12/16/2021

COVID-19 Diet Analysis

MGSC 661 Final Project

Jiahua Liang 260711529

Hisham Salem 261003138

DESAUTELS FACULTY OF MANAGEMENT

MCGILL UNIVERSITY

Table of Contents

1. Introduction

COVID-19 is a major disease that has affected the lives of countless people throughout the years 2020 till now. Unfortunately, a lot of people were tested positive for COVID-19, some of them even lost their lives. For those who survived, don't be frustrated, the worst is behind, everything will be going well again, just keep calm and recover. But is there a non-pharmaceutical treatment that can help recover? A healthy diet! [5][6]

"Health requires healthy food." Said Roger Williams. Eating well is important for people with COVID-19 as their bodies need energy, protein, vitamins, and minerals to recover. Having a good intake of protein and energy-rich foods is helpful for rebuilding muscles, maintaining the immune system, and increasing energy levels. Therefore, in this project, we would like to study the potential effect of protein intake from different kinds of food on the COVID-19 recovery rate. In other words, what kinds of food may help COVID-19 patients to recover?

The data we selected was the COVID-19 Healthy Diet Dataset that is available on Kaggle. It is a suite that contains four datasets about fat intake (%), protein intake (%), energy intake (kcal), and food supply quantity (kg) from distinct categories of food in the diet of different countries in the world, as well as the population and COVID-19 related data of each country. Specifically, we focused on the Protein Supply Quantity Data for this project. This dataset has 170 rows corresponding to 170 countries. The columns contain information about the percentage of protein intake from various categories of food based on the common diet of each country. Some columns contain data about COVID-19 confirmed rate, deaths rate, recovery rate, and active rate of each country, and other health problems such as obesity and Undernourished rate.

There are two main parts to this project. One part is the regression modeling for predicting the COVID-19 recovery rate using protein intake from different food as predictors. The other part is the descriptive data analysis and recommendation of diet. We began our analysis by preprocessing the data and selecting the most key features to the target variable. Then, we explored the selected features and visualized their distributions. Before modeling, we investigated potential model issues like collinearity and outliers. After eliminating these issues, we built different regression models and compared their performance to select the best one. For descriptive analysis, we also performed PCA and data aggregation to help with our analysis.

2. Data Description

2.1 Data Preprocessing

First, we checked if there were any missing values in the dataset. Then, we imputed the missing values instead of dropping the rows because we had a limited number of observations in our dataset. Instead of filling the NAs simply with the zeros or column mean, we used MICE (Multivariate Imputation via Chained Equations) package in R for the imputation. This technique assumes that the missing data are missing at random, and it applies a regression model (for numerical values) on the values in other columns to predict the missing values, which can well resemble the nature of the dataset. Next, we subset the cleaned dataset with all the columns for protein intake, and the column for COVID-19 recovery rate, which was ready for feature selection.

2.2 Feature Selection

Feature selection is the process of removing insignificant features that do not contribute much to the performance of our model. Instead, it is best to identify and remove the insignificant predictors, which will allow our model to train faster, reduce complexity, and make the model easier to understand. To do so, we used two main different packages in R which are Random Forest and Boruta. Please refer to the appendix to see the output for the features selection method. We primarily used Boruta for feature selection and use Random Forest to validate our selection.

2.2.1 Boruta Feature Selection

Boruta is a feature selection wrapper algorithm built around a random forest classification algorithm where it captures all the relevant features for the target variable. First, the algorithm adds randomness by creating shuffled copies of all features, shadow features which are the blue boxes. Then it utilizes the random forest on the extended dataset and applies the feature importance measure to evaluate the importance of each feature. At every iteration, it checks a real feature has higher importance than the best shadow features. Finally, the algorithm stops when all features are confirmed or rejected and distinguishes which features are good, tentative, or bad in green, yellow, and red, respectively (Appendix B Figure 2). We selected the features that are in green and yellow.

2.2.2 Random Forest Feature Selection

Random forest, by default, computes the variable importance by finding the mean decrease in impurity (Gini Importance). At every split, the improvement by the split is what evaluates the importance of the attribute which will be then accumulated separately in all trees for every variable. From there it shows the feature importance on a graph. (Appendix B Figure 3) We compared the results of Boruta and Random Forest and it turned out that these two techniques selected the same set of features even though the order of importance was a bit different. The selected features are shown in the following section.

2.3 Distribution of Selected Features

Before analyzing the relationship between our target variable and predictors, it is important to understand and analyze the predictors themselves. To visualize the nature of each predictor, we created boxplots, density plots, and histograms to best visualize the distribution of the data (Appendix B Figure 6-10). The following tables as well will give a detailed summary of each column (Table 1) of the data as well as the Skewness. Negative scores reflect left skewness while positive scores represent right skewness (Table 2).

Table 1: Summary of Selected Features (percentage intake)

Data	Minimum	1st Quartile	Median	Mean	3rd Quartile	Max
Animal Products	4.46	14.46	21.85	21.23	28.30	35.79
Vegetable Oils	0.00	0.01	0.02	0.02	0.03	0.11
Animal Fats	0.00	0.02	0.06	0.11	0.12	0.98
Vegetal Products	14.22	21.70	28.15	28.77	35.54	45.55
Fish & Seafood	0.06	1.34	2.60	3.36	4.53	18.08
Milk Excluding Butter	0.26	2.42	5.60	5.92	8.50	16.48
Oil Crops	0.01	0.38	0.80	1.37	1.79	8.06
Eggs	0.04	0.43	1.12	1.16	1.68	3.57

Table 2: Skewness

Variable	Skewness	Skewness Type
Recovered	1.7	Highly Skewed
Animal Products	-0.21	Symmetrical

Vegetable Oils	1.92	Highly Skewed
Animal Fats	2.83	Highly Skewed
Vegetal Products	0.21	Symmetrical
Fish & Seafood	1.82	Highly Skewed
Milk Excluding Butter	0.45	Symmetrical
Oil Crops	2.19	Highly Skewed
Eggs	0.61	Moderately Skewed

3. Modelling & Results

3.1 Modelling Issues

After preprocessing our data, we decided to ensure that our predictors are providing accurate results and avoiding bias. Now tree-based models such as gradient boosting and random forest are very robust models, however, we believe that it is best to perform the adjustments to get the best model. To do so, the following tests will be performed.

3.1.1 Correlation and Multicollinearity

To gauge the correlation among variables, we performed a Pearson correlation test amongst all the selected features. Collinearity happens when two numerical values are highly correlated. Therefore, this would affect the significance of a predictor in the model. To avoid that we built a correlation matrix between predictors. From the map, we were able to detect some strong correlation between our variables vegetal products and animal products being negatively correlated by 1 [1][4].

To confirm this, we ran a VIF test, and we realized that there was a strong correlation between the two variables. We, therefore, decided to drop this variable and the model collinearity problem has been solved effectively.

3.1.2 Outlier Detection

Outliers are observations that lie outside the overall pattern of the distribution, these values can significantly affect our model's accuracy and therefore we need to remove these rows. Therefore, we decided to run a Bonferroni outlier test between the target variable and the predictors, and we were able to detect 1 outlier which was subsequently removed.

3.1.3 Log Transformation

From Table 2, the target variable "Recovered" and some of the features are highly right-skewed. In skewed data, the tail region may act as an outlier for the statistical models, which may adversely affect the model's performance, especially for regression-based models. After doing some research online, we concluded that tree-based models are robust to skewed predictors but may still suffer from skewed target variables because tree-based models make predictions by average similar target values. This can lead to highly skewed predictions (predictions could be extremely far off) if the outlier effect is present, which would reduce the model performance. Therefore, we used to apply nature log transformation only on the target variable to reduce the skewness. We also added 1 to every value of the target variable before applying the log transformation to avoid getting infinite output by applying log on a zero value or exceedingly small values [3].

3.2 Regression Models

We wanted to build a regression model that can use the information about protein intake from various kinds of food to predict the COVID-19 recovery rate so that one could use this model to evaluate how helpful a diet with a certain protein profile is for recovering from COVID-19. To practice what we learned in the second half of the course, we built three tree-based regression models and compared their performance in terms of MSE. We split the data into training and test set at a ratio of 0.8, and we made predictions and calculated MSE using the test data. Note that we needed to apply inverse log transformation and minus 1 to both the predicted values and the transformed target variable before we calculated the MSE.

3.2.1 Decision Tree

The most important hyper-parameter of a decision tree in R is the complexity parameter (cp) which determines the depth of the tree. We needed to determine the optimal cp value to build the optimal tree for our data. Therefore, we first built an overly complicated tree with a cp value of 0.0000001 and observed the out-of-sample performance of the tree as a function of cp value. We picked the optimal cp value (0.08) that leads to the lowest out-of-sample error. Then, we built the optimal tree with the training data using the optimal cp. In addition to the predictive model, we also built a decision tree fitted with the entire dataset for data analysis purposes given the fact that the decision tree has good interpretability, and we can easily visualize the tree in R. So, we

can observe the splitting criterion at each node and determine what predictors are important and how they predict the target variable.

3.2.2 Random Forest

Based on what we covered in class, we only focused on the most important hyperparameter of Random Forest, the number of trees in the forest (ntree) while leaving other parameters as default. Since the bootstrapping technique used by Random Forest is stochastic, we set a fixed seed to reproduce the same forest when we compared the results. We started tuning the ntree parameter by trying values on a large scale such as 100, 500, 1000, 2000, 10000, etc. After identifying the general trend of the model performance for the value of ntree, we zoomed in and fine-tuned it with smaller intervals like 10. The optimal ntree was determined by the balance of out-of-bag (OOB) MSE and the test set MSE. Finally, the optimal number of trees in the forest was found to be 440, and we built the forest with the training data and calculated the MSE using the test data.

3.2.3 Gradient Boosting

Gradient Boosting Forest is similar to Random Forest but each tree in it is a simple tree, and the trees are built sequentially so that they learn from the previous tree. All simple trees come together to form an intelligent forest. Therefore, in addition to the number of trees (n.trees) parameter, we also tuned the interaction.depth parameter which determines the complexity of each simple tree. The tuning of the n.trees parameter followed the same procedures as for Random Forest. For interaction.depth parameter, we tried the values 2, 3, 4, and 5. The optimal set of parameters was found to be n.trees = 430 and interaction.depth = 3. We built the Gradient Boosting Forest fitted by training data using these two optimal parameters and calculated the MSE using the test data.

3.3 Regression Model Comparison

Given the nature of the data and the modeling of the model issue corrections (skewness correction, Multicollinearity, Outlier Test), the models have improved significantly, especially gradient boosting. Based on Table 3 below, Gradient boosting is our champion model for regression task. The result is consistent with what we have learned in class that Gradient Boosting forest is an improvement of Random Forest which in turn is an improvement of single decision tree. Boosting has become one of the two most powerful AI predictive techniques along with

Neural Networks. Note that the MSEs of all the models are not decently low, this is because of the nature of the data we used and the limited number of observations in the dataset.

Table 3: Model comparison using test set MSE			
Model	Decision Tree	Random Forest	Gradient Boosting
Before Corrections	4.36	2.71	2.23
After Corrections	2.39	1.65	1.48

3.4 Principal Component Analysis

The main purpose of doing principal component analysis (PCA) was to perform descriptive analysis on what are the good sources of protein that could help recover from COVID-19. PCA is a low-dimensional representation of the data that captures as much information as possible. It allows us to quickly analyze the distribution of observations in lower dimensions. First, we categorized the observations into three categories based on the recovery rate, namely high, medium, low recovery rate. We used the quantiles of the recovery rate column to delineate these three categories. Given that the recovery rate column is highly right-skewed (most countries have low recovery rates), the recovery rates below the median (~0.5 %) were categorized into low recovery rates. The ones from the median to the 3rd quantile (3%) were considered medium recovery rates. And the rest, above the 3rd quantile, were all high recovery rates. Then, we perform PCA on the features only, excluding the target variable. It turned out that the first two principal components capture about 65% of the variation of the data which is enough for the analysis. So, we plotted the PCA result with only the first two principal components (Figure 1) for analysis.

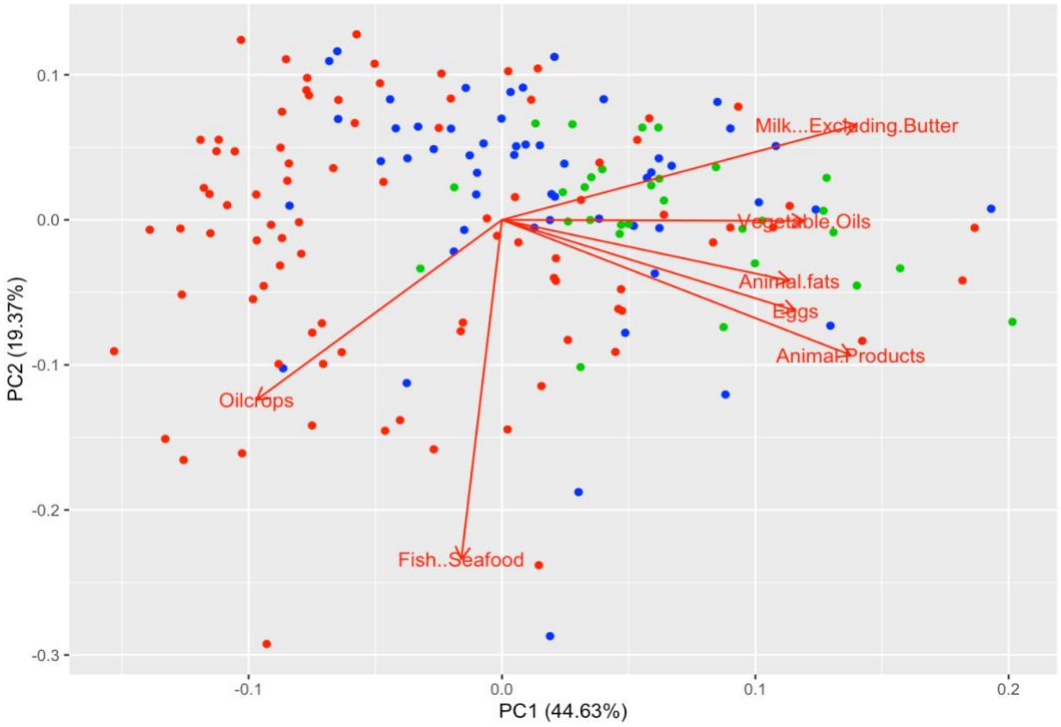

Figure 1: PCA plot with 2 principal components indicating High (Green), Medium (Blue), and Low (Red) Recovery Rate

4. Conclusion

One goal of this project was to investigate the protein profile for diets that help COVID-19 patients recover. In the PCA plot in Figure 1, the red arrows represent the variables and indicate the direction at which the variables increase. The center of the arrows represents the average of each variable. Most green points, corresponding to high recovery rate, are located at the above-average region for the variables Milk-Excluding Butter, Vegetable Oils, Animal fats, Eggs, and Animal Products and low region for Oil crops comparing to the red points, corresponding to low recovery rate. Meanwhile, Fish & seafood does not contribute much to the recovery rate. Therefore, we can conclude that diets with high protein intake from Milk-Excluding Butter, Vegetable Oils, Animal fats, Eggs, and Animal Products and low protein intake from Oil crops would be helpful for recovery from COVID-19. Appendix A provides an exhaustive list of food in each food categories which allows people from different countries and different regions to choose their diets despite that the food availabilities are different across countries and regions.

The other goal of this project was to build a model that can predicts the recovery rate of a country based on the diet of the country. Different regression models were tested to reach the final model that maximizes accuracy. Our lowest MSE for the recovery regression was 1.48 from Gradient Boosting which was the most accurate model we can achieve with this dataset.

We wanted to push the boundaries of this project therefore we decided that despite the lack of hierarchy on the data we wanted to be able to analyze the data on a higher level of analysis. In this scenario, we wanted to analyze landmass but no column has a landmass. To bypass this, we researched and found a package called countrycode which converts country names into landmasses. Using this package, we were able to use the landmasses at a higher level and aggregate the data to a higher level of analysis. This approach enables us to use a bottom-up modeling approach of predicting the recovery rate of each country then aggregating it if we want to. This also enables us to run more analysis more holistically if we decide to take a top-down approach as well (Table 4). Based on the aggregation we believe that European countries have the highest recovery rate. Overall Europe has the highest consumption of all products except for vegetal products and oil crops.

Table 4: Average data over landmasses (in percentage)

	Africa	Americas	Asia	Europe	Oceania
Recovered	0.35	1.46	1.33	3.02	0.69
Animal Products	12.97	24.56	19.58	28.43	26.15
Vegetable Oils	0.01	0.03	0.01	0.04	0.02
Animal fats	0.03	0.08	0.07	0.26	0.13
Vegetal Products	37.03	25.44	30.42	21.57	23.85
Fish & Seafood	2.76	3.02	3.85	2.96	7.02
Milk-Excluding Butter	2.89	6.11	5.10	10.63	3.75
Oil crops	2.22	1.04	1.31	0.49	2.50
Eggs	0.40	1.38	1.30	1.75	0.91

Appendix

A. The food in each selected feature (food category)

Variable	What each food category is composed of (from FAO.org)
Animal Products	Aquatic Animals, Others; Aquatic Plants; Bovine Meat; Butter, Ghee; Cephalopods; Cream; Crustaceans; Demersal Fish; Eggs; Fats, Animals, Raw; Fish, Body Oil; Fish, Liver Oil; Freshwater Fish; Marine Fish, Other; Meat, Aquatic Mammals; Meat, Other; Milk - Excluding Butter; Molluscs, Other; Mutton & Goat Meat; Offals, Edible; Pelagic Fish; Pig meat; Poultry Meat
Vegetable Oils	Coconut Oil; Cottonseed Oil; Groundnut Oil; Maize Germ Oil; Oil crops Oil, Other; Olive Oil; Palm Oil; Palm kernel Oil; Rape and Mustard Oil; Rice bran Oil; Sesame seed Oil; Soyabean Oil; Sunflower seed Oil
Animal Fats	Butter, Ghee; Cream; Fats, Animals, Raw; Fish, Body Oil; Fish, Liver Oil
Vegetal Products	Alcohol, Non-Food; Apples and products; Bananas; Barley and products; Beans; Beer; Beverages, Alcoholic; Beverages, Fermented; Cassava and products; Cereals, Other; Citrus, Other; Cloves; Cocoa Beans and products; Coconut Oil; Coconuts - Incl Copra; Coffee and products; Cottonseed; Cottonseed Oil; Dates; Fruits, Other; Grapefruit and products; Grapes and products (excl wine); Groundnut Oil; Groundnuts (Shelled Eq); Honey; Infant food; Lemons, Limes and products; Maize and products; Maize Germ Oil; Millet and products; Miscellaneous; Nuts and products; Oats; Oil crops Oil, Other; Oil crops, Other; Olive Oil; Olives (including preserved); Onions; Oranges, Mandarines; Palm kernels; Palm Oil; Palm kernel Oil; Peas; Pepper; Pimento; Pineapples and products; Plantains; Potatoes and products; Pulses, Other and products; Rape and Mustard Oil; Rape and Mustard seed; Rice (Milled Equivalent); Rice bran Oil; Roots, Other; Rye and products; Sesame seed; Sesame seed Oil; Sorghum and products; Soyabean Oil; Soyabeans; Spices, Other; Sugar (Raw Equivalent); Sugar beet; Sugar cane; Sugar non-centrifugal; Sunflower seed; Sunflower seed Oil; Sweet potatoes; Sweeteners, Other; Tea (including mate); Tomatoes and products; Vegetables, Other; Wheat and products; Wine; Yams
Fish & Seafood	Cephalopods; Crustaceans; Demersal Fish; Freshwater Fish; Marine Fish, Other; Molluscs, Other; Pelagic Fish
Milk Excluding Butter	Just Milk and no butter
Oil Crops	Coconuts - Incl Copra; Cottonseed; Groundnuts (Shelled Eq); Oil crops, Other; Olives (including preserved); Palm kernels; Rape and Mustard seed; Sesame seed; Soyabeans; Sunflower seed
Eggs	Eggs

B. Figures

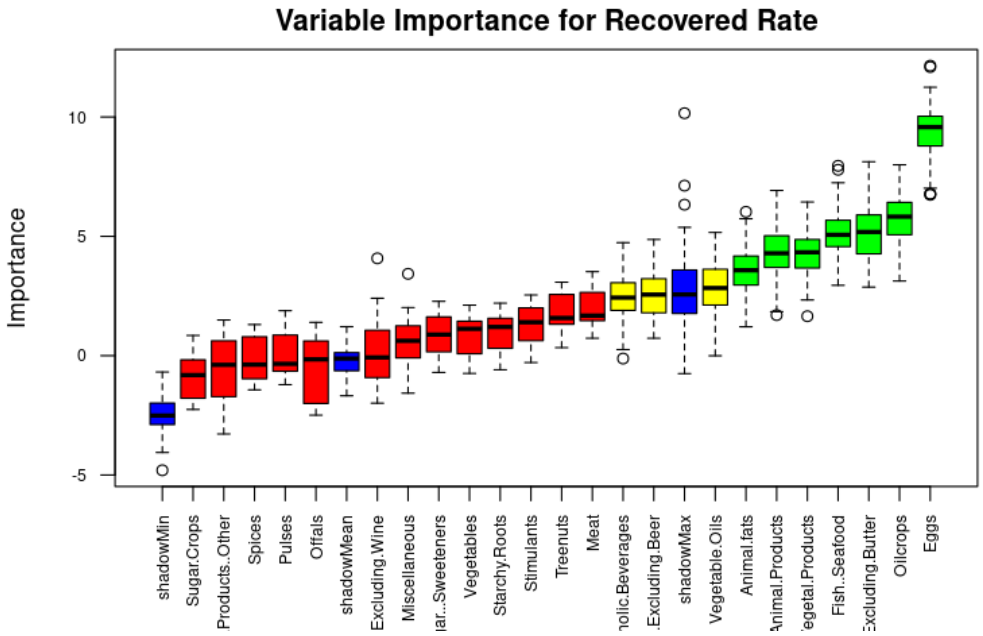

Figure 2: Feature Selection Using Boruta

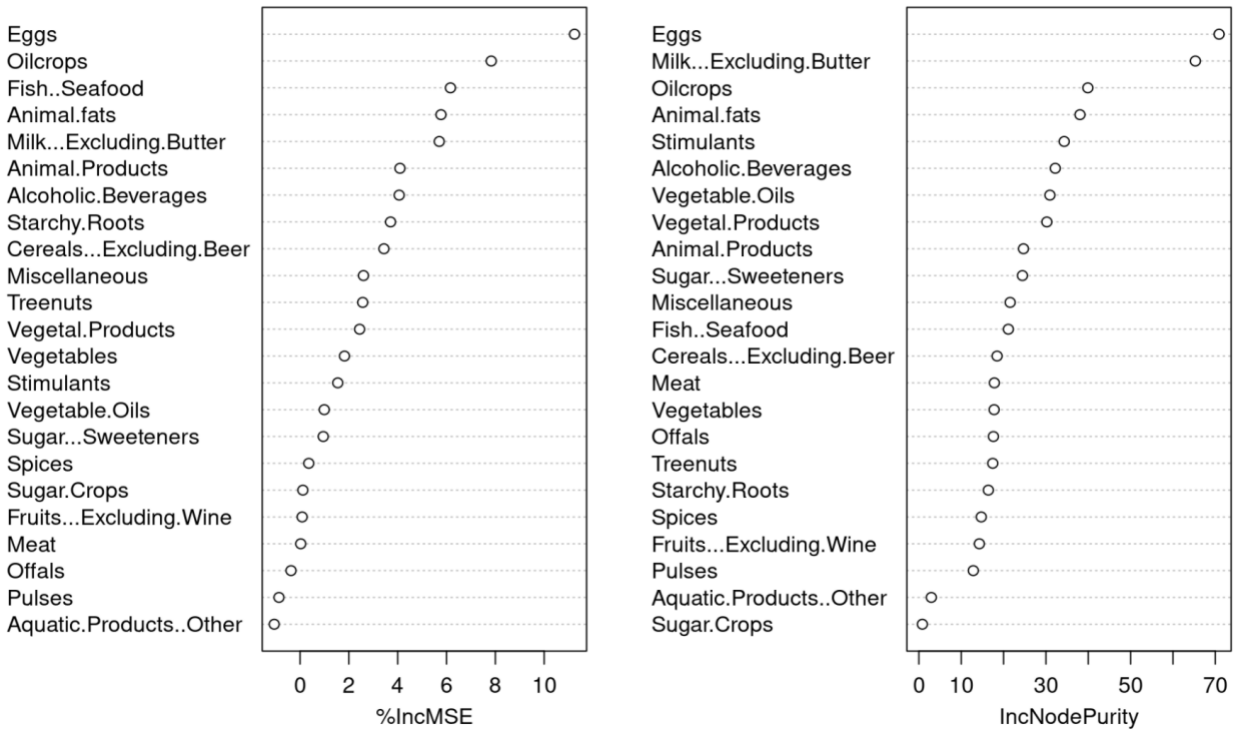

Figure 3: Feature Selection Using Random Forest

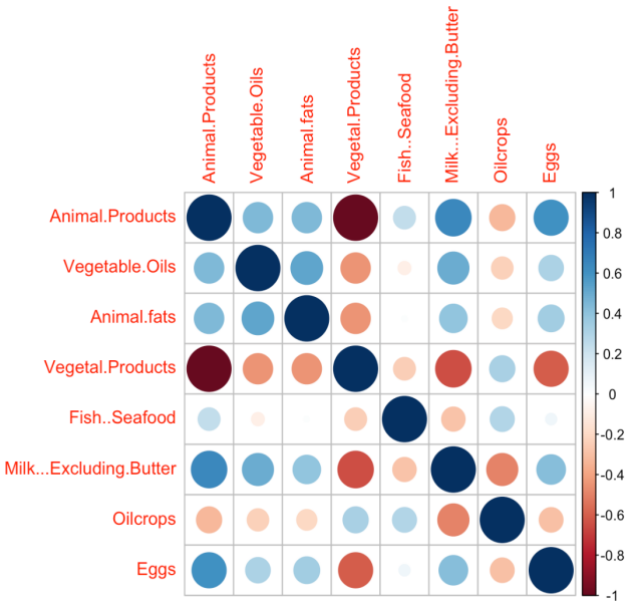

Figure 4: Correlation Coefficient Heap Map for Collinearity Test

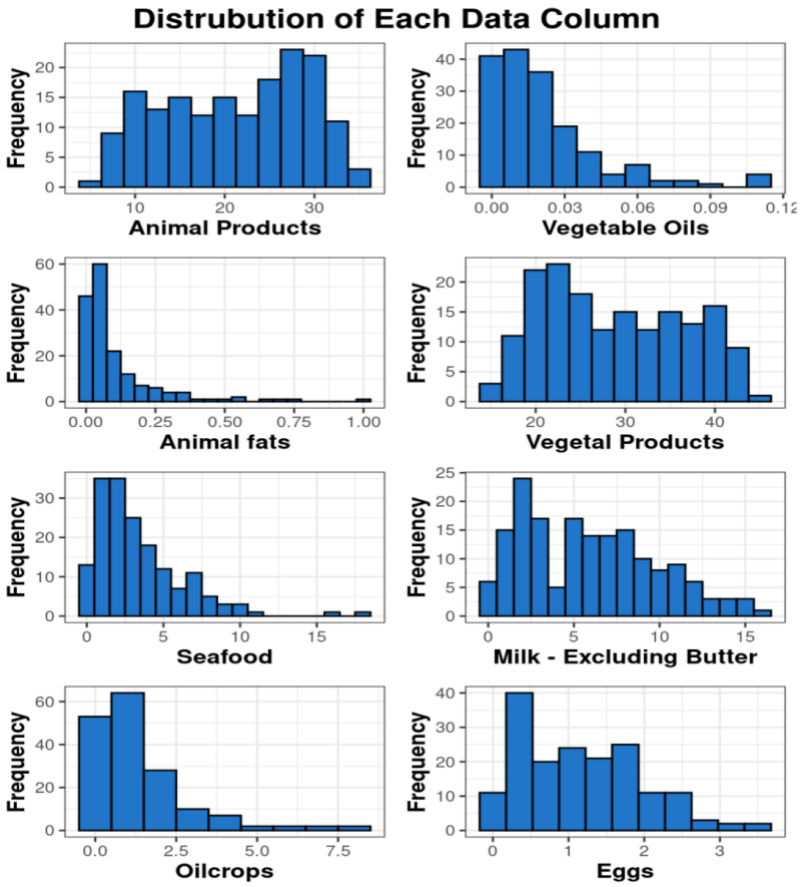

Figure 5: Histogram of Predictors (% protein intake from food categories)

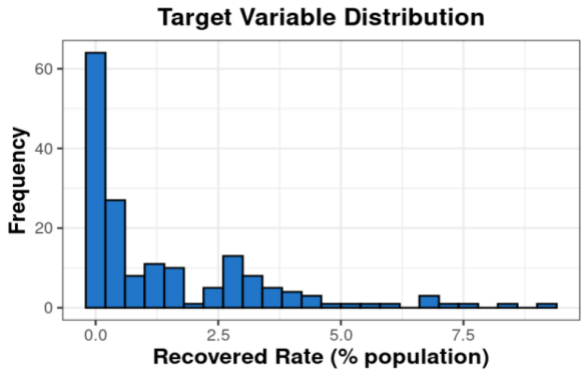

Figure 6: Histogram of Target Variable (% Recovery rate of population)

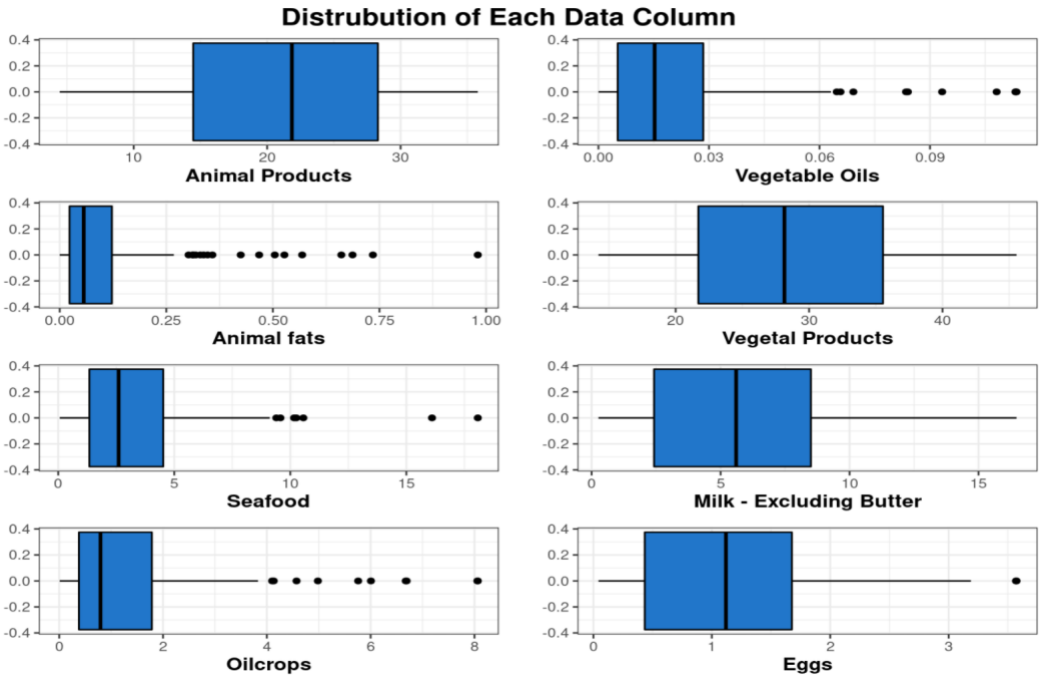

Figure 7: Box Plot of Predictors (% protein intake from food categories)

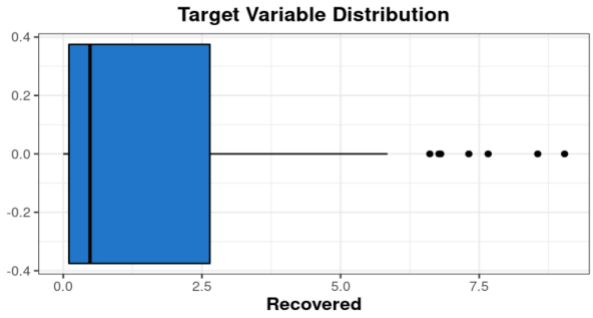

Figure 8: Box Plot of Target Variable (% Recovery rate of population)

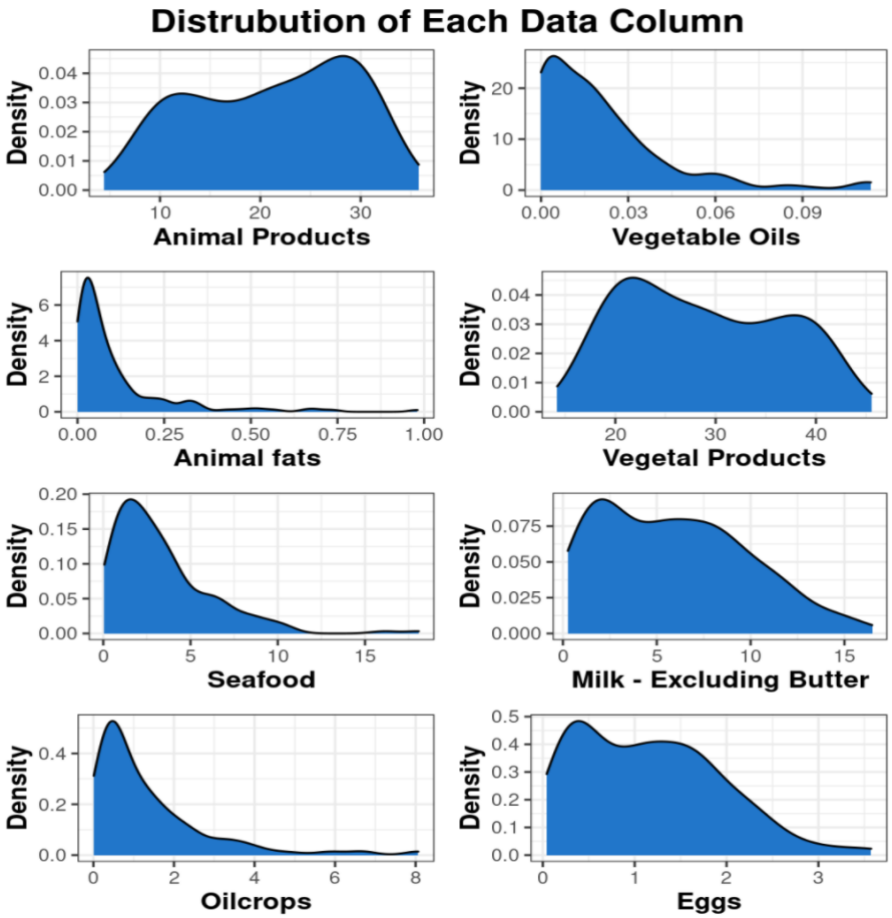

Figure 9: Density Plot of Predictors (% protein intake from food categories)

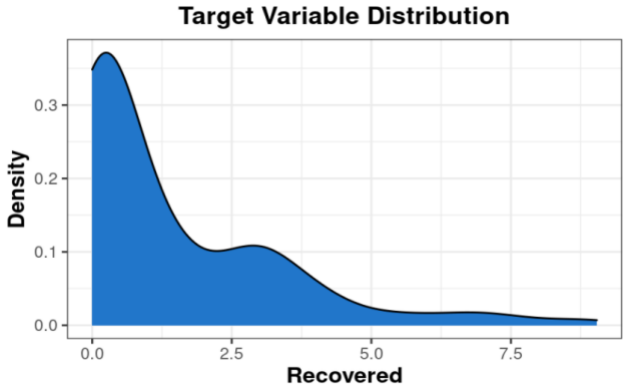

Figure 10: Density Plot of Target Variable (% Recovery rate of population)

Bibliography and References

[1] "Comparing Random Forest and Gradient Boosting." *Medium*, 1 Nov. 2021, towardsdatascience.com/comparing-random-forest-and-gradient-boosting-d7236b429c15#: ~:text=This%20means%20tree. Accessed 16 Dec. 2021.

[2] "Feature Transformations in Data Science: A Detailed Walkthrough." *Analytics Vidhya*, 6 May 2021, www.analyticsvidhya.com/blog/2021/05/feature-transformations-in-data-science-a-detailed-walkthrough/. Accessed 16 Dec. 2021.

[3] "Machine Learning - Why Log-Transform to Normal Distribution for Decision Trees?" *Cross Validated*, stats.stackexchange.com/questions/385231/why-log-transform-to-normal-distribution-for-decision-trees.

[4] "Classification - Should One Be Concerned about Multi-Collinearity When Using Non-Linear Models?" *Cross Validated*, stats.stackexchange.com/questions/266267/should-one-be-concerned-about-multi-collinearity-when-using-non-linear-models#: ~:text=Multi-collinearity%20will%20not%20be%20a%20problem%20for%20certain. Accessed 16 Dec. 2021.

[5] Evans, Nick. "How Nutrition Can Aid Recovery from COVID-19." *Nursing Standard*, vol. 35, no. 8, 5 Aug. 2020, pp. 35–37, 10.7748/ns.35.8.35.s16. Accessed 27 Aug. 2020.

[6] Wu, Guoyao. "Dietary Protein Intake and Human Health." *Food & Function*, vol. 7, no. 3, 2016, pp. 1251–1265, pubmed.ncbi.nlm.nih.gov/26797090/, 10.1039/c5fo01530h.