STAT 425: Applied Regression and Design
Spring 2021

Final project report

COVID-19 Healthy Diet and Recovery Rate

Nursultan Baitlessov

Sandip Sonawane

University of Illinois, Urbana-Champaign

Table of Contents

1. Introduction

COVID-19 has caused the first severe pandemic of the XXI century that brought around 3.3 million fatal outcomes around the world as of May 2021. In this project, we will answer the following utterly important question during a pandemic time: Is a diet considered an important factor to protect ourselves from COVID-19? To answer this question, we will analyze the effect of a diet in different countries on COVID-19 recovery rate. The conclusion from this analysis will help us to understand what should be the optimal diet to mitigate the negative health effects from COVID-19. The dataset for this project was collected by Maria Ren and taken from Kaggle. This data was collected with an aim to learn more about how a healthy diet could help combat the COVID-19. The implications of our analysis could be crucial to understand what people should have in their diet to increase the possibility of recovery from COVID-19.

2. Exploratory Data Analysis

Originally, the dataset contained 31 variables, but we use 26 variables in our analysis. We have excluded the variables called *Population, Death, Confirmed and Active*, because they were highly correlated with our response variable called *Recoved*. In addition, a variable containing undernourished rates was excluded from our model, as it was highly correlated with a variable containing obesity rates for different countries. The variables we included in our analysis provide information about the country, percentage of energy intake in kcal from different types of food, obesity rate, and recovery rates from COVID-19. The following table briefly describes the type and a brief description of these variables:

Table 1. Variables description

Variable	**Type**	**Description**
Country	Categorical	Country in which observation was taken
Alcoholic.Beverages	Numerical	Percentage of energy intake from alcoholic beverages
Animal.Products	Numerical	Percentage of energy intake from animal products
Animal.fats	Numerical	Percentage of energy intake from animal fats
Aquatic.Products..Other	Numerical	Percentage of energy intake from aquatic products
Cereals...Excluding.Beer	Numerical	Percentage of energy intake from cereal, excluding beer
Eggs	Numerical	Percentage of energy intake from eggs
Fish..Seafood	Numerical	Percentage of energy intake from fish seafood

Fruits...Excluding.Wine	Numerical	Percentage of energy intake from fruits, excluding wine
Meat	Numerical	Percentage of energy intake from meat
Milk...Excluding.Butter	Numerical	Percentage of energy intake from milk, excluding butter
Miscellaneous	Numerical	Percentage of energy intake from miscellaneous products
Offals	Numerical	Percentage of energy intake from offals
Oilcrops	Numerical	Percentage of energy intake from oilcrops
Pulses	Numerical	Percentage of energy intake from pulses
Spices	Numerical	Percentage of energy intake from spices
Starchy.Roots	Numerical	Percentage of energy intake from starchy roots
Stimulants	Numerical	Percentage of energy intake from stimulants
Sugar.Crops	Numerical	Percentage of energy intake from sugar crops
Sugar...Sweeteners	Numerical	Percentage of energy intake from sugar and sweeteners
Treenuts	Numerical	Percentage of energy intake from treenuts
Vegetal.Products	Numerical	Percentage of energy intake from vegetal products
Vegetable.Oils	Numerical	Percentage of energy intake from vegetable oils
Vegetables	Numerical	Percentage of energy intake from vegetables
Obesity	Numerical	Obesity rate in percent
Recovered	Numerical	Percentage of people recovered from COVID-19

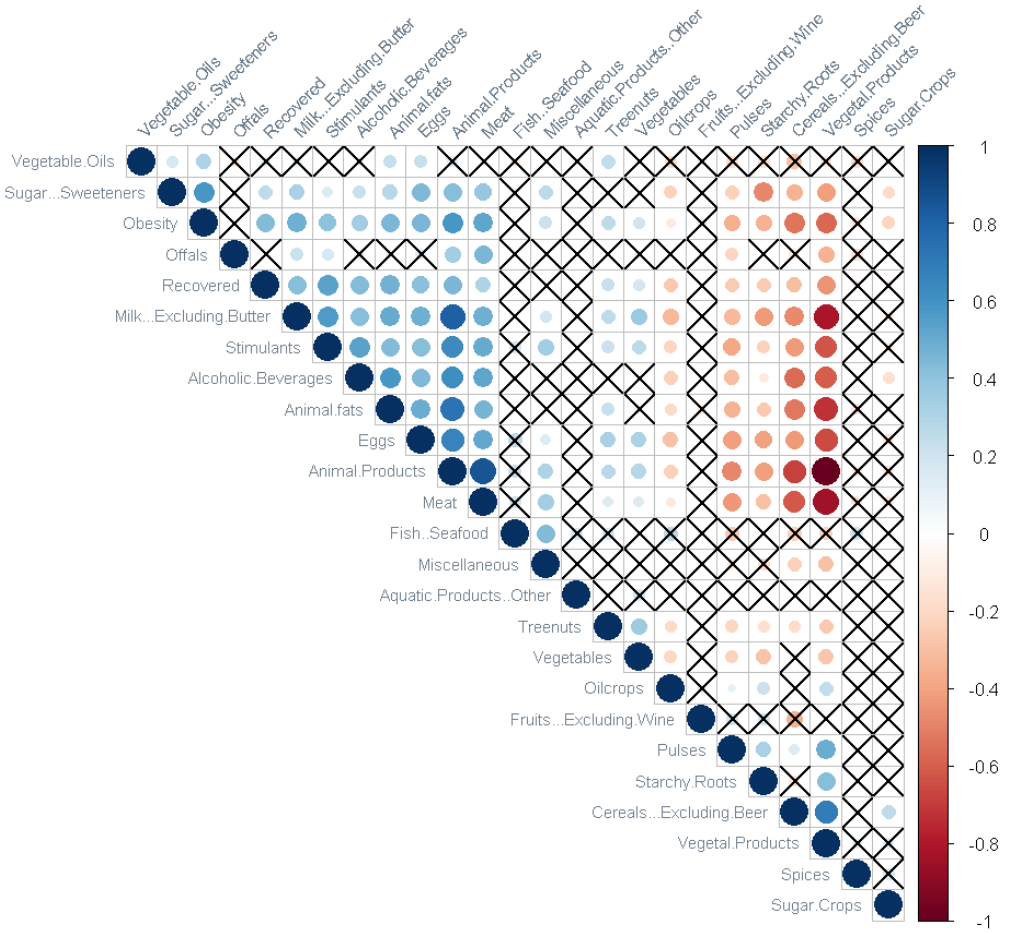

Figure 1. Correlation matrix of all variables

The above figure demonstrates a correlation magnitude between any two individual variables. Here, our target variable is Recovered. It is interesting to note that Vegetal.Products and Animal.Products (Milk, Meat, Eggs, Animal Fats) have a strong negative correlation. There are a few variables, e.g. Fish..Seafood, Miscellaneous and Aquatic.Products.Other that have a negligible effect on the recovery rate. From statistics theory, the linear model performs better on the test data when we have less parameters. Hence, we can remove the variables that do not have a linear relationship with the target variable. By looking at the correlation plot, we can see that the following variables in the training data have insignificant correlation with the recovery rate at 1% significance level and are denoted by the “crossed” symbol: 1. Aquatic..Products..Other; 2. Fish..Seafood; 3. Miscellaneous; 4. Fruits..Excluding..Wine; 5. Spices; 6. Sugar.Crops. Therefore we can remove these variables from our training data.

When the feature variables are highly correlated, they have almost the same effect on the target variable and cause a problem of multicollinearity in a linear model. In such cases, the inverse of our design matrix does not exist and we get infinitely many solutions for beta parameters of our linear model. Hence, in order to build a linear model, we need to ensure that our design matrix does not have highly correlated features. We have already mentioned that Animal.Products and Vegetal.Products are highly correlated according to the correlation plot and if we include both of these variables in our linear model it will

cause a multicollinearity. Since Animal.Products (e.g. Meat, Aquatic.Animals, Eggs, Milk, etc.) and Vegetal.Products (e.g. Alcohol, Oilcrops, Pulses, etc.) are just a combination of many variables that we already have in our model, we will drop these two variables to avoid multicollinearity problems. This step will ensure that the inverse of our design matrix exists.

After removing the mentioned variables, we obtain the following correlation plot:

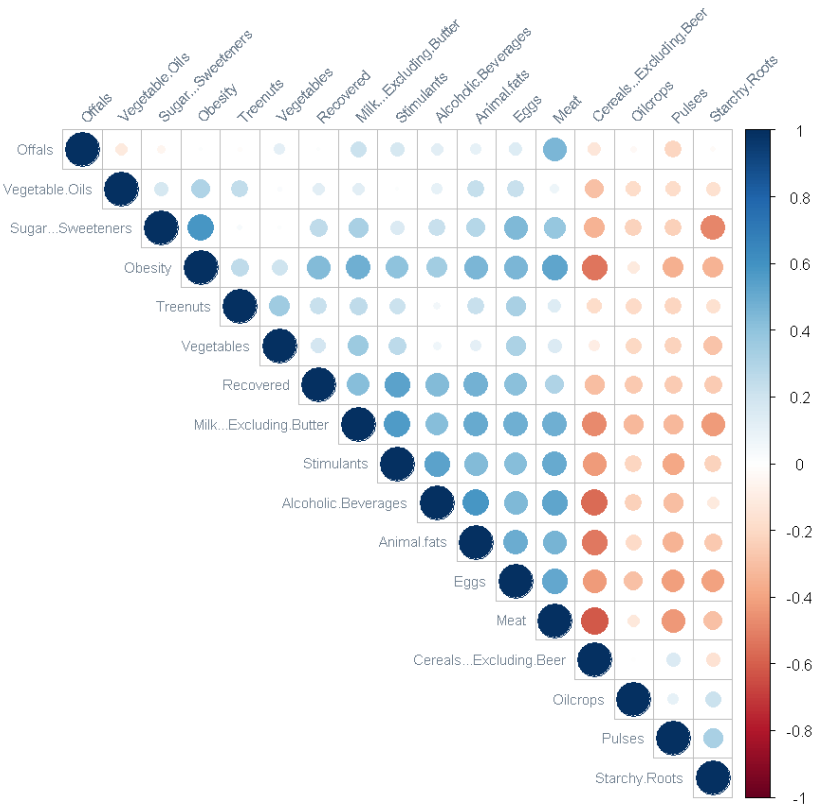

Figure 2. Correlation matrix of chosen variables

From Figure below, we can see that most of the countries include Vegetal Products (variable 8) in their diet. The second major food type included in the diet is cereals excluding beer. Animal Products is the third most included food item in the diet by various countries.

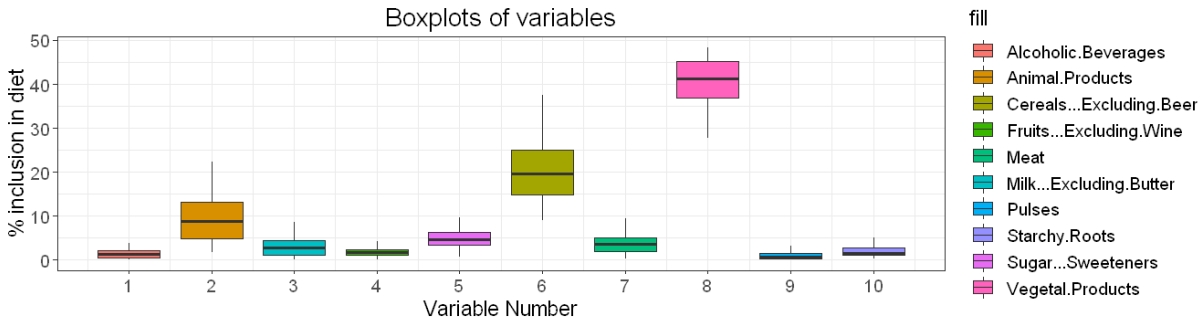

Figure 3. Boxplots of variables

3. Methodology

In this project, we are trying to solve a regression problem. There are two methods of regression: parametric regression and non-parametric regression. We will first build parametric regression models and test their performance. Afterwards, we will build non-parametric regression models. Finally, we will compare performance for all models.

3.1 Linear Regression

The first model that we have used in our project is the most simple, which is a linear regression model. To assess the performance of the model, we separated our data into training (80%) and testing (20%) sets. After we constructed the model with the response variable *Recovered* and with explanatory variables containing all the variables, except those that were excluded in Exploratory Data Analysis part, we also excluded *Cereals...Excluding.Beer* variable, which had a very high Variance Inflation Factor (VIF) = 17.78, to avoid multicollinearity in our model. Afterwards, we performed regression diagnostics and found that the residuals were not normally distributed according to the Shapiro-Wilk test and heteroskedasticity was present in our model according to Breush-Pagan test. These two factors violate the assumptions for linear regression modelling, so we had to use a transformation of the response variable and robust standard errors to avoid misleading results. The inverse Box-Cox transformation was used and 1 was added to remove the possibility of dividing by 0 in testing set prediction, so we got 1/(Recovered + 1) as our Response variable. Using the robust standard errors to combat heteroskedasticity did not change the significance of our variables and considering the fact that we have sufficiently large training sample size (n = 131), we expect that the results are precise enough with non-robust standard errors.

1) The formula for our linear regression model is:

1 / (Recovered + 1) ~ Alcoholic.Beverages + Animal.fats + Eggs + Meat + Milk...Excluding.Butter + Offals + Oilcrops + Pulses + Starchy.Roots + Stimulants + Sugar...Sweeteners + Treenuts + Vegetable.Oils + Vegetables + Obesity

Train RMSE: 0.224
Test RMSE: 0.239

We got a good performance for this model. Next, we will assess the performance of the reduced models using variable selection by stepwise elimination using BIC and AIC criterias, lasso regression and Principal Components regression.

2) The formula for linear regression with variables selected using BIC stepwise elimination:

1 / (Recovered + 1) ~ Oilcrops + Stimulants + Obesity

Train RMSE: 0.241
Test RMSE: 0.251

3) The formula for linear regression with variables selected using AIC stepwise elimination:

1 / (Recovered + 1) ~ Animal.fats + Eggs + Meat + Oilcrops + Stimulants + Obesity

Train RMSE: 0.228
Test RMSE: 0.232

4) The Lasso regression is used for variable selection and it shrinks the estimated coefficient towards 0, which helps us to choose the coefficients that are different from zero and make our estimated coefficients to work better on testing set, but at the same time avoid over-fitting. The formula for model with variables selected using Lasso with lambda = 0.71:

1 / (Recovered + 1) ~ Alcoholic.Beverages + Animal.fats + Eggs + Meat + Milk...Excluding.Butter + Oilcrops + Starchy.Roots + Stimulants + Sugar...Sweeteners + Treenuts + Vegetable.Oils + Vegetables + Obesity

Train RMSE: 0.227
Test RMSE: 0.241

5) Next, we use Principal Components Analysis to reduce the dimensionality of our dataset by transforming our set of variables to a smaller set that contains most of the information of the larger set. In Principal Components regression, we do not need to worry about the multicollinearity problem, as the Principal Components are mutually orthogonal. 10 Principal Components were selected for Principal Components regression using Cross-Validation:

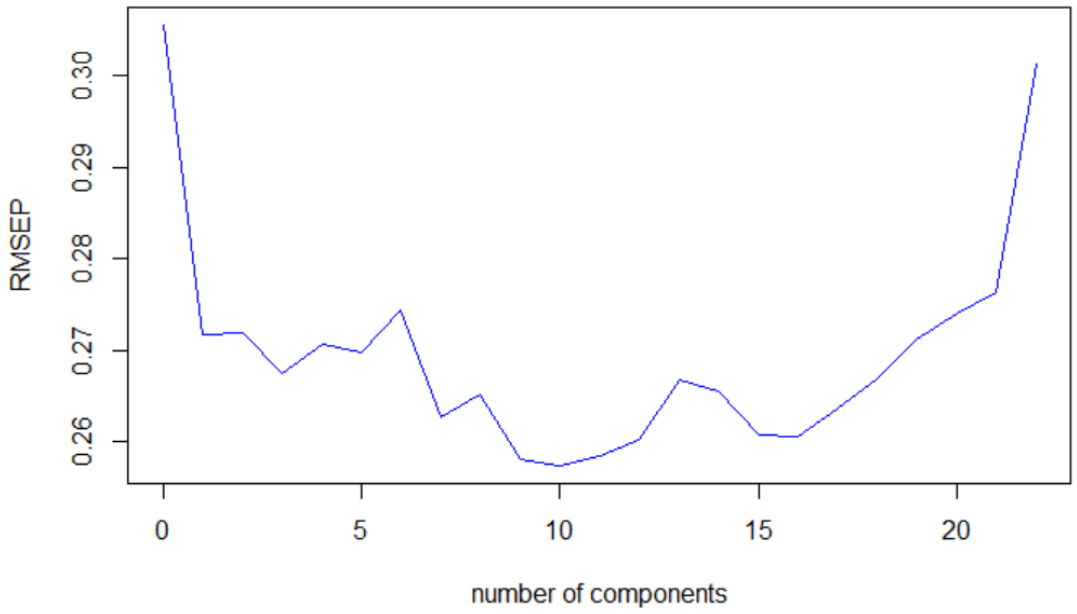

Figure 4. Root Mean Squared Error of Prediction (RMSEP) vs. number of components

The model with 10 Principal Components has yielded the following results:

Train RMSE: 0.236
Test RMSE: 0.252

From these results using different techniques to select variables, we can see that the linear regression model with variables selected using AIC criteria performed the best both in terms of Train RMSE and Test RMSE. Therefore, we are going to interpret the results from this model.

Table 2. Coefficients for linear regression model with variables selected using AIC criteria

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.831390   0.055457  14.992  < 2e-16 ***
Animal.fats -0.036455   0.018276  -1.995 0.048287 *
Eggs        -0.185796   0.084214  -2.206 0.029224 *
Meat         0.033251   0.012860   2.586 0.010885 *
Oilcrops     0.047403   0.014977   3.165 0.001954 **
Stimulants  -0.234440   0.077239  -3.035 0.002934 **
Obesity     -0.010224   0.002627  -3.892 0.000162 ***
```

By looking at the Table above, we can see that having animal fats, eggs and stimulants in the diet is predicted to positively affect the recovery rate from COVID-19, whereas meat and oilcrops are expected to negatively impact the recovery rate. In regards to obesity rate, we get a counter-intuitive result, as the higher obesity rate should decrease the recovery rate. This could happen, as we could have an omitted variable bias, as the error term could contain the variable that is highly correlated with obesity rate. Therefore, we disregard the implication regarding the positive impact of high obesity rate. The model's coefficients can be interpreted in the following way: 1 unit increase in percentage of energy intake from stimulants holding other factors constant is predicted to decrease 1/(Recovered+1) by 0.2344 that means that this change increases the recovery rate.

3.2 Non-Parametric Regression Models

In non-parametric regression methods, the relationship between predictor and response variables is defined based on information derived from the data unlike an equation in parametric (multiple linear) regression methods. We will train below four non-parametric methods and assess their performance in this section.

1. Kernel Regression
2. KNN Regression
3. Decision Tree Regression
4. Extreme Gradient Boosting Regression

The dataset is split into training(80%) and testing sets(20%). The training set is further split into estimation(80%) and validation set(20%). For cross-validation, 5 fold cross-validation is done..

3.2.1 Kernel Regression

Kernel regression is a non-parametric regression algorithm. We will use the Gaussian kernel for our model. Here, the tuning parameter is lambda, which is also called bandwidth. The bandwidth controls the flexibility of the model. Smaller values of bandwidth makes the model more sensitive and it becomes overfit if the value of bandwidth is too small. Values of bandwidth are tuned from 0.01 to 0.1 with a difference of 0.005. Below results are obtained for reduced and full model.

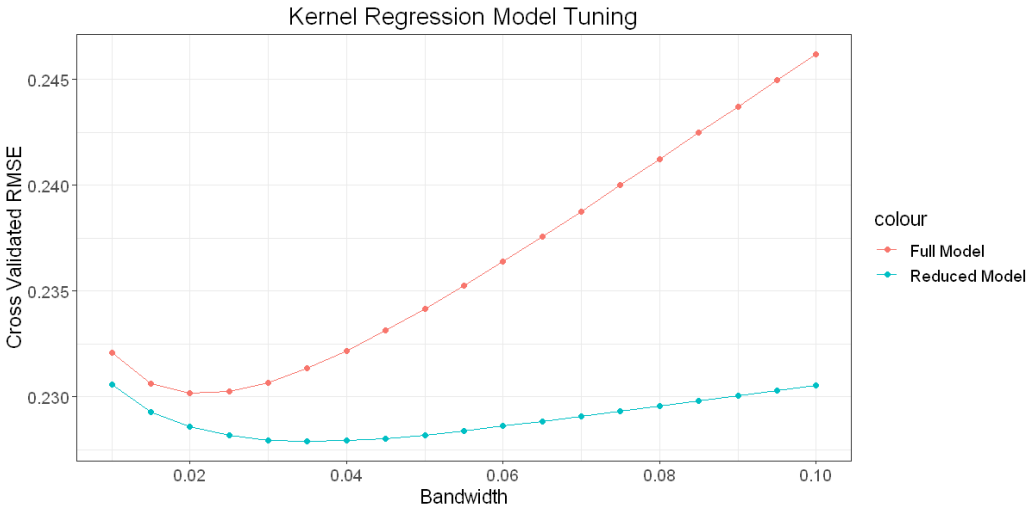

Figure 5. Kernel Regression Model Tuning

We can observe from the above plot that cross-validated RMSE is reduced slowly as the value of bandwidth increases initially. After a certain threshold, increasing bandwidth does not result in reduction in cross validated RMSE.

Table 3. Kernel Regression Model Results

Model	**Formula**	**Best Bandwidth**	**Cross Validation RMSE**	**Test RMSE**
Kernel Full Model	*1/ (Recovered+1) ~.*	0.02	0.230	0.252
Kernel Reduced Model	*1 / (Recovered+1) ~ Animal fats + Sugar Sweeteners + Oil Crops + Stimulants + Obesity + Eggs + Tree Nuts*	0.035	0.227	0.263

3.2.2 KNN Regression

K-nearest neighbours (KNN) Regression is a very simple nonparametric regression algorithm. The prediction for a new record is made based on values of K-nearest points based on cartesian distance by default. The performance of KNN models depends upon the value of K. Smaller values of K tend to overfit the data, resulting in low bias but high

variance. Choice of K should be done such that we achieve low bias as well as low variance considering bias-variance trade off. It is also a good idea to choose minimum features for the KNN algorithm since it avoids the curse of dimensionality.

The dataset is split into training(80%) and testing sets(20%). The training set is further split into estimation(80%) and validation set(20%). For cross-validation, 5 fold cross-validation is done. Values of K are tuned from 1 to 30. Below results are obtained.

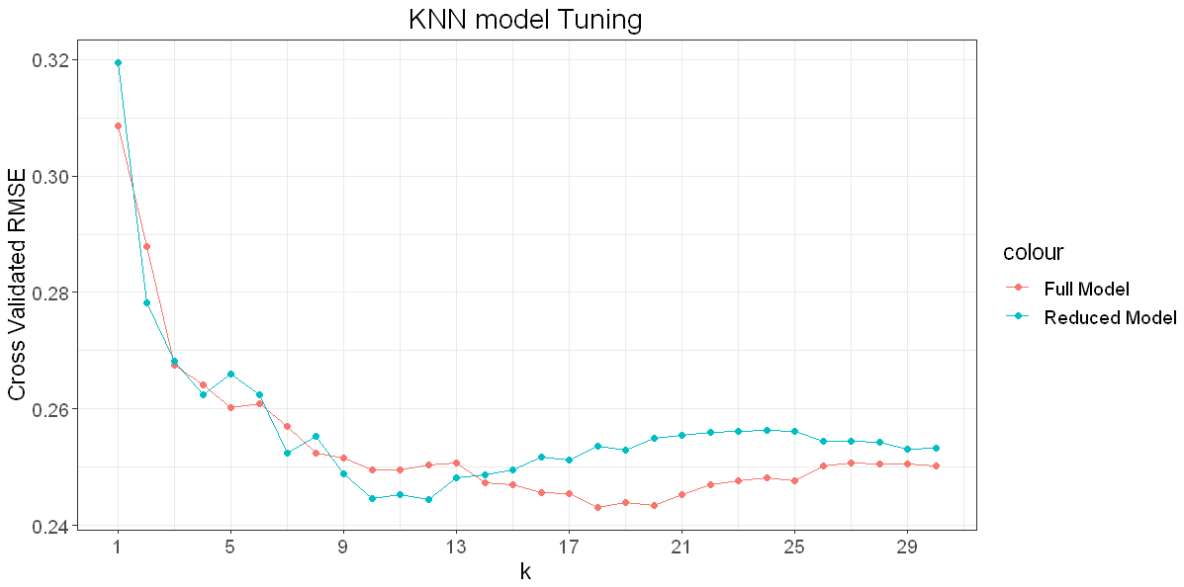

Figure 6. KNN model tuning

When the value of k is low, the model does not perform well on the validation set. As the value of K increases, cross-validated RMSE keeps reducing and it becomes flat after the value of k crosses a certain threshold. Here, the threshold is 26.

Table 4. KNN Model Results

Model	**Formula**	**Best K**	**Cross Validation RMSE**	**Test RMSE**
KNN Full Model	*1/ (Recovered+1) ~.*	18	0.243	0.25
KNN Reduced Model	*1/ (Recovered+1) ~ Tree Nuts Alcoholic.Beverages + Oil Crops + Stimulants + Obesity + Eggs*	12	0.244	0.26

3.2.3 Decision Tree Regression

Decision tree regression is another nonparametric regression algorithm. For a decision tree, to avoid overfitting, and to be able to generalize well, we prune the decision tree using below criteria.

minimize{RSS + cp TreeSize}*

This idea is similar to the regularization problem solved by using ridge or lasso regression. When we use a lower value of complexity parameter (cp), a deeper tree having many nodes is created. Here, we have tuned the values of cp from 0 to 0.15. First, a decision tree model considering all features was trained. Below features were identified as important.

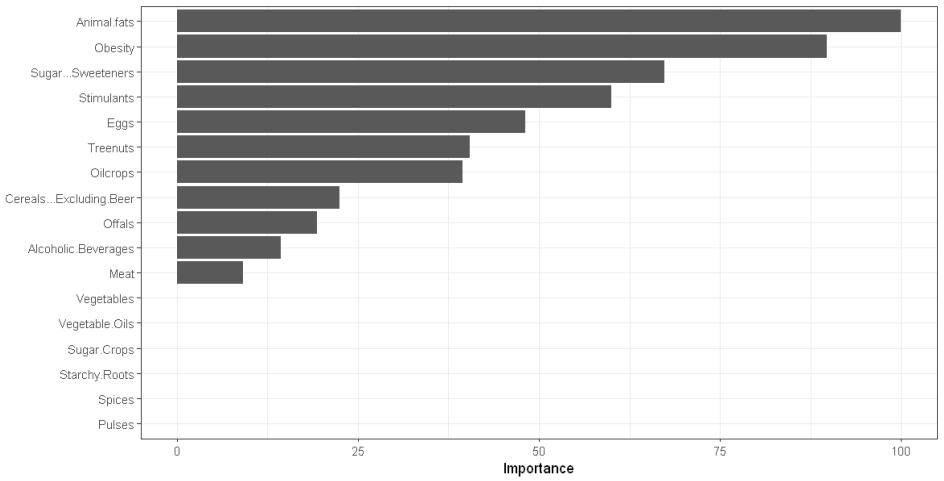

Figure 7. Decision tree variable importance

After looking at the variable importance plot, first 6 variables were chosen for training a reduced model. Below plot shows variation of cross validated RMSE with values of cp for full and reduced models.

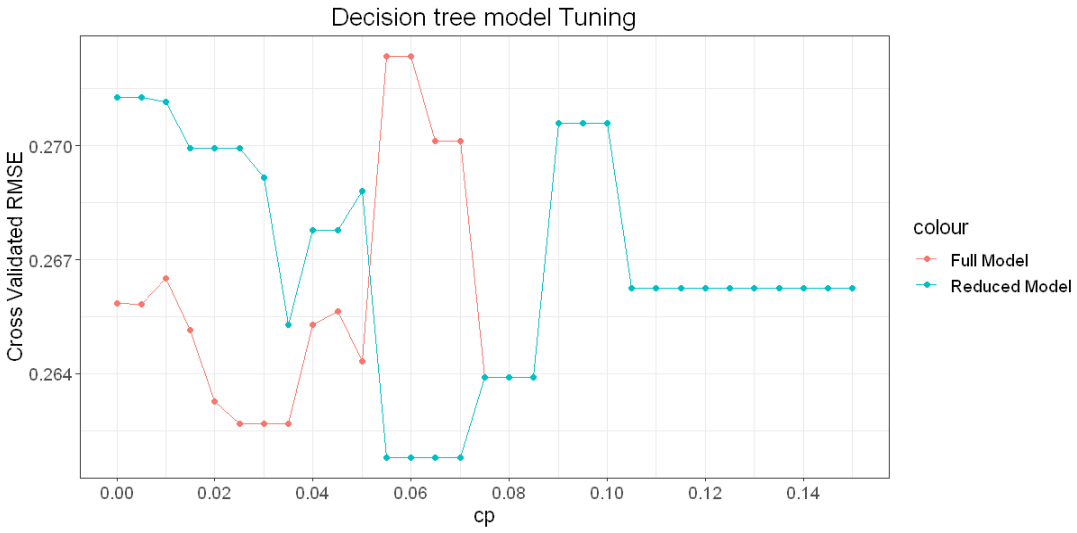

Figure 8. Decision tree model tuning

Below figure shows decision boundaries for recovery prediction based on the reduced model with value of cp equal to 0.07. We can see that the decision tree is not very large or complex and also gives lower test RMSE.

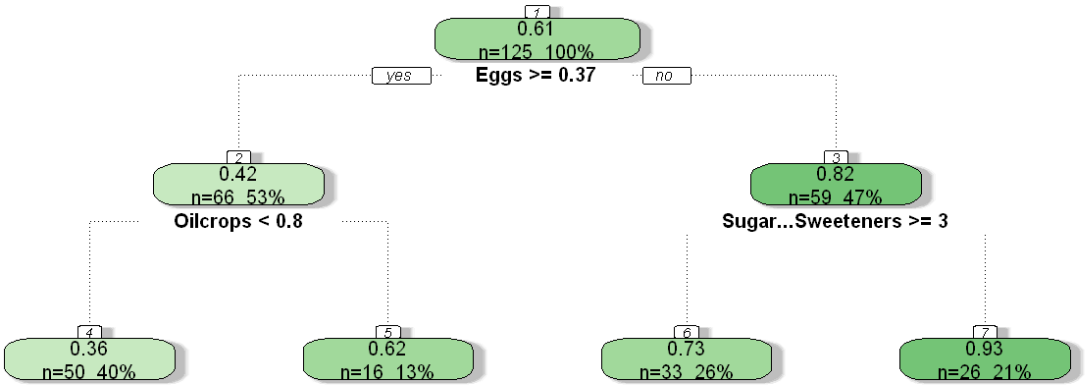

Figure 8. Decision tree decision boundary

Table 5. Decision Tree Model Results

Model	**Formula**	**Best cp**	**Cross Validation RMSE**	**Test RMSE**
Decision Tree Full Model	*1/ (Recovered+1) ~.*	0.035	0.262	0.266
Decision Tree Reduced Model	*1 / (Recovered+1) ~ Tree Nuts Alcoholic.Beverages + Oil Crops + Stimulants + Obesity + Eggs*	0.05	0.261	0.258

3.2.4 XGBoost Regression

XGBoost (Extreme Gradient Boosting) is an ensemble learning algorithm. The model learns from previously grown trees and builds subsequent trees. After training with all variables, below variables are found important in predicting recovery.

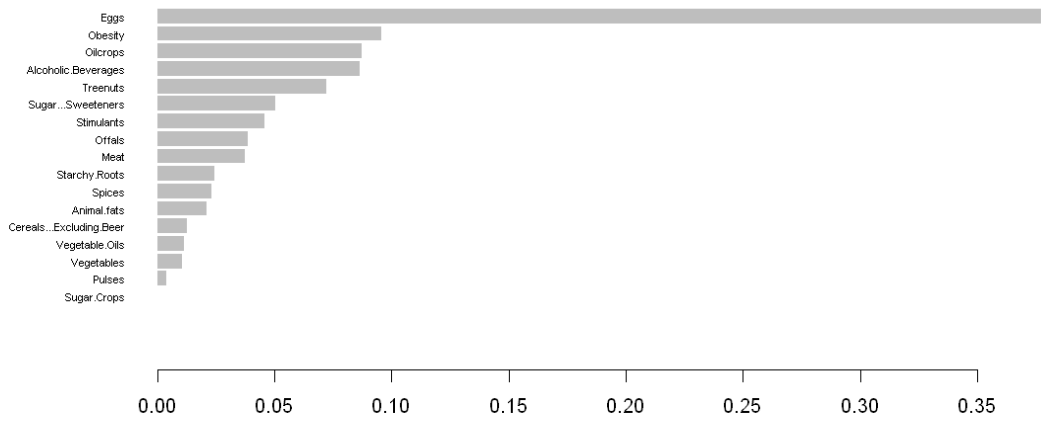

Figure 9. XGBoost variable importance

Table 6. XGBoost Model Results

Model	**Formula**	**Cross Validation RMSE**	**Test RMSE**
XGBoost	*1/ (Recovered+1) ~.*	0.035	0.249

From the table, it can be seen that the model obtained very low cross-validation RMSE as well as lowest test RMSE among all the non-parametric models tried. This proves that ensemble models perform better than an individual model.

4. Discussion and conclusions

As we may already know from nutrition policies and promotion, the diet plays an important role in building a strong immunity to resist different illnesses and viruses including COVID-19. In this project, we analyzed the effect of diet in different countries on COVID-19 recovery rate. We built and trained parametric as well as non-parametric regression models. Below are the key outcomes from this analysis.

- Eggs is the most important factor which is related to recovery followed by obesity.
- Extreme Gradient Boosting based model achieved lowest cross-validated RMSE.

Through this project, we implemented the concepts learned from the course Applied Regression and Design (STAT 425). The learnings from this project will help us solve industry problems through regression techniques.