COVID-19 and Nutrition in Food

Linlin Li, Mingxuan Yang, Yijie He, Zilin Yin

November 16th
https://github.com/biostat823-FinalProject/Game-of-Data

1 Background

COVID-19, a disease caused by a new type of coronavirus, has become a major global human threat that has turned into a pandemic. During this period, the role of the immune system has attracted people's attention. Some researchers identified some important nutritional considerations for the prevention and management of COVID-19 diseases, including the role of select micronutrients (vitamins and minerals), phytochemicals and probiotics in conferring protection against both viral infection and pathogenicity. [?]. Based on these studies, some "experts" and articles have urged people to buy supplements or eat particular foods to enhance their immune system. However, as highlighted by the World Health Organization, a healthy lifestyle makes all bodily functions work better, including immunity [?]. Having a healthy diet, including lots of fruits and vegetables, is a key component of a healthy lifestyle and plays a vital role in supporting a well-functioning and effective immune system to help protect against infection and other diseases [?], which indicates that people may not need to eat such supplements. Therefore, facing these two completely different claims, people, including us, have a big question: Can these supplements and food really strengthen people's immune systems? Is it necessary for people to eat these supplements or particular foods?

In this project, we are going to provide some exploratory analysis about COVID-19. Then, we will try to build machine learning models to examine the relationship between the recovery rate of COVID-19 and food. Finally, we will evaluate the performance of the models and provide insights from the results.

2 Objective

The objective of this project is to build a dashboard to visualize current COVID-19 situation in the world and monitor the trend of the pandemic, and to find the relationship between food intake from different food products and recovery rate, and determine if a healthy diet could increase the possibility for one to get recovered from COVID-19.

3 Data Description

3.1 Healthy Diet Dataset

The Healthy diet dataset has 170 rows and 32 columns. Each row in the dataset has all of the information for a certain country. The first columns contains the names of all of the countries. From column 2-26, each column contains the percentage of food intake in kilograms for certain food type in different countries. The last 6 columns contain information like percentage of confirmed cases, death cases, recovery cases, active cases, total population, and one last column for unit declaration.

4 Exploratory Analysis

4.1 Investigating Economic Impacts

It is a common belief that the speed and severity of disease transmission in a country are often related to its economic status. Figure **??** and **??** displays the relationship between COVID-19 and economy. From these graphs, we found that:

- In the first quarter of 2020, more than half of the countries/regions were affected by COVID-19, of which China suffered the most economic impact, while the United States suffered only economic stagnation.
- In the second quarter of 2020, most of the G20 countries suffered a severe economic blow due to COVID-19, and China is the only country that has recovered its economy.

Figure 1: COVID-19 and economic impact in the first quarter of 2020 on G20 countries.

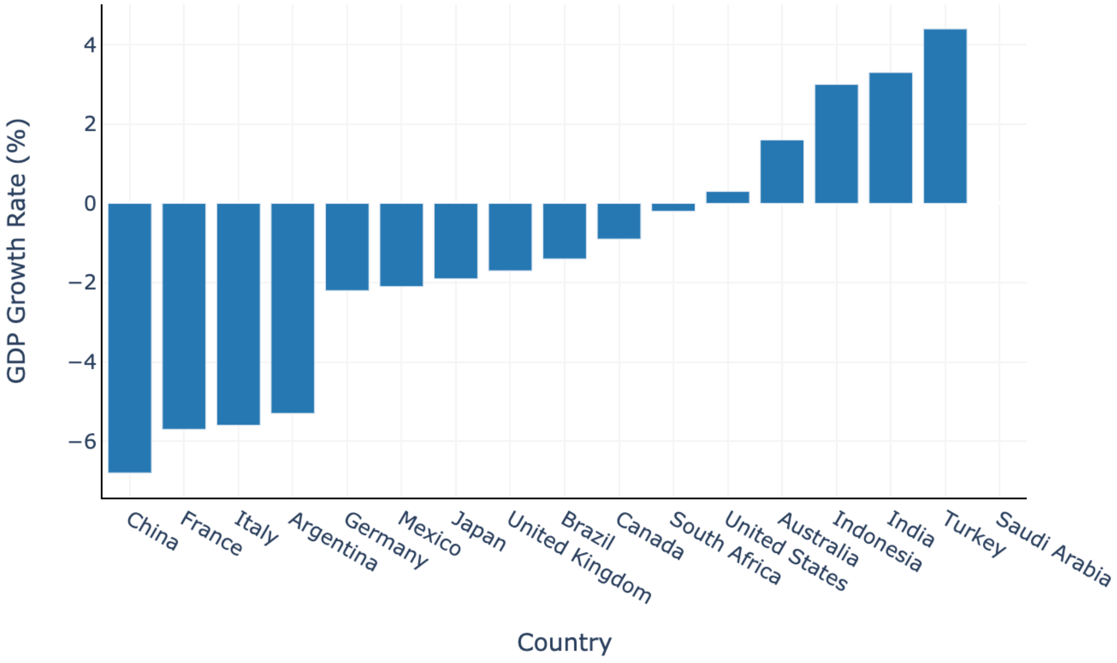
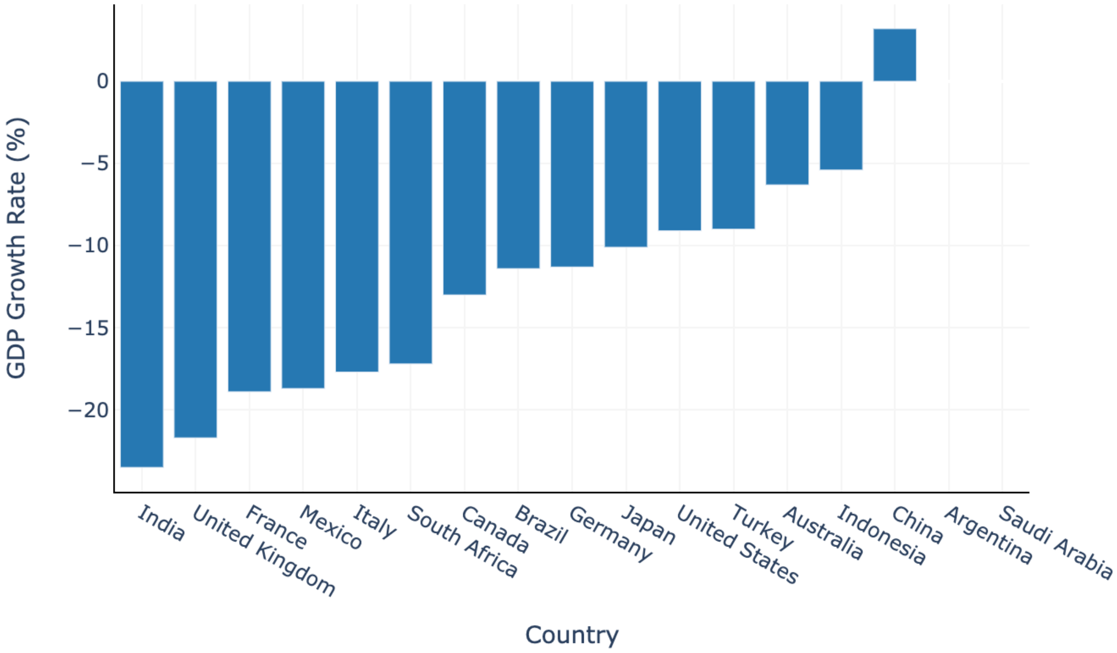


Figure 2: COVID-19 and economic impact in the second quarter of 2020 on G20 countries.



4.2 Investigating Dietary Factors

Next, we will study the relationship between the recovery rate of COVID-19 and food. For example, Figure **??** shows the relationship between animal fats and COVID-19 recovery rate in different regions.

- People in Africa have the lowest average animal fats intake around the world and a really low recovery rate of COVID-19.
- People in Asia and Pacific as well as Arab States have a moderately low animal fats intake and a moderately low recovery rate.
- People in Europe have a more diverse eating habits and the recovery rate also varies a lot.
- People in Middle East and South/Latin America have a moderately low animal fats intake, but the recovery rate has a pretty wide range.
- People in North America has the highest animal fats intake with a close to 50% recovery rate.

Figure 3: The relationship between animal fats and COVID-19 recovery rate.

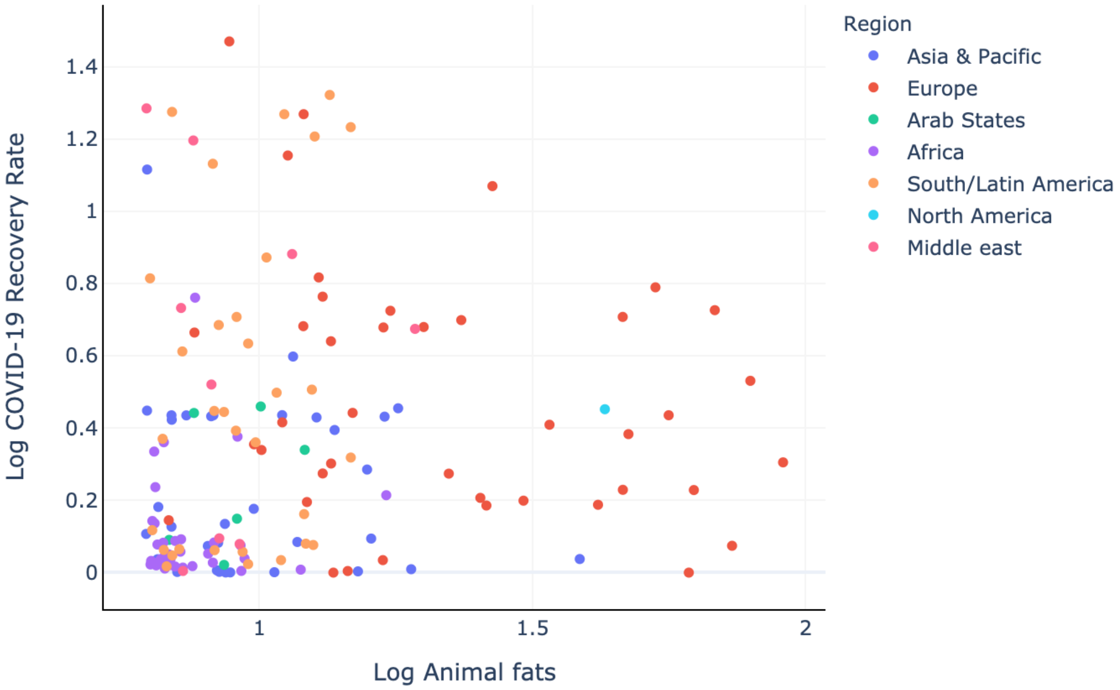


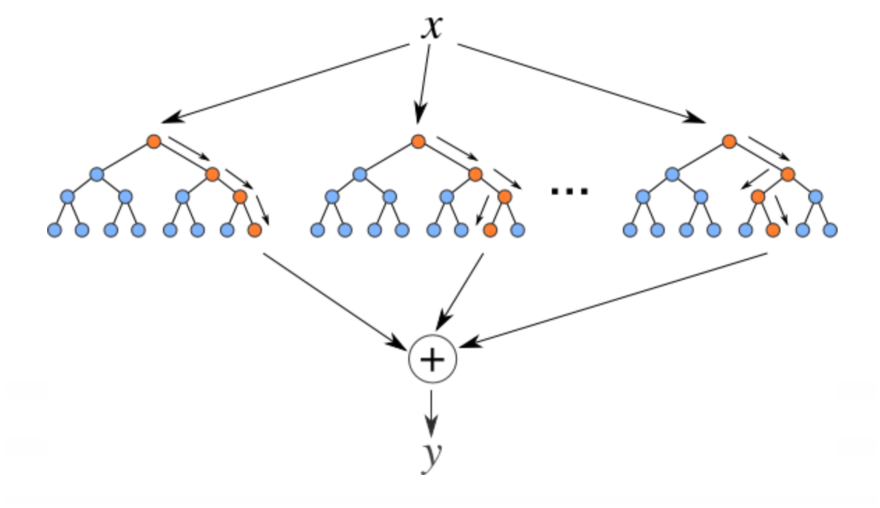
Figure **??** indicates that the intake of animal fats is likely to be related to the recovery of COVID-19. We will build machine learning models to help us understand it better in the following sections.

5 Models

5.1 Random Forest Regression

The first model we used to investigate the relationship between food intake and recovery of COVID-19 is random forest regression model.

Figure 4: Random forest regression model architecture.

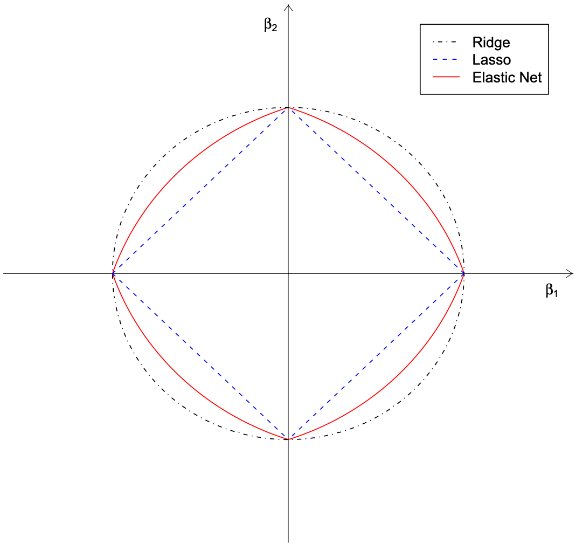


Random forest regression is a ensemble learning method for regression. As shown in Figure 4, x is the input data, and y is the true outcome. A random forest regression model combines the predictions of its sub-trees into a single one in order to get a more accurate result. Note that each sub-tree itself is a machine learning model, normally a decision tree model. Overall, it has advantage in telling the relative importance of each feature.

5.2 Elastic Net Regression

The second model we used is elastic net regression model.

Figure 5: Elastic net regression geometry.



As shown in Figure 5, Elastic net regression is a combination of both Lasso and Ridge regression. It has advantage in regularization and can effectively prevent over-fitting, especially when the number of columns is big or when the features are highly correlated to each other. It also make the model easier to understand as it assign little weights or even zero weights to the parameters that hardly influence the result. Hence, we wanted to see the effect of using elastic net regression model on the task.

Elastic net regression model's regularization effect can be explained with its loss function.

$$\hat{\beta} = \arg\min_{\beta} \|y - X\beta\|^2 + \lambda_2\|\beta\|^2 + \lambda_1\|\beta\|_1 \tag{1}$$

As shown in equation (1), it has two additional terms compared to traditional mean square error function. This combines feature elimination from Lasso and feature coefficient reduction from the Ridge model to improve model's predictions. One thing to be noticed is that l1 ratio is used to determine the weights the elastic net model gives to its Lasso part and Ridge part respectively. With different l1 ratio, the model performance could vary significantly. Later in the hyperparameter selection section, we will discuss how we come up with a good l1 ratio.

5.3 Gradient Boosting

Gradient boosting is an ensemble learner that combines weak learners such as decision trees into one prediction. Boosting, unlike bagging in random forest, grows trees sequentially on a modified version of the data, and generalizes them by allowing optimization of an arbitrary differentiable loss function. The sudo code for gradient boosting algorithm can written as:

1. Set $\hat{f}(x) = 0$ and $r_i = y_i$, for all i in the training set
2. For b = 1, 2, 3 ... B, repeat:
   (a) Fit a tree, $\hat{f}^b$ with d splits (d+1 split nodes) to the training data
   (b) Update $\hat{f}(x)$ by adding a shrunken version of the new tree
   $$\hat{f}(x) = \hat{f}(x) + \lambda\hat{f}^b(x) \tag{2}$$
   (c) Update the residuals
   $$r_i = r_i + \lambda\hat{f}^b(x_i) \tag{3}$$
3. Output boosted model
$$\hat{f}(x) = \sum_{1}^{B} \lambda\hat{f}^b(x) \tag{4}$$

From the sudo code, it is important to note that the performance of the model depends on the choice of shrink size λ and the number of trees we are growing. There is a tradeoff between these two values. Moreover, the attributes of each tree grown in the algorithm affect the quality of the aggregated model. Therefore, tuning for the parameters aforementioned was essential.

6 Hyperparameter Selection

We used GridSearchCV with 5-fold cross validation for parameter tuning for all of the three models. Specifically, when building the random forest regression model, we looked into number of estimators, maximum number of features, maximum depth, minimum samples split and minimum samples leaf because they are the key factors that influence the model performance. The random forest regression model was optimized to have 100 estimators, 'auto' for the maximum number of features, None for maximum depth, 2 for minimum samples split and 1 for minimum samples leaf.

When building the elastic net regression model, we looked into maximum number of iteration, alpha and l1 ratio, which are the key factors that determine the model performance. The elastic net regression model was optimized to have 1000 for maximum number of iteration, 1 for alpha and 0.5 for l1 ratio.

The optimal gradient boosting model was chosen to have 100 trees and shrink size of 0.1. In addition each tree has 2 minimum split and 1 minimum samples in the leaf.

7 Result Interpretation

We compared the performance of the three models based on squared error loss and r2 score for test data, as shown in the table below.

	r2	MSE
Random Forest	-4.031518	0.04868
Elastic Net	-6.227299	0.04995
Gradient Boosting	-3.802199	0.05166

Hence, random forest performs best in terms of squared error loss. However, all three models have negative r2 values, which suggests the models predict badly for recovery rate from COVID given the nutrition data. The bad performance might be due to the limitation of our dataset, which only contains food data while does not take the social-economic status of each country. Recovery from COVID should not be solely predicted with nutrition information. Nonetheless, Our models might still reveal the association between some nutrition factors and COVID-19 recovery rate, given the variance importance plots.

Figure 6: Variance importance plot of random forest

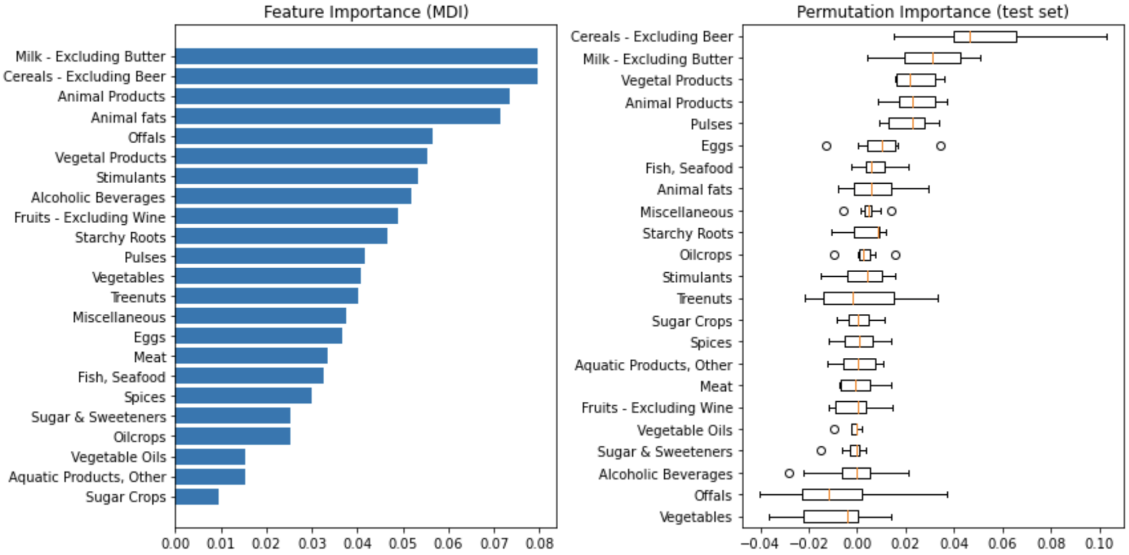
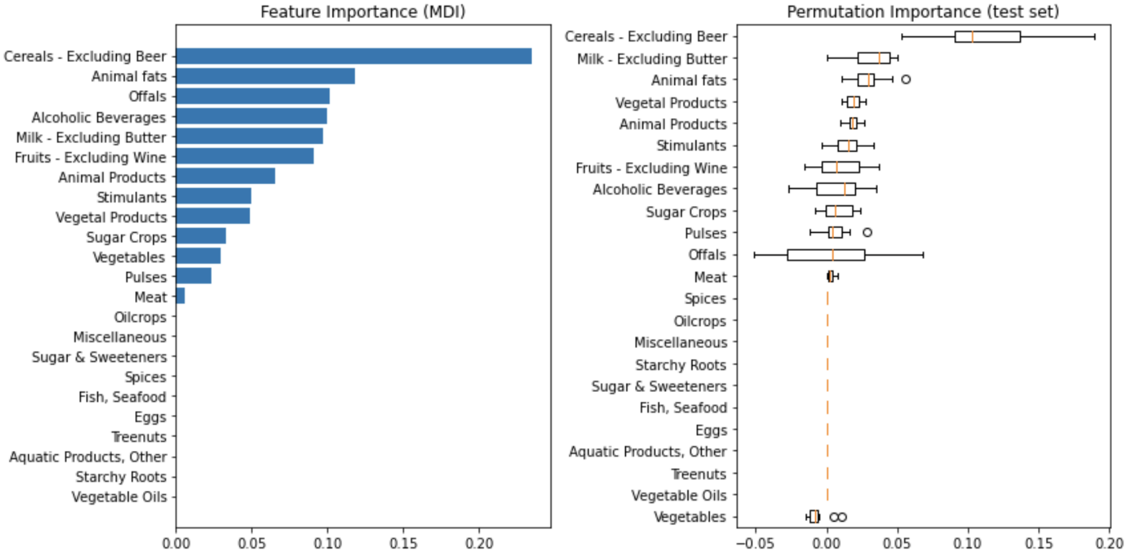


Figure 7: Variance importance plot of gradient boosting



From the figures above, we see that the consumption of cereals, milk, vegetal and animal products appears to be top permutation importance in both models. This finding concord with the WHO instruction. A healthy lifestyle can help build a stronger immune system and promote disease recovery.