the data was below the median of each column added to 1.5 times the standard deviation.

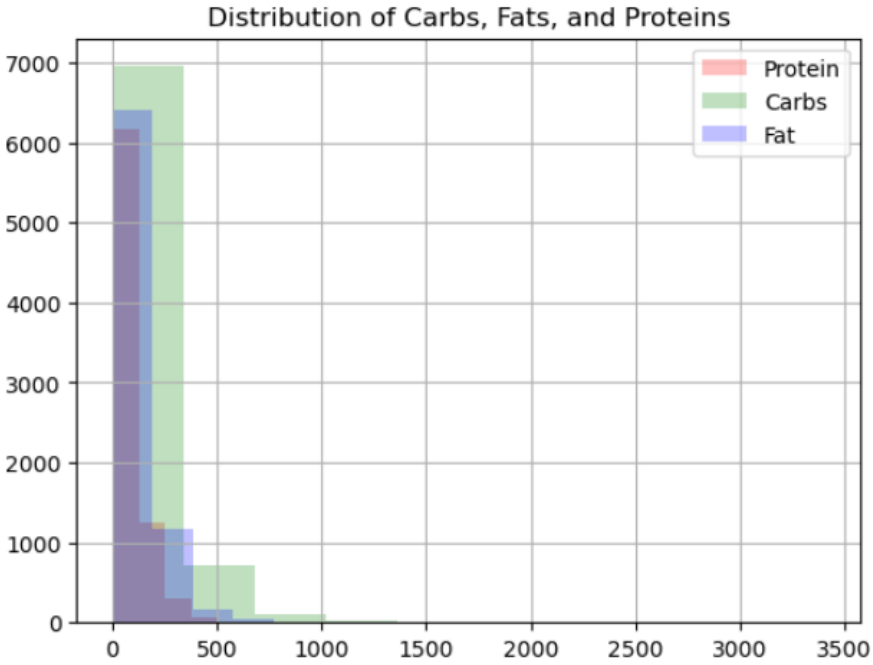

Figure 2

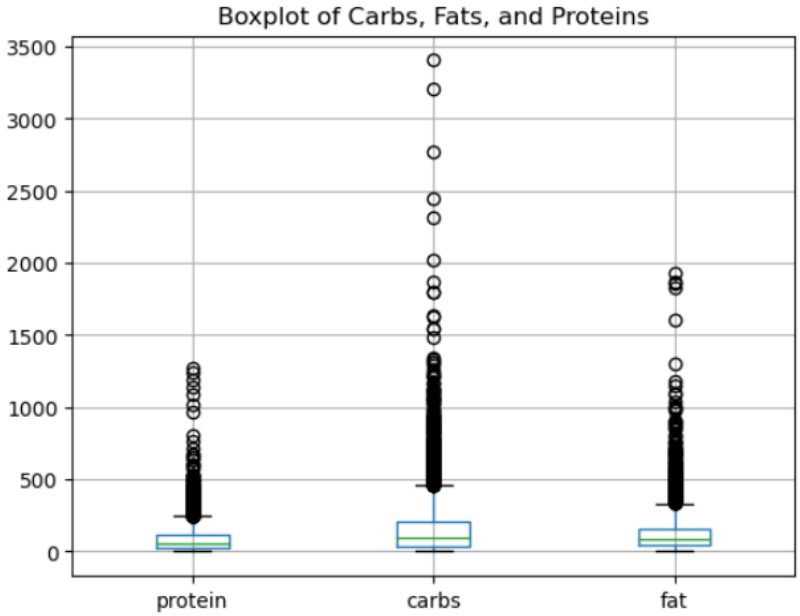

Figure 3

At this point the data has gone from 7806 data points to 4669. After some consideration on the fact that this dataset is being used for healthy means, I had removed the very high category entirely from the data; Protein values that were very high were just marked as high as protein is just healthy for you and there are no low protein diets, while very high values were removed from both carbs and fat. Additionally I had removed all 0 values from fat as well dropping the dataset down to 2853 data points.

Model Selection

When determining which model to use, there was a large question on what I wanted to do with the dataset at the time. After a lot of deliberation, I had concluded that I would like to create a recommendation system with the data as a way to spread the kinds of food. After discussion with others, I concluded I would manually create a Naives-Bayes model as it was not reasonable to create the model through a prebuilt function.

When creating the model, I had to create the mathematical functions for the NB classifier and from there had to decide what exactly was going to be recommended. Initially I had considered doing it by a recipe basis, but with there being little to no duplicate recipes, there was no way to create a functional model. From there the decision was to make a model that recommends based on regional cuisine, as there were only 19 unique cuisines.

```
len(df['Recipe_name'].unique())
```

```
2636
```

```
df['Cuisine_type'].unique()
```

```
array(['american', 'mexican', 'chinese', 'mediterranean', 'italian',
       'french', 'south east asian', 'nordic', 'eastern europe',
       'british', 'caribbean', 'south american', 'middle eastern',
       'asian', 'japanese', 'central europe', 'world', 'indian', 'kosher'],
      dtype=object)
```

Figure 4

The model was manually created by creating a naives bayes function with 4 helper functions to define the individual variables. These functions were run 9 times for each cuisine with a format of ‘cuisine|nutrition category’ where nutrition is either fat, carb, or protein and category is high, medium, low, and all of this data was stored into a dictionary and then converted to a dataframe. This newly created dataframe was the model which had percentage values for the probability of that combination to be found in the model based on the naives bayes function.

To test the model, the probabilities for each carb, fat, protein were multipled together for each cuisine and my metric of success was if it was within the top 5 of recommended cuisines. Because there was simply too much data to manually test, a small sample of 15 instances were used and compared with an accuracy of 9/15. One major issue noted while performing these tests was the large skew of american and mediterranean items in the data set; in every test, both were always top 2 as they just had the largest amount of data points.

Takeaways

The model created gives a recommendation for regional cuisines for people looking to try out new food and ideally healthy food given their requirements of nutritional values in fat, carbs, and proteins. While the model only gives a cuisine type, this can then be expanded upon as the user can do research in the future on that regional food. Additionally with future adjustments made to the system, the ideal would be that the user can then look through the options of food within that cuisine type.

The issue with the model comes with the fact there is such a large imbalance in the data. Regardless of the combination of categories for nutrition, American and Mediterranean cuisine will always be mentioned as the best 2 options. While there are a variety of options in these cuisines, there is still a fundamental error with the model that would need to be improved upon before making this usable for public use

Future Research

The food recommender model is fully capable of recommending cuisines that will help people find both healthier food and more variety in their food. Two major improvements that can be done to improve the product is to find more recipes to reduce the skew, and to create a function that will show food options of the recommended cuisine. After the model is used to find the cuisine there can easily be a UI created that will allow the user to find options within the model itself without having to use a secondary source for more research. Additionally by adding extra data points and evening out the amount of options within the cuisines, the model will bring more variety and accuracy to the recommendations.