Foods that combat COVID-19

Determining the relationship between foods and Covid-19

Arash Behpour
Computer Science
Virginia Tech and Polytechnic
Institute and State University
Blacksburg, VA
arashb97@vt.edu

ABSTRACT

This research is motivated on the ability of a person to combat against Covid-19 with their diet. This research endeavor attempts to find the relationship between foods that people in different countries have eaten and their status with Covid-19. I conduct an analysis to see which countries have had higher confirmed recoveries / confirmed Covid-19 cases while looking at those countries most common food intakes that might lead to their high recovery rate. I weight the food groups that are more often eaten within the top recovered countries as a better combative against Covid-19. With the food group-Covid-19 relationship information I use semi-supervised label propagation technique to classify common food products for the food groups. Once a relationship is established, I then model common food products with their food groups in a social network system. I show the best food groups to eat within the social network system for visualization. Finally, I use 10-fold cross validation to evaluate by discrepancy on the label propagation model's ability to classify food products within the best food groups to eat. This research will determine what food groups and some foods that can best combat against Covid-19.

CCS CONCEPTS

• Semi-Supervised learning • Social Network Analysis • Cross Validation • Label propagation • Pearson correlation

KEYWORDS

Covid-19, Diet, Health

Problem Statement

The major problem being tackled is what foods contribute to the most Covid-19 recoveries which would hopefully lead to lower Covid-19 death related cases. The sub problem being investigated is then what specific types of foods can be recommended to eat to have a healthy diet against Covid-19. The theme then is, can a person's diet and health influence their combative immune system against Covid-19 through a data analytics perspective?

Problem Domain: The domain resides in analyzing people's health and diet against Covid-19 information (confirmed cases and number of people recovered). The practical impact is to encourage people to start incorporating foods in their diet that might help their body against Covid-19. In society we are taking external precautions against Covid-19 by wearing masks and keeping distance from one another. However, the question then becomes are there ways we can take internal precautions for our body to help against Covid-19? Some ways we can strengthen our body is by exercise and by the foods we eat. This research paper will delve into the internal measures one might take to fight Covid-19 using the foods we eat. The data that will be used for this research endeavor is a Covid-19 diet dataset by country [1] and nutritional values for common foods and products [2].

Problem Solution: The approach to solving this research is split into 2 stages. First stage is to first analyze the Covid-19 diet dataset by country [1] to get a view on what the data represents. I as the researcher will interpret the data when looking at Covid-19 confirmed and recovered cases. I will look for food groups (meats, fruits, etc.) that contain a high Covid-19 recovered over confirmed cases ratio. These food groups are compared to one another by their quantity of food supply intake percentage. Based on my results the second stage will then attempt to find real food products that a person can eat and recommend it to them.

I am also keeping in mind that correlation does not affect causation. Meaning that just because there is a certain food group that has a high number of recovered cases does not mean by eating this food it will cause the person to recover. This study is based on data analytics and the interpretation of the results might be by eating a certain food group it can help combat against Covid-19.

This approach is ideal because I will determine the data relationship between Covid-19 and foods to answer the question on what groups of foods are more beneficial to eat to combat Covid-19. The second stage finds a couple common food products in dataset2[2] that correspond to the food groups in dataset1[1]. I then represent this relationship as a social network and perform semi-supervised technique label propagation to classify the rest of the food products in dataset2[2] and recommend the food products corresponding to the best food groups determined in the first stage. This then answers the question on what foods to eat and

what are their corresponding nutrients (Vitamin D, Calcium, etc.). The reason I do not build a recommendation system by collaborative filtering because I do not want to recommend based of nutrient information but rather food group information (connected to Covid-19) extracted from the first dataset. The evaluation plan measures the discrepancies of food product-food group classification to show how accurate the best food group products are being recommended.

1 Data Description 1

The first dataset is a Covid-19 diet dataset by country [1]. This dataset contains 2 different csv files. The rows of the first file contain the countries and the columns contain the food groups with their corresponding food intake percentage and Covid-19 data (shown in figure 1). Covid-19 data is shown by percentage of each country's population. The second file contains a column of the food groups and another column for a sub sample of food products that correspond to its food group (shown in figure 2). Having data from countries all over the world helps remove location dependency in my analysis. However, this data analysis will be on a group(country's) data and not individual(people) data. This dataset helps me solve the first stage in determining the relationship between Covid-19 and food groups.

	Alcoholic Beverages	Animal fats	Animal Products	Aquatic Products, Other	Cereals - Excluding Beer	Eggs	Fish, Seafood	Fruits - Excluding Wine	Meat	Milk - Excluding Butter	...
0	0.0014	0.1973	9.4341	0.0	24.8097	0.2099	0.0350	5.3495	1.2020	7.5828	...
1	1.6719	0.1357	18.7684	0.0	5.7817	0.5815	0.2126	6.7861	1.8845	15.7213	...
2	0.2711	0.0282	9.6334	0.0	13.6816	0.5277	0.2416	6.3801	1.1305	7.6189	...
3	5.8087	0.0560	4.9278	0.0	9.1085	0.0587	1.7707	6.0005	2.0571	0.8311	...
4	3.5764	0.0087	16.6613	0.0	5.9960	0.2274	4.1489	10.7451	5.6888	6.3663	...

Figure 1.

Cereals - Excluding Beer	Barley and products; Cereals, Other; Maize and products; Millet and products; Oats; Rice (Milled

Figure 2.

The second dataset is nutritional values for common foods and products [2]. This dataset contains one csv file where each row contains a unique food product and the columns correspond to the nutrition information (calories, total fat, protein, Vitamin D, etc.) per 100 grams of that specific product (shown in figure 3). This data allows me to find real food products that are within the best food groups selected from the first dataset. For example, this dataset will help find food products that are within the meats or vegetables food groups.

	name	serving_size	calories	total_fat	saturated_fat	cholesterol	sodium	choline	folate	folic_acid	...	fat	saturated_fatty_acids	monounsaturated
0	Cornstarch	100 g	381	0.1g	NaN	0	9.00 mg	0.4 mg	0.00 mcg	0.00 mcg	...	0.05 g	0.009 g	
1	Nuts, pecans	100 g	691	72g	6.2g	0	0.00 mg	40.5 mg	22.00 mcg	0.00 mcg	...	71.97 g	6.180 g	
2	Eggplant, raw	100 g	25	0.2g	NaN	0	2.00 mg	6.9 mg	22.00 mcg	0.00 mcg	...	0.18 g	0.034 g	
3	Teff, uncooked	100 g	367	2.4g	0.4g	0	12.00 mg	13.1 mg	0	0	...	2.38 g	0.449 g	
4	Sherbet, orange	100 g	144	2g	1.2g	1mg	46.00 mg	7.7 mg	4.00 mcg	0.00 mcg	...	2.00 g	1.160 g	

Figure 3.

2 Data Preprocessing 2

For the first dataset I remove unnecessary data columns that will not help in solving my problem. I remove the data columns 'Obesity', 'Undernourished', 'Active', and 'Unit (all except Population)'. I remove the 'Obesity' and 'Undernourished' columns because I am not looking at the relationship between obesity nor undernourished population with Covid-19. I remove the 'Active' column because I cannot make a confident conclusion based on the number of active Covid-19 cases and their corresponding relationship with food intake. I removed the 'Unit (except for Population)' column because I know that all the columns in the first dataset are percentages except for the population and names of the countries. For the second dataset I remove the units from the values within the nutrient columns. I do this because I when I measure similarity, I will compare the numerical values corresponding to each nutrient column, so I do not care that nutrients have different units within the dataset.

I also combine the information given in the food intake percentage per country csv file with the food group description csv file. This is necessary to connect the food groups data in the first dataset with the common food products data in the second dataset. I do this by text comparison which is by looking to see if the substring of a sub sample food product given in the first dataset is in the string name of the common food product in the second dataset. For example, 'oats' is a sample food product in the 'Cereals – Excluding Beer' food group from the first dataset which would be assigned to food product 'Quaker Oats' from the second dataset. This pre-processing is needed to label some data points before using the label propagation semi-supervised technique in the model building portion of this research project.

3 Data Exploration 3

For the first dataset I find the relationship with food groups and Covid-19 for each country. For example, Afghanistan's highest food intake percentage is from 'Vegetal Products' and 'Cereals – Excluding Beer' (shown in figure 4). Adding all the food intake percentages for each food group will add up to 100%. Each country data point has the percentage of people who recovered from Covid-19, the percentage of people confirmed to have Covid-19, and the population of the country. The first dataset then shows what most common food groups people ate in a specific country and its relationship to the recovered Covid-19 cases.

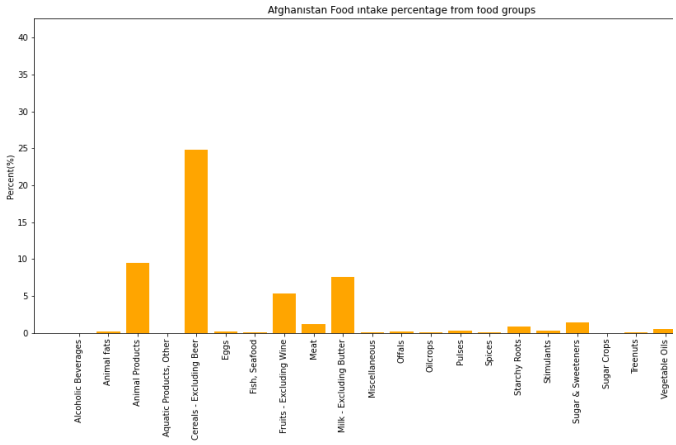

Figure 4.

To find the food group and Covid-19 relationship, I sort the food intake percentage values of food groups for the top Covid-19 recovered countries. To find the top recovered countries I use the information on Covid-19 recovered population and confirmed population (shown in figure 5). If I only looked at the recovered population and not the percentage recovered from confirmed cases, then there would be a preference for bigger countries because they have more people to be to recover.

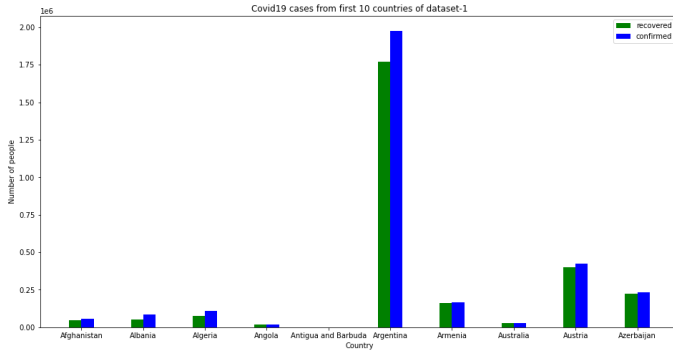

Figure 5.

For the second dataset each food product is a node within a social network. Its edges are weighted by its corresponding similarity to another food product. This similarity is based off Pearson correlation which is determined by the food products nutritional value (proteins, Vitamin D, etc.). Each food product node is connected to its most similar food products.

4 Model building 4

First, I establish that the top Covid-19 recovered countries are the top 10% of countries that have the best (recovered percentage / confirmed percentage) ratio. The (recovered percentage / confirmed percentage) ratio can be seen from the recovered and confirmed population in figure 5. The best (recovered percentage / confirmed percentage) ratio is 1. This means that 100% of the confirmed Covid-19 cases of people were all able to recover. If a country contains a ratio of 1 it makes that country more valuable to look at and determine what are the highest food group intakes that people in that country normally eat. The top 10% of countries with the best ratios are shown below in table 1.

Top 10% of countries:	'Vanuatu', 'Samoa', 'Tajikistan', 'Iceland', 'Grenada', 'Nepal', 'Djibouti', 'Uzbekistan', 'Central African Republic', 'Guinea', 'Saudi Arabia', 'Azerbaijan', 'India', 'Georgia', 'Croatia', 'New Zealand', 'Kyrgyzstan'

Table 1.

After determining the top 10% of countries, I then look at the top 8 food groups that have the highest food intake percentage within that country. I count and total up the occurrences to find the most common food groups that were eaten within the top 10% of recovered countries. The food groups with the most occurrences are then determined to be the best food groups to eat and combat against Covid-19. I take the top 2 highest food group occurrences to be the best food groups. The best food groups that were determined are 'Vegetal Products' and 'Animal Products' (shown in table 2).

Count Occurrences:	{22: 17, 2: 16, 4: 15, 21: 10, 9: 10, 7: 6, 15: 5, 12: 2, 8: 2, 0: 1, 17: 1}
Best food group indices:	22, 2
Best food group names:	'Vegetal Products', 'Animal Products'

Table 2.

After discovering the best food groups, I looked to learn more about their relationships with the recovered/confirmed ratio for all the countries. For 'Vegetal Products' it seems that having a higher food intake percentage can result in a higher recovered/confirmed ratio (shown in figure 6). Meaning the more a country ate Vegetal products the more likely they would have a higher recovered/confirmed ratio. For the second-best food group more countries had higher recovered/confirmed ratio when the food intake percentage of 'Animal Products' were between 5-15% (shown in figure 7). When 'Animal Products' food intake went above 15% there appeared to have some countries that had low recovered/confirmed ratios. This can imply that containing 'Animal Products' in a person's diet is important but they should not look to making their whole diet based on only eating animal products like meat and butter.

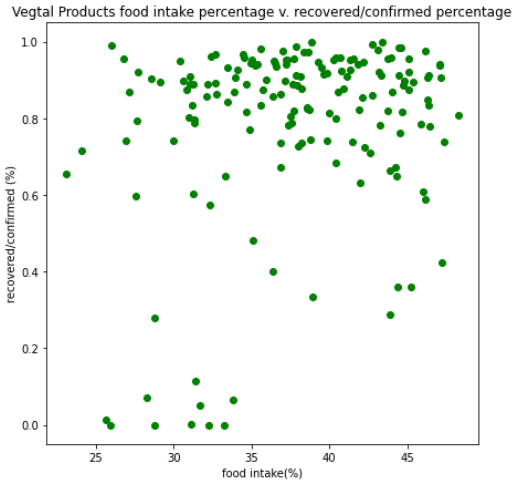

Figure 6.

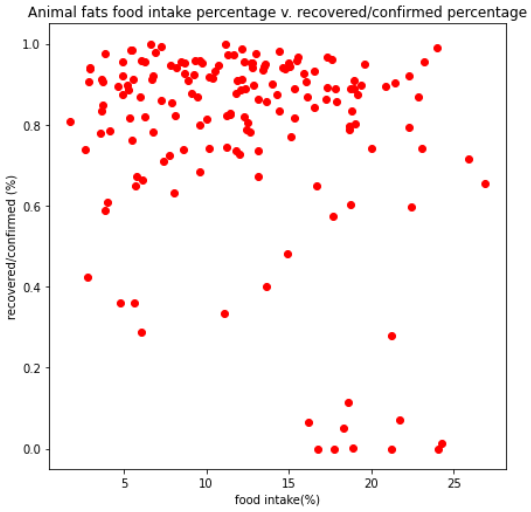

Figure 7.

After determining the best food groups, I extract a sub sample of the common food products from the second dataset [2] to use. The reason I extract a sub sample is because when I later train my label propagation model it takes a very long time when using the whole common food products dataset [2]. Like mentioned in the data pre-processing section I perform a text comparison to map food groups in the first dataset to common food products in the second dataset. For example, 'Apples would be mapped to the 'Fruits' food group. I do not map all food groups in my model because I mainly care about the best food groups and the other food groups can be all placed under one same label (Other food groups). This mapping is based off the string name of the common food product. Ideally, I would want to have a nutritionist assign each common food product to their corresponding food group. However, I do not have access to a professional nutritionist, and it would take too much time to look through the food products and assign food groups to it. The mapping process does not map all the food products to a food group, so there will be unlabeled food products.

The next process is to use Pearson Correlation between food products to find what each food product is strongly similar too. Pearson Correlation is a technique used to measure similarity between items, in this case food products. Each food product is an x-variable and y-variable where the nutrient information are the values of these variables (shown in figure 8) [3].

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient
x_i = values of the x-variable in a sample
$\bar{x}$ = mean of the values of the x-variable
y_i = values of the y-variable in a sample
$\bar{y}$ = mean of the values of the y-variable

Figure 8.

I calculate the Pearson Correlation for each food product extract the 5 most similar food products for each food product that must be strongly correlated (0.5 <= Pearson Correlation value <= 1). The 5 most similar food products will indicate the edges that are built in the social network representation of the food products. The weights of these edges will be the Pearson Correlation values.

Using the food products as nodes and the Pearson similarity between other food products as edges, I construct a social network graph. The graph contains 1000 sub sampled common food products and 3608 edges. While building the graph each node is labeled to its assigned food group and if it does not it is considered an unlabeled food product which needs to be assigned. I add the edges between food products that are strongly correlated and assign the weight of the edge to be the Pearson Correlation value calculated between those food products. After constructing the graph there will be labels for the best food group ('Vegetal Products' and 'Animal Products') mapped to common food products, labels of the other food groups mapped to common food products, and unlabeled food product data that has no food group assignment yet (shown in figure 9).

Figure 11. and Figure 15. Graph Key	
	Vegetal Products
	Animal Products
	Other food groups
	Unlabeled food product

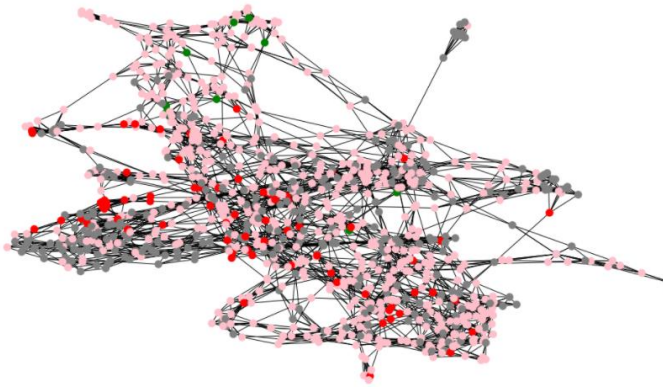

Figure 9.

Since I have represented my data as a social network graph, I then use semi-supervised technique called Label Propagation. Label Propagation is an iterative algorithm where it assigns labels to unlabeled data points by propagating labels through the dataset. The required assumption for Label Propagation is that an edge connecting two nodes carry a notion of similarity which I established using the Pearson Correlation and top 5 similar food products. The concept behind the Label Propagation algorithm is that each unlabeled data point will take random walk paths in the graph until they reach an absorbing state which is a labeled data point. There can be multiple random walk paths for a single data point. A label for an unlabeled data point is determined by the label that resulted in a majority of the random walk paths. The mathematical formulation is shown below in figure 10 [4]. The T matrix is the degree matrix of the graph that is being propagated on (in figure 10). The degree matrix is made using the weight of the edge and the degree of the node being examined.

$$y_i[c] = \sum_{j \in X_l} T_{ij}{}^{t}\, y_j[c]$$

The probability of node $x_i \in X_u$ to have label c

The probability P(i→j) to jump from node x_i and end up in node x_j in t steps. We can define the number of steps to be a large number (infinity).

$$\hat{Y} = T^{t\to\infty} Y$$

Vector of labels we need to get | Matrix of probabilities to end up on different nodes | Vector of labels on those nodes

$$\begin{bmatrix} \hat{Y}_l \\ \hat{Y}_u \end{bmatrix} = T^{t\to\infty} \begin{bmatrix} Y_l \\ 0 \end{bmatrix}$$

$\hat{Y}_u$ is what we are interested in.

Figure 10.

The reason Label Propagation is an iterative algorithm is because the T matrix that contains the probabilities to go from a labeled/unlabeled node to a labeled/unlabeled node will be iteratively multiplied to itself until convergence as shown in Figure 11 [4]. After the T matrix has converged, I can determine the labels for the unlabeled data by looking at the probabilities from unlabeled nodes to labeled nodes and pick the largest probability. The largest probability will correspond to the labeled node that ended in the most random walk paths for the corresponding unlabeled node. Finally, the unlabeled node is assigned the label of the largest probability labeled node (shown in figure 12).

- T_{ll} — Probability to get from labelled nodes to labelled nodes
- T_{lu} — Probability to get from labelled nodes to unlabelled nodes
- T_{ul} — Probability to get from unlabelled nodes to labelled nodes
- T_{uu} — Probability to get from unlabelled nodes to unlabelled nodes

$$T = \begin{bmatrix} T_{ll} & T_{lu} \\ T_{ul} & T_{uu} \end{bmatrix} = \begin{bmatrix} I & 0 \\ T_{ul} & T_{uu} \end{bmatrix}$$

$$\lim_{t\to\infty} T^t = \begin{bmatrix} I & 0 \\ T_{ul} & T_{uu} \end{bmatrix} \times \begin{bmatrix} I & 0 \\ T_{ul} & T_{uu} \end{bmatrix} \times \begin{bmatrix} I & 0 \\ T_{ul} & T_{uu} \end{bmatrix} \times \dots$$

$$= \begin{bmatrix} I.I + 0 + 0 + \cdots & 0 + 0 + 0 + \cdots \\ T_{ul} + T_{ul}.T_{uu} + T_{ul}.T_{uu}{}^2 + \cdots & T_{uu}.T_{uu}.T_{uu}.\dots \end{bmatrix}$$

$$= \begin{bmatrix} I.I + 0 + 0 + \cdots & 0 + 0 + 0 + \cdots \\ (I + T_{uu} + T_{uu}{}^2 + \cdots).T_{ul} & T_{uu}.T_{uu}.T_{uu}.\dots \end{bmatrix}$$

$$= \begin{bmatrix} I & 0 \\ \left(\sum_{t=0}^{\infty} T_{uu}{}^t\right).T_{ul} & T_{uu}{}^{\infty} \end{bmatrix}$$

Model this as the sum of a geometric series x^n where $|x|<1$. We get the sum as $(1-x)^{-1}$.

Since we self-multiply T_{uu} and values in T_{uu} are less than 1, it will reach 0.

$$= \begin{bmatrix} I & 0 \\ (I - T_{uu})^{-1}.T_{ul} & 0 \end{bmatrix}$$

Figure 11.

$$\begin{bmatrix} \hat{Y}_l \\ \hat{Y}_u \end{bmatrix} = \begin{bmatrix} I & 0 \\ (I - T_{uu})^{-1}.T_{ul} & 0 \end{bmatrix} \begin{bmatrix} Y_l \\ 0 \end{bmatrix}$$

$$\hat{Y}_u = (I - T_{uu})^{-1}.T_{ul} \times Y_l$$

$$Label\ of\ x_i \in X_u = \mathbf{argmax}\ \hat{Y}_u[i]$$

Figure 12.

After performing Label Propagation on the food products graph, I classified all the unlabeled food products (shown in figure 13). By performing the food group semi-supervised classification my model can recommend common food products within the best food groups to eat against Covid-19.

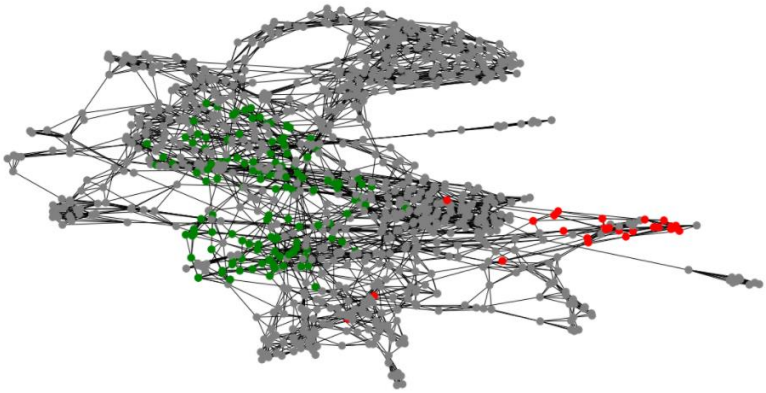

Figure 13.

5 Model Evaluation 5

I use 10-fold cross validation on the label propagation model to check the best food groups recommended food products. I did not do cross validation on the first stage because there was not a lot of data points (170 for each country) and I wanted a global analysis

of the food group and Covid19 data so that my model is not bias toward a certain region in the world. I checked the accuracy of the model by seeing the difference from the whole dataset model (after label propagation classification) and between subsets generated from cross validation. My model achieved an accuracy of 85%. The evaluation technique looked for discrepancies with the foods that were being recommended to eat to combat Covid19. Food products in 'Vegetal Products' and 'Animal Products' food groups were recommended to eat while other food groups were not. However, one aspect to note is that some food groups can be intertwined within food products. Like potato chips uses ingredients Potatoes ('Starchy Roots') and Canola oil ('Vegetal Product') shown in table 3. But it also depends on the type of chips food product as well.

Example Vegetal Products:	'Peas, with salt, drained, boiled, cooked, sprouted, mature seeds', 'KASHI, Frozen Entree, Chicken Fettuccine, STEAM MEAL', 'Cookies, baked, refrigerated dough, peanut butter', "CRACKER BARREL, from kid's menu, macaroni n' cheese plate", 'Beverage, not chocolate, dry, milkshake mix', 'SPAGHETTIOS, easy open, SpaghettiOs Original', 'Pigeon peas (red gram), with salt, boiled, cooked, mature seeds', 'Snacks, baked, restructured, white, potato chips', 'MORNINGSTAR FARMS Roasted Garlic & Quinoa Burger, unprepared, frozen', 'Beans, low sodium, canned, mature seeds, great northern'

Table 3.

After evaluating the model, I use a tree classifier and convert the assigned food product labels (food groups) to a simpler classifier output. I do this to be able to use a SHAP plot to explain why certain food products are chosen as the best food products. SHAP is a python library that can help explain tree models. The way SHAP works is that it interprets the nutrients as features for the food products and weighs its importance when my model decides if a food product is best to eat against Covid-19 or not. I made the output equal to 1 on food products within the best food groups and 0 for other food products within the other food groups.

Then I split the food products labeled data with 80% train and 20% test. The tree classifier model fits on the training data and returns its predicted outputs given the test data. What was found was that nutrients like potassium, lutein zeaxanthin, and saturated fatty acid were the nutrients that were common among the best food products (shown in figure 14).

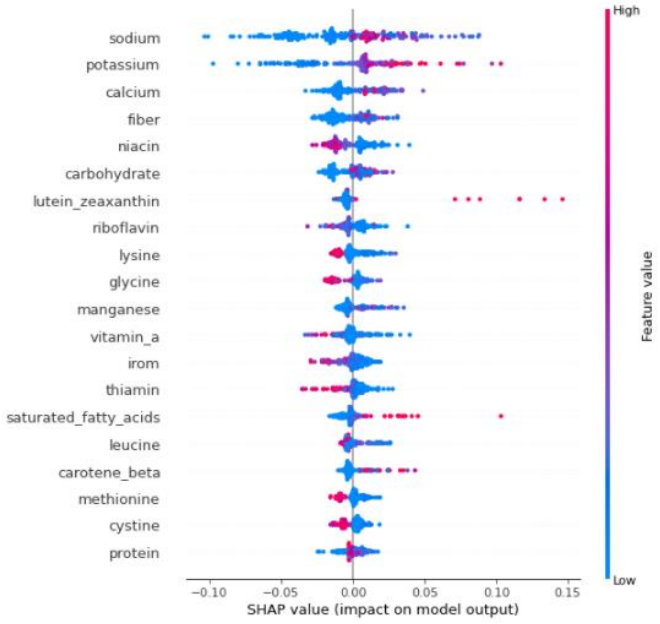

Figure 14.

6 Insights 6

Vegetal products and animal products are the most common food groups that were consumed within the top 10% of recovered countries. Food products people should look to eat should have high levels of potassium, lutein zeaxanthin, and saturated fatty acid nutrients. Potassium helps a person's muscles work, including the muscles that control a person's heartbeat and breathing. Lutein zeaxanthin is a potent antioxidant found in vegetables. Saturated fatty acid comes from animal sources and has antimicrobial properties, makes up a majority of human cell membranes and improves the immune system.

7 Lessons Learned 7

Vegetal products and animal product foods might be better for people to eat and recover against Covid-19 compared to other food groups (fruits, grains, sugars/sweeteners, etc.).

This research experiment would have been more ideal if the data points from the first dataset [1] were individual versus group(country) data points. I believe this research goal is better representative if the data points were from the food intake percentages of individual people. Because there is a risk of the Ecological fallacy were just because there is an inference in the group level (countries) does not mean it applies to the individual level (people). In addition, there could be other factors on why a country had a better recovery rate (Ex. resources in U.S. versus Peru)

The initial labeling to combine both datasets before creating the label propagation model would be better if it were not based on

the food products text name. It would be better to have a specialist like a nutritionist to classify food products. Or I would need to do research on each individual food product within the dataset which takes a lot more time. However, some common food products can overlap in the food groups they belong to. Meaning that a food product can have multiple food groups representing it. For example, a hamburger could contain buns ('Cereal – Excluding Beer'), lettuce (Vegetables), burger ('Meats') and sauce ('Vegetal Products'). This research endeavor was a global overview into the types of foods people should eat to combat against Covid-19.

ACKNOWLEDGMENTS

I want to thank my professor Chandan Reddy for teaching me new skills in data analytics and guiding me through this research endeavor. I want to thank both my teacher assistant's Ming Zhu and Aman Ahuja for helping me with my questions and concerns.

REFERENCES

[1] M. Ren, "COVID-19 healthy Diet Dataset," 07-Feb-2021. [Online]. Available: https://www.kaggle.com/mariaren/covid19-healthy-diet-dataset?select=Food_Supply_Quantity_kg_Data.csv. [Accessed: 21-Mar-2021].

[2] A. Antonov, "Nutritional values for common foods and products," *Kaggle*, 18-Jun-2019. [Online]. Available: https://www.kaggle.com/trolukovich/nutritional-values-for-common-foods-and-products. [Accessed: 21-Mar-2021].

[3] "Pearson correlation coefficient," 24-Apr-2021. [Online]. Available: https://en.wikipedia.org/wiki/Pearson_correlation_coefficient#:~:text=Pearson's%20correlation%20coefficient%20is%20the,product%20of%20their%20standard%20deviations. [Accessed: 03-May-2021].

[4] C. Reddy, "Semi-Supervised Learning," in *Data Analytics,* 03-May-2021.