Final report

Analysis on the factors that affect COVID-19 death rate

김경수

이채영

전민식

1.Introduction

The world is suffering from COVID-19 pandemic. The vaccine is yet to be made and many scientists are expecting a long term until the crCeation of a successful vaccine. The COVID-19 is continuously mutating and worldwide scientists are having a tough time creating an effective vaccine that can meet the standards of clinical trial.

We came to a conclusion then that the best way to deal with the COVID-19 for now would be to get the factors that affect the death rate of COVID-19. If we can find out the factors that can decrease the death rate of the coronavirus, it would be a meaningful analysis in the current global situation. By properly processing the data collected from various countries, it would be possible to make a model that would contribute to the decrease of death rate. Judging that this study would be meaningful if only with proper processing of data and modelling, we decided to do an analysis on the topic.

2. Project objectives

Based on the datasets on health_indicators, age, and death_rate of COVID-19, we wanted to conduct an analysis about the relationship between the factors and the death rate of COVID-19. Therefore we set our objective as identifying the factors that affect the death rate among the ones given in the raw data. Then, we decided to use them to build a relatively accurate model to forecast the death rate given the statistics by using decision tree.

3. Theoretical Analysis

a) one-way ANOVA and post-hoc analysis

i) ANOVA

ANOVA is used to find if certain factor is significant. ANOVA doesn't show the exact values, but it is used to decide only if it is significant or not. The null hypothesis is set when the means of all the groups are the same. On the other hand, alternative hypothesis is set when the means are varied.

Depending on the p-value and significance level alpha, we can decide whether the null hypothesis is rejected or not. When the p-value is smaller than alpha, the null hypothesis is rejected and if not, it is accepted. If the null hypothesis is rejected, we have to conduct post-hoc analysis and

investigate what is the exact difference in the means within the group. In R programming, ANOVA can be conducted by using str() of the lm()value or by using the aov() function. Depending on the library one picks, there are various ways to do ANOVA.

ii) Post-hoc analysis

Post-hoc analysis used for mean comparison come in different shapes such as FIsher's LSD and Tukey's HSD. In this analysis, we only used scheffe test. Scheffe Test is useful when the number of samples are different in each level(category). Moreover, when there are more than 2 groups, scheffe test is preferable among all the other tests. We divided the death rate into three levels, and due to the fact that the number of samples vary among the levels, we chose scheffe test.

iii) Pearson correlation coefficient

Correlation coefficient shows statistical relationship between variables. The values of correlation coefficient range between -1 and 1. If the value is zero, it would mean that there is no significant relationship between the values. The positive values imply positive correlation and the negative values imply negative correlation.

b) Data pre-processing

i) Min-max normalization

Min-max is one of the most common ways of normalizing the data. For all the features in the data, set the minimum value as 0 and maximum value as 1 and convert all the other values in between to the floating points between 0 and 1. If the value on the raw data is x, the converted value is (x- xmin)/(xmax - xmin).

This preserves the characteristic of each dataset and only changes the scale, which makes it an advantage when dealing with data with large size by preventing oversampling.

ii) Synthetic Minority Oversampling Technique(SMOTE) - Oversampling

If the data is not normally distributed, anomaly detection is likely to happen, which means that that overfitting would occur. There are numerous algorithms that regulate the number of samples within a class. Most famous methods are undersampling and oversampling. Since there were 5 samples within level3, we used oversampling and SMOTE algorithm[2].

SMOTE algorithm is one of the most popular oversampling methods using KNN algorithm. First, it selects one minority class sample. After finding 3 K-nearest neighbors, it generates synthetic minority instances. Since it doesn't just replicate variables, it can reduce the problem of over-estimating.

c) Decision Tree

i) Strategy for finding optimal tree

1) Complexity parameter

(a) The complexity parameter is proportional to the tree size.

(b) In the process of tree constructing, we conducted a 5-fold CV on a train set and measured its performance. The model for the optimal complexity parameter was chosen as the optimal model for our analysis.

ii) Evaluation

1) Accuracy(Performance)

(a) Construct a model on a train set and evaluate the model's performance using accuracy. This would be an indicator that takes sensitivity and specificity into account.

Accuracy = (TP + TN)/(TP + TN + FP + FN)

where: TP = True positive; FP = False positive; TN = True negative; FN = False negative

2) Gini index(Variable importance)

(a) Gini index is a measurement for the variance of k classes and is defined as the following. p_mk is the ratio of m'th area of k'th class[3]. If p_mk is close to 0 or 1, the Gini index's value is small, so we can conclude that if the gini index is small, the node is pure.

$$G = \sum_{k=1}^{K} p_{mk}(1 - p_{mk})$$

So when we increase the height of the tree, we can measure the decreasing of gini index and can indirectly decide the importance of a variable.

iii) Ensemble method

Decision tree's interpretability is superior to that of other classification models. But due to the fact that that variance is large, the accuracy of prediction decreases. We applied several techniques to make the variance smaller and thus increased the performance[4].

1) Bootstrap aggregating(Bagging)

(a) It builds k independent bootstrapped training sets from one single training set and train each model with each training set. Thus it predicts the data with the majority class. It could reduce the variance by averaging the uncorrelated errors in independent samples.

2) Randomforest

(a) This method is similar with bagging, only that the difference is the number of predictors. When generating each bagging tree, it randomly selects the root p out of full p predictors. Since bagging trees could include correlated trees due to the existence of some dominating predictors, this method efficiently reduces the variance by preventing similar trees.

3. Data sets

a) Raw data

i) Country Health Indicators[5]

This dataset is a collection of data about several indicators related to the health of a country. There are a total of 180 countries, and 70 variables about health. The variables can be divided into the following areas : Country Region(col 1), COVID-19 facts(2~8, 67~70), Death Causes statistics (9~18), Other fatalities (19,20), Food Sources (21~41), Health care systems (43~48), School closures (49, 63~66), BCG data(50, 51), CIA factbook statistics (People/Society facts) (52~62). The data used in this study is Death Causes statistics, Food sources, Health care systems, and factbook statistics.

Death Cause Statistics: This is the data showing the proportion of each disease in DALYs. DALYs is an indicator which represents the loss of life expectancy caused by each disease. Data includes DALYs of cardiovascular disease, cancer, etc.

Food Sources: This is data representing the country's food consumption in thousand tons. There are variables such as vegetable oils, alcoholic beverages, and starchy roots.

Health care systems: It contains data on the country's health care system. There are a number of beds per 10,000 people, the density of hospital(the number of hospitals per 100,000 people), the number of surgeons/gynecologists, and the number of doctors per million people.

Factbook Statistics: It is a fundamental data of a country. But the only variable we used in this study is the median age of each country.

ii) COVID-19 cases: worldwide[6]

This dataset is a collection of data about the number of COVID-19 cases and deaths worldwide. We used this dataset to calculate the death rate of COVID-19. Because the number of confirmed cases increases everyday, we fixed the date on May 31.

iii) Population ages 65 and above(%)[7]

The dataset is about the percentage of the elderly population in each country, over 65. We used the data of 2018, and tried to analyze the death rate of COVID-19 according to the proportion of elderly in the population.

b) Pre-processed data

Since the death rate is a dependent variable, we eliminated the samples with 0 or NA death rate. Then we processed the data by deleting the lower 30% of the death rate, deciding that those variables would be a noise.

To make a prediction model using decision tree, we went through the preprocess step written in the theoretical analysis. First step was min-max normalization. After that, we used 'inner join' to join the several datasets with country names, and joined disease, food, health care system and age data. Also, we rounded all the datas to second decimal place.

To group the countries according to the death rate, we made a new variable named "level", which indicates the level of death rate. Death rate 0~4% is level 1, 4~8% is level 2, and 8%~ is level 3. The reason why we introduced the level of death is because a model that exactly predicts the death rate with high performance is hard to build. And it is also meaningless, because important thing is the degree of death, not the exact death rate.

The raw data contained data of 180 countries, but after the processing of data, the remaining data was like the following.(level 1 : 24, level 2 : 12, level 3 : 5)

Since there were only 41 samples left, we made the sample number of each level(24 each) same through oversampling.

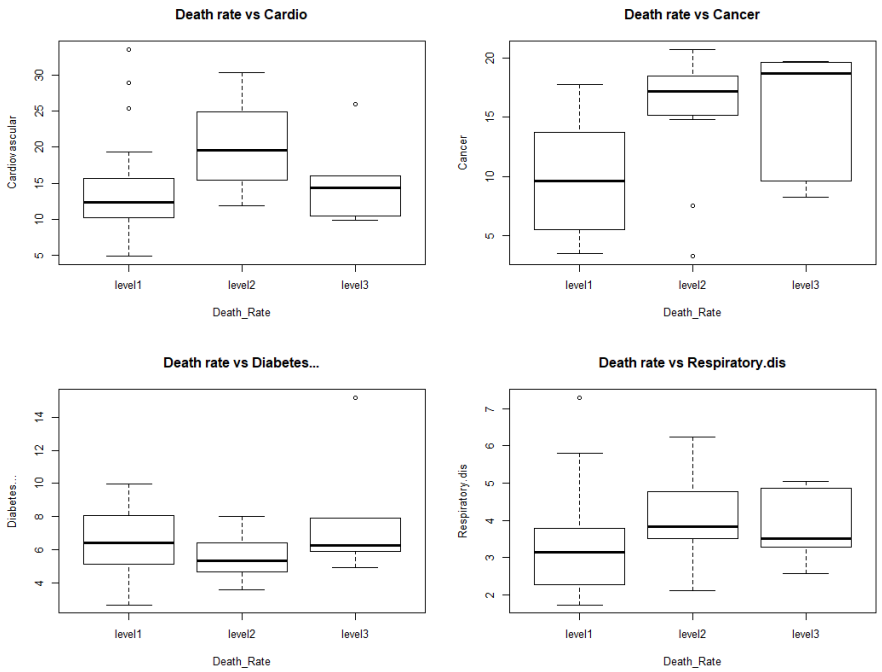
Final data set includes level (column 1), location (2), death rate(3), disease data (4~13), food source data (14~34), Health care systems (35~40), over65 percentage(41), and median age(42).

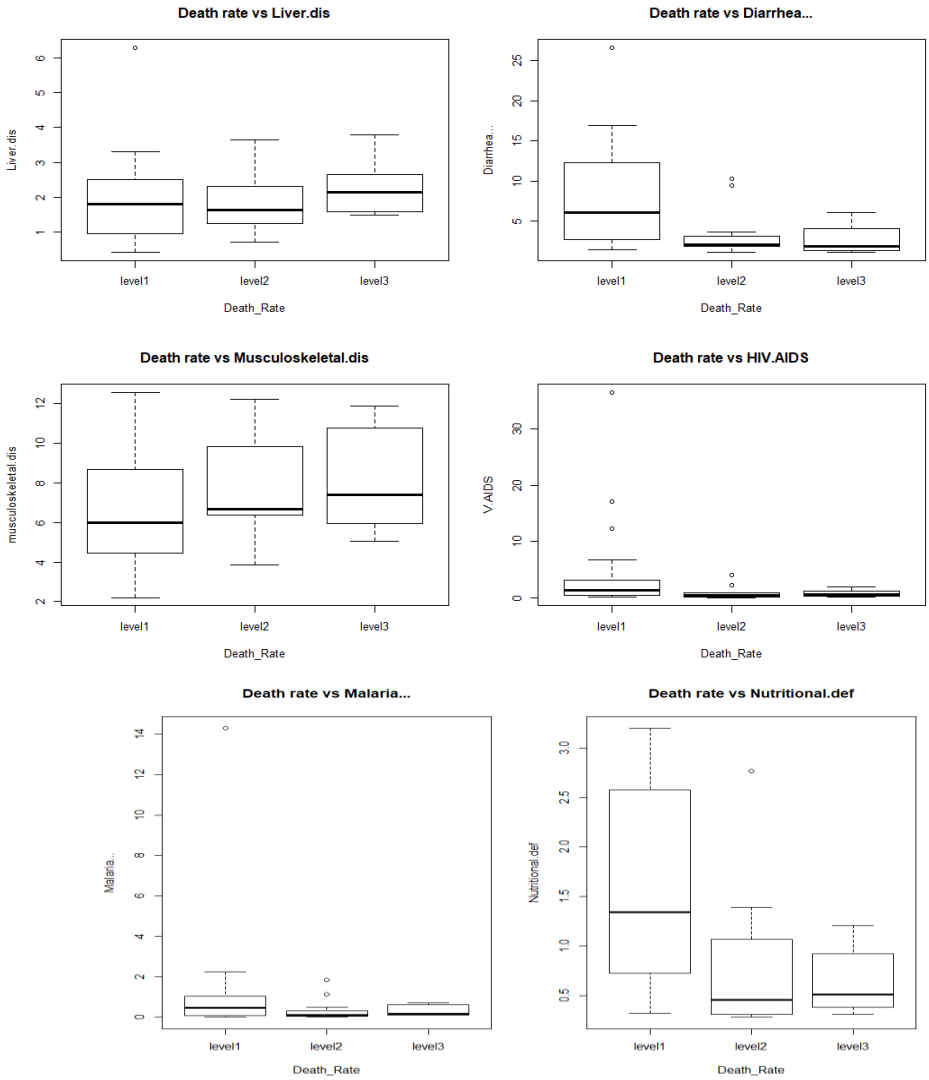
4. Problem definition & Analysis

a. Analyze the relationship between each of the factors(variables) in the 4 datasets(classes)

1) Death cause statistics

a) Boxplots (level vs. factors)





b) Result of ANOVA

-H0: average of the levels within each content is equal

-significance level: 0.05

Cancer: 0.00294

Above is the p-values of ANOVA test for each factors, which are less than 0.05. We can see that only 'Cancer' factor has passed the ANOVA test.

c) Result of post-hoc analysis

(Significance level: 0.05)

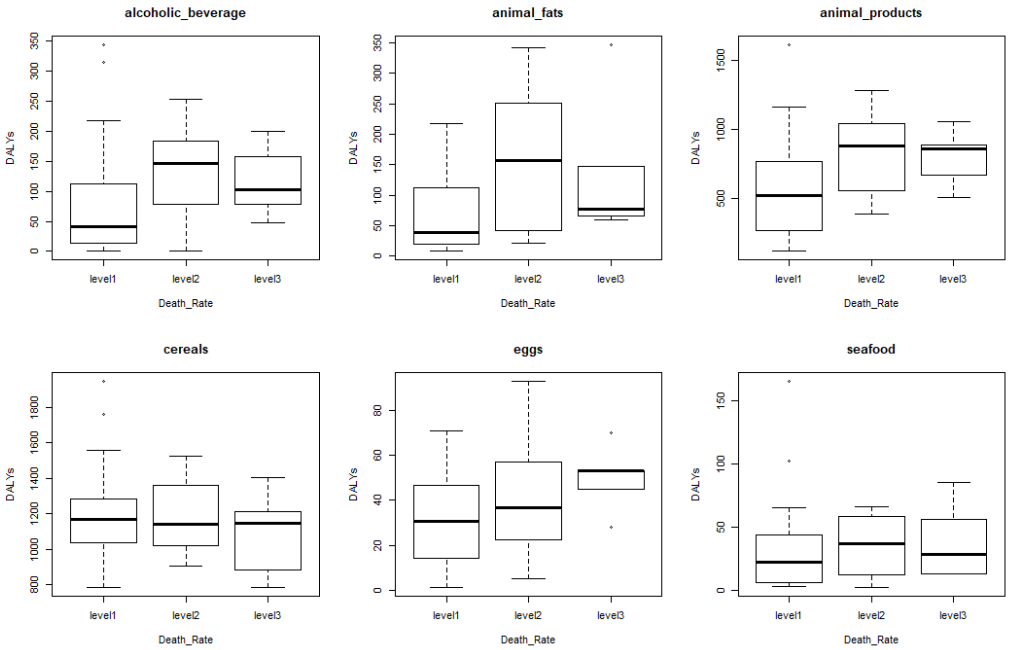
Cancer:

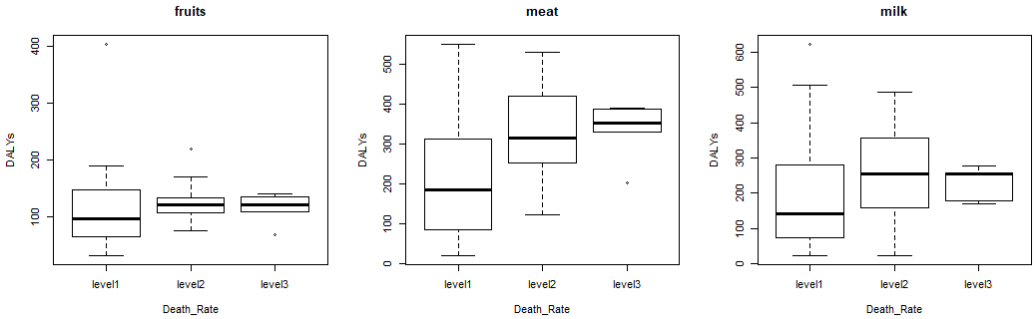
	y[, i]	groups
level2	15.45417	a
level3	15.16200	ab
level1	9.63625	b

Above is the post-hoc analysis result for 'cancer'. The result tells us that level 1 and level 2 are significantly different, but level 2 and level 3, level 1 and level 3 are not significantly different. This implies that cancer affects the COVID-19 death rate.

2) food sources

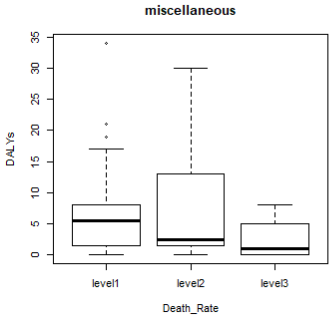
a) Boxplots(level vs. factors)

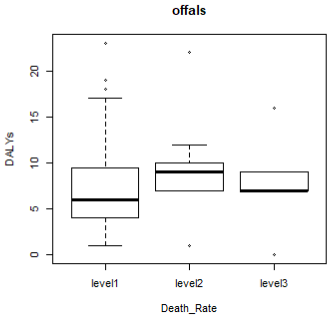


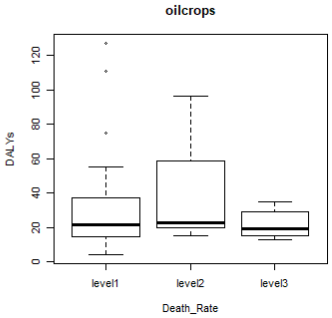

fruits
DALYs
Death_Rate
level1
level2
level3

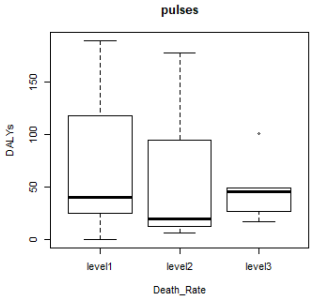
meat
DALYs
Death_Rate
level1
level2
level3

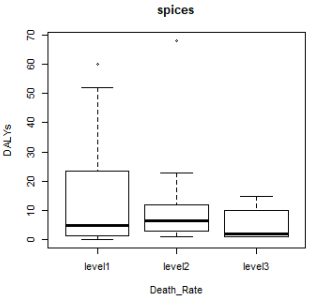
milk
DALYs
Death_Rate
level1
level2
level3

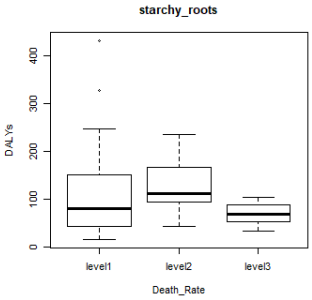

miscellaneous
DALYs
Death_Rate
level1
level2
level3

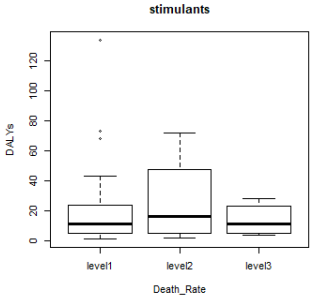

offals
DALYs
Death_Rate
level1
level2
level3

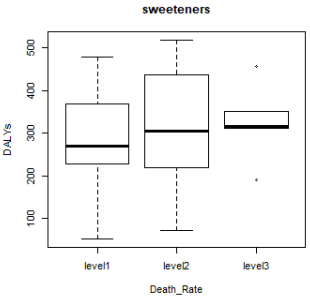

oilcrops
DALYs
Death_Rate
level1
level2
level3

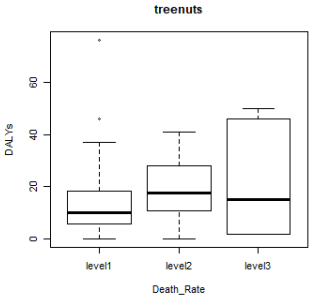

pulses
DALYs
Death_Rate
level1
level2
level3

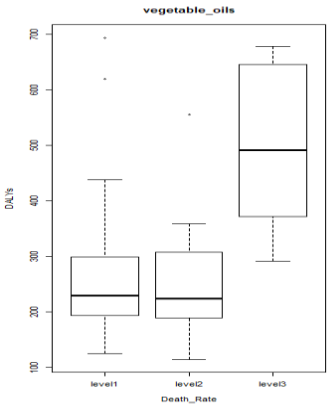

spices
DALYs
Death_Rate
level1
level2
level3

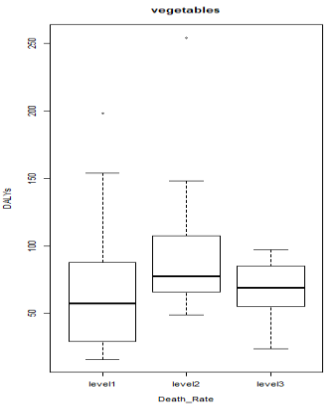

starchy_roots
DALYs
Death_Rate
level1
level2
level3

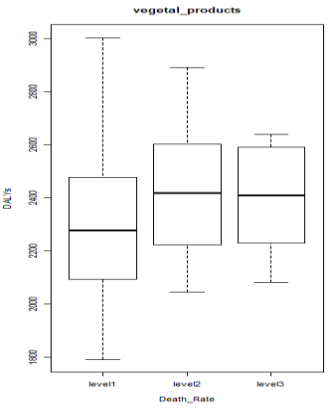

stimulants
DALYs
Death_Rate
level1
level2
level3


sweeteners
DALYs
Death_Rate
level1
level2
level3


treenuts
DALYs
Death_Rate
level1
level2
level3


vegetable_oils
DALYs
Death_Rate
level1
level2
level3


vegetables
DALYs
Death_Rate
level1
level2
level3


vegetal_products
DALYs
Death_Rate
level1
level2
level3

b) Result of ANOVA

-H0: average of the levels within each content is equal

-significance level: 0.05

animal_fats: 0.0161

vegetable_oils: 0.00522

Above is the p-values of variables which passed the ANOVA test. Only two variables, animal fats and vegetable oils, had passed the ANOVA test, so we can claim that the consumption of animal fat or vegetable oil is related to the death rate. Further analysis is done with post-hoc analysis.

c) Result of post-hoc analysis

(Significance level: 0.05)

animal_fats:

	y[, i]	groups
level2	154.08333	a
level3	139.40000	ab
level1	67.58333	b

vegetable_oils:

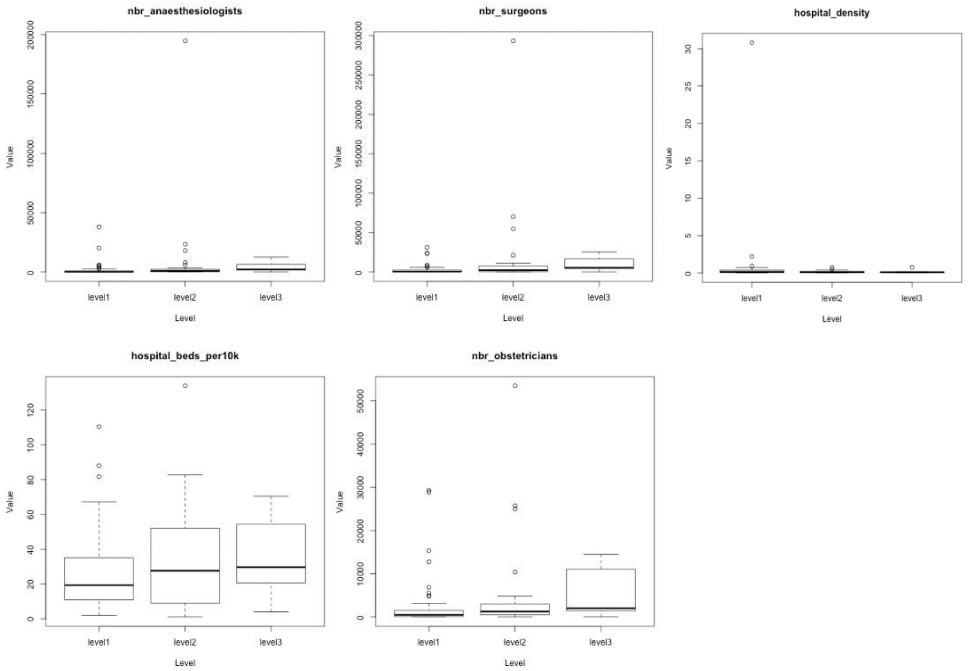
	y[, i]	groups
level3	496.2000	a
level1	276.8750	b
level2	257.4167	b

First, as a result of post-hoc analysis for animal fats, there is a significant difference between level1 and level2, and there is a slight difference between level1 and level3, level2 and level 3. Second, as a result of post-hoc analysis for vegetable oils, we can group the levels into two, level 3, and level 1,2. There are significant differences between two groups.

We can conclude that both animal fats and vegetable oils perform a role in the death of COVID-19.

3) Health care systems

a) Boxplot (level vs. factors)



b) Result of ANOVA

-H0: average of the levels within each content is equal

-Significance level: 0.05

nbr_surgeons: 0.271

nbr_obstetricians: 0.173

nbr_anaesthesiologists: 0.292

none of the factors rejected null hypothesis

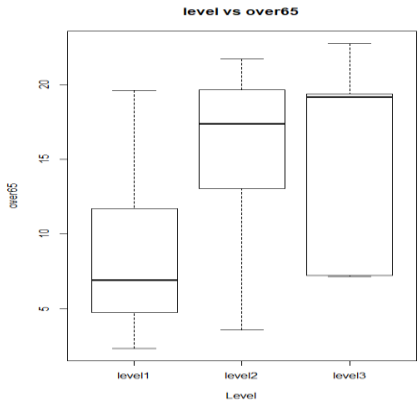
c) Result of post-hoc analysis

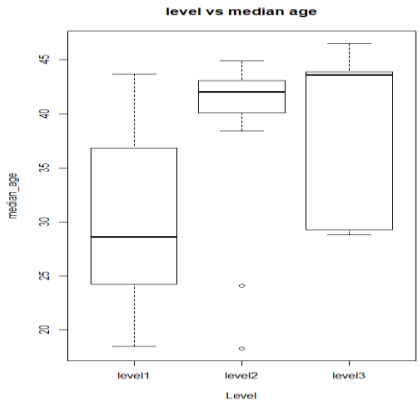
(Significance level: 0.05)

unable to conduct

4) factbook statistics(median age), population ages 65 and above

a) Boxplot (level vs. factors)





b) Result of ANOVA

-H0: average of the levels within each content is equal

-Significance level: 0.05

over65 : 0.000822

median age : 0.00409

Both factors, over65 and median age rejected the null hypothesis. They both passed the ANOVA test since the p-value was smaller than 0.05.

c) Result of post-hoc analysis

(Significance level: 0.05)

over65:

	over65	groups
level2	15.435000	a
level3	15.134000	a
level1	8.240833	b

median age:

	median.age	groups
level2	38.88333	a
level3	38.42000	ab
level1	29.91667	b

First, as a result of post-hoc analysis for over65, we could group the levels into two, level 1 and level 2,3. There are significant differences between two groups.

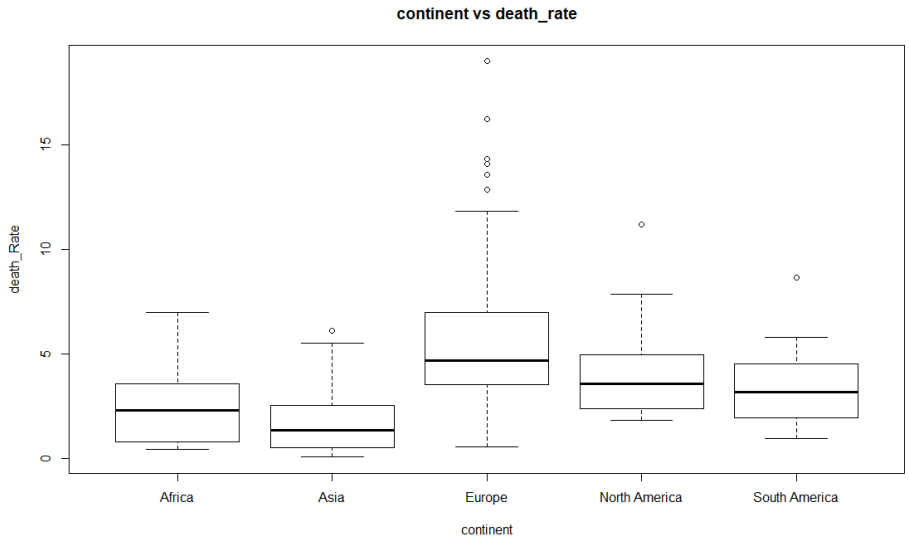
Second, as a result of post-hoc analysis for vegetable oils, we could see that the levels can be grouped into three. Level 1 and

level 2 had significant differences. Level 2 and level3, level 1 and level 3 also had certain level of differences between them.

The conclusion of above result is that both the percentage of population over65, and median age is related to COVID-19 death rates.

5) death rate depending on the continent (continent vs. death rate)

a) boxplot



b) result of ANOVA

-H0: average of the levels within each content is equal

-Significance level: 0.05

p-value : 3.21e-06

The null hypothesis is rejected.

c) Result of post-hoc analysis

(Significance level: 0.05)

	death_rate	groups
Europe	6.403505	a
North America	4.378223	ab
South America	3.611183	ab
Africa	2.573784	b
Asia	1.889986	b

The death rate of Africa and Asia didn't have significant difference between them, but Europe was significantly higher. North America and South America were to a limit similar to both sides.

6) Further analysis for food sources (animal fats & vegetable oils)

Through the above process, we identified specific variables which are related to COVID-19 death level: Animal fats, Vegetable oils, Cancer, Median age, over 65 percentage. However, we thought that the food values are somewhat correlated with other kinds of variables. Animal fats and vegetable oils are both a fat product, which tends to be consumed more in a wealthy country. And because wealthy countries have higher level of medical welfare, their median age and over 65 percentage will be higher than other countries. Also, in developing countries, other kinds of disease usually have a greater impact on death than cancer. Accordingly, further analysis of relation between food sources(animal fats, vegetable oils) and remaining factors(median age, over 65 percentage, and cancer DALYs) were conducted. After leveling the factor, ANOVA test and post-hoc analysis was done among levels.

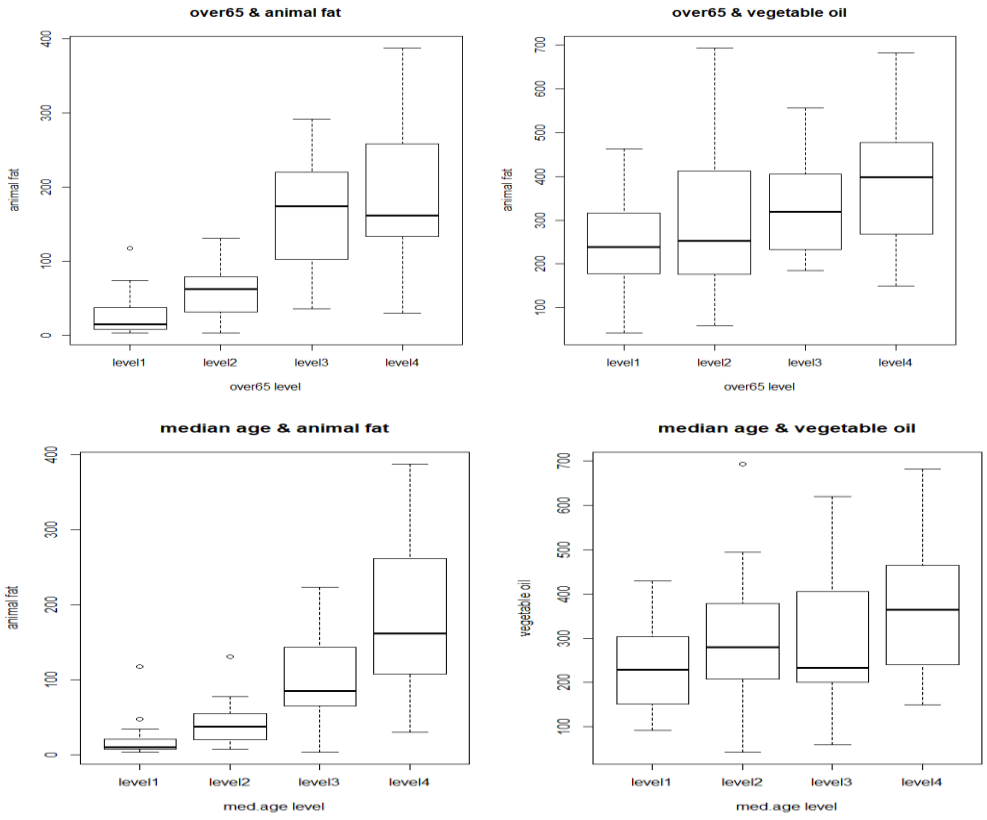
Levels of each factor are set as below.

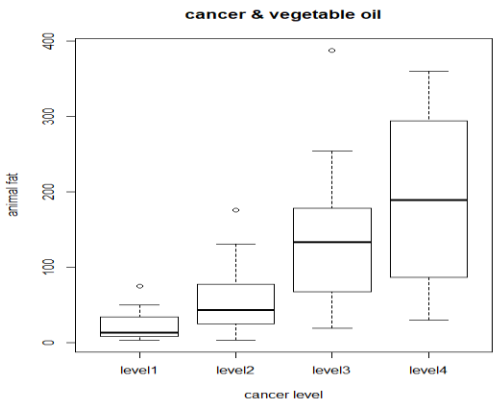
Over65 => level 1: ~6, level 2 : 6 ~ 12, level 3 : 12~18, level 4: 18~

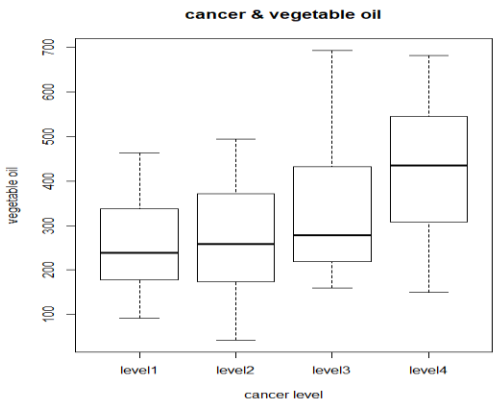
median age => level 1:~24, level 2 : 24 ~ 32, level 3 : 32~40, level 4: 40~

Cancer =>level 1: ~6, level 2 : 6 ~ 12, level 3 : 12~18, level 4: 18~

a) Boxplots(level vs. factors)







b) ANOVA

-H0: average animal_fat/vegetable oils consumption is equal for each level

-significance level: 0.05

Variables on the front are leveled variables.

over 65 & animal fat : 2.74e-14
over 65 & vegetable oils : 0.0039
median age & animal fat : 6.3e-12
median age & vegetable oils : 0.0189
cancer & animal fat : 4.26e-11
cancer & vegetable oils : 0.00118

All of the factors rejected the null hypothesis, and came to a conclusion that There are relation between every two factors. All results above are passed to the post-hoc analysis stage.

c) Post-hoc analysis

(Significance level: 0.05)

over 65 & animal fat :

	animal_fats	groups
level4	191.73913	a
level3	163.72727	a
level2	57.66667	b
level1	27.27586	b

over 65 & vegetable oils :

	vegetable_oils	groups
level4	395.1739	a
level3	340.4545	ab
level2	302.5000	ab
level1	247.9310	b

median age & animal fat :

median age & vegetable oils :

	vegetable_oils	groups
level4	379.0000	a
level2	296.2963	ab
level3	293.1250	ab

	animal_fats	groups
level4	183.13793	a
level3	101.31250	b
level2	41.81481	bc
level1	21.66667	c

cancer & animal fat :

	animal_fats	groups
level4	198.00000	a
level3	137.12500	a
level2	56.53846	b
level1	21.76190	b

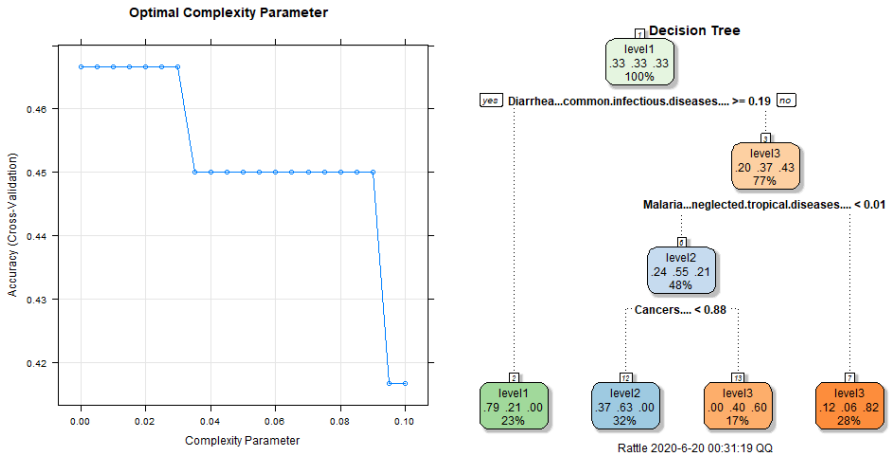
cancer & vegetable oils :

	vegetable_oils	groups
level4	427.6250	a
level3	335.9583	ab
level2	266.2692	b
level1	259.8095	b

In the above result, means with the same group letter are not significantly different. All the result indicates that the levels can be grouped into multiple groups according to the factors. We can draw a conclusion that there is relation between food source(animal fats, vegetable oils consumption) and median age, over 65 population, and cancer.

b. Construct & Evaluate the prediction model through Decision Tree

1. Prediction model for disease features



Optimal complexity parameter is 0.03. And optimal decision tree is above. But its accuracy is only 0.4167. Test results are below.

Confusion Matrix and Statistics

Reference

Prediction	*level1*	*level2*	*level3*
level1	*2*	*0*	*1*
level2	*2*	*0*	*0*
level3	*0*	*4*	*3*

Overall Statistics

Accuracy : 0.4167

95% CI : (0.1517, 0.7233)

No Information Rate : 0.3333

P-Value [Acc > NIR] : 0.3685

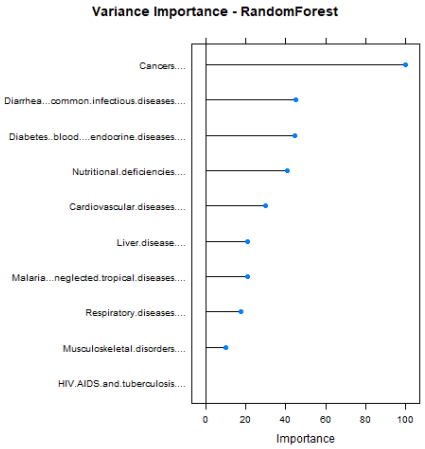
Kappa : 0.125

Mcnemar's Test P-Value : 0.0719

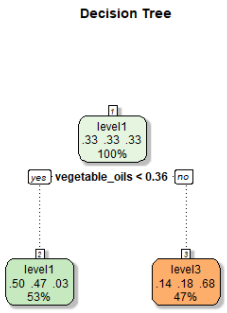
As we apply ensemble methods, the model's performance becomes higher. We could conclude that the last model(randomforest) is optimal.

Decision tree	Bagging	RandomForest
0.4167	0.8333	1

For that model, the plot of variable importance is below. Top 3 predictors are Cancer, Diarrhea, and diabetes. This is a consistent result since cancer is the only feature that passes the ANOVA & Scheffe test.



2. Prediction model for food features

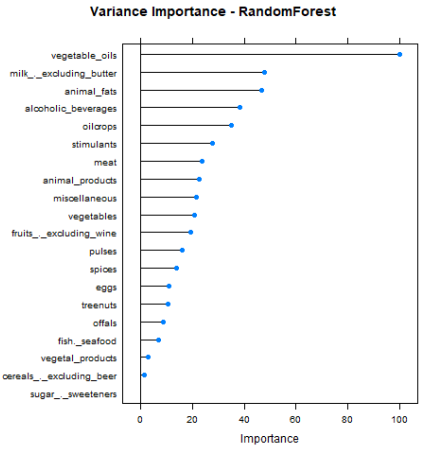


Optimal complexity parameter is 0.03. And optimal decision tree is above. But its accuracy is only 0.5833. Test results are below.

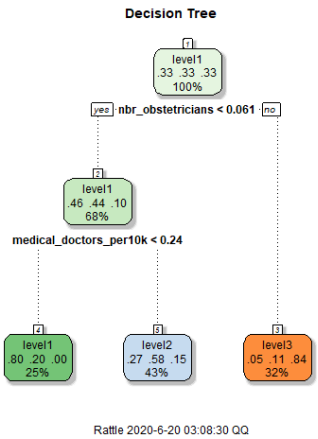
As we apply ensemble methods, the model's performance becomes higher. We could conclude that the last model(randomforest) is optimal.

Decision tree	Bagging	RandomForest
0.5833	0.5833	1

For that model, the plot of variance importance is below. Top 3 predictors are vegetable_oils, milk_excluding_butter, and animal_fats. This is a consistent result since vegetable_oils, and animal_fats are the only feature that pass the ANOVA & Scheffe test.



3. Prediction model for health care systemfeatures

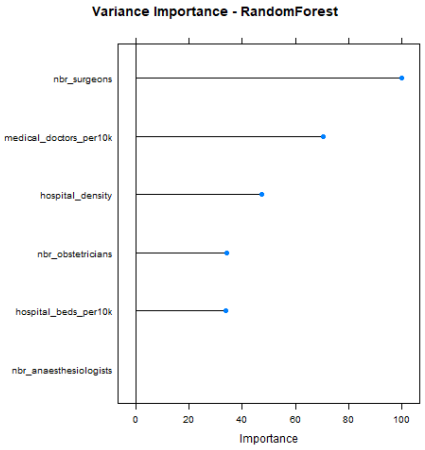


Optimal complexity parameter is 0.155. And optimal decision tree is above. But its accuracy is only 0.75. Test results are below.

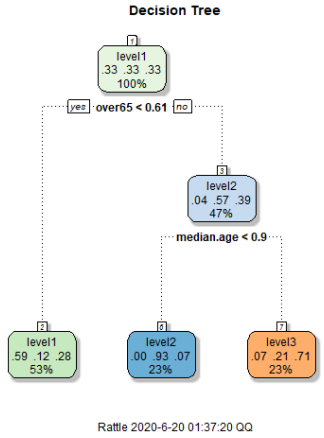
As we apply ensemble methods, the model's performance becomes higher. We could conclude that the last model(randomforest) is optimal.

Decision tree	Bagging	RandomForest
0.75	0.75	1

For that model, the plot of variance importance is below. Top 3 predictors are number of surgeons, medical doctors' density, and hospital density. But since no features passed the ANOVA & Scheffe test, we could say that this model isn't reliable. It is over-trained.



4. Prediction model for age features

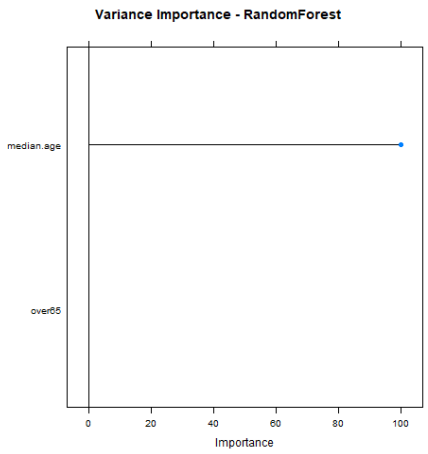


Optimal complexity parameter is 0.12. And optimal decision tree is above. But its accuracy is only 0.533. Test results are below.

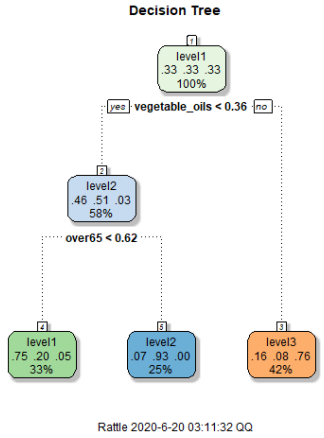
As we apply ensemble methods, the model's performance becomes higher. We could conclude that the last model(randomforest) is optimal.

Decision tree	Bagging	RandomForest
0.5833	0.75	1

For that model, the plot of variance importance is below. According to our model, median age made more attributes to death rate increment than over65.



5. Prediction model for all features

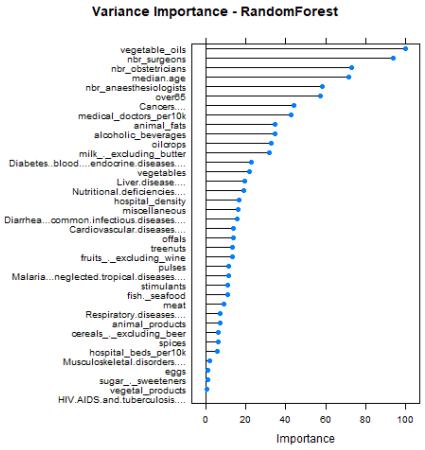


Optimal complexity parameter is 0.1. And optimal decision tree is above. But its accuracy is only 0.6667. Test results are below.

As we apply ensemble methods, the model's performance becomes higher. We could conclude that the last model(randomforest) is optimal.

Decision tree	Bagging	RandomForest
0.6667	0.8333	1

For that model, the plot of variable importance is below. Contrary to the inference step, there exist 4 health-care related features among top 6 predictors. At the inference step, any health care related features passed the ANOVA test.



5. Conclusion

The purpose of this study is to identify the variables which are highly correlated with COVID-19 death rate, and to create a predictive model with good performance reflecting them. The variable selection and model creation process could help with

research on COVID-19 and the factors critical to death can be specified. Also we will be able to identify countries which are vulnerable to COVID-19.

The first step was to infer the relationship. By specifying not only the data that is generally related to death rate such as age, but also other variables such as food, disease, hospital, we have identified relationships with features that we do not generally know. For a total of 39 variables, we conducted ANOVA test, and then used the scheffe test for the post-hoc analysis. The variables which passed both tests and concluded to be relevant to death rate were as follows: Cancer(DALYs), animal fat, vegetable oil, median age and over 65 population. Furthermore, two food variables, animal fat and vegetable oils, were considered to be related with other three variables because they are related with the wealth of the country. We conducted the ANOVA and post-hoc analysis for them, and all the variables had passed the test and concluded that they are relevant. There were many features predicted to be correlated with COVID-19 death rate before analysis. But as a result, except for the age variable which is easily predictable, only the importance of cancer was additionally known.

Next step is to make a prediction model. We chose the decision tree as a method because our goal was to build a model which is both highly predictable and highly interpretable. Normally, when the complexity of the prediction model decreases the bias increases, and when the model is overtrained and complexity increases the variance increases. This is called bias-variance tradeoff, and it is crucial to set an appropriate complexity which can minimize the error. In this study, we used 5-fold cross validation to find the optimal model, and made a lot of effort to control the tree depth(cp) during decision tree creation, and ratio of selected predictor during randomForest creation. Additionally, we build a high-predictable model with accuracy 1 for the test set, using ensemble methods such as bagging and randomForest. However, the result of the prediction model was inconsistent with the result of inference done before. The only variable in the health care category which was related to death rate was cancer in the inference step, but four kinds of diseases were in the top 6 important variables in the optimal decision tree we've constructed.

We thought that this happens because the over-estimation had occurred during the sampling process. Since one country is equal to one sample, and there were almost 40 variables, the number of samples has reduced significantly after excluding all the NA terms in each variable. The number of samples remaining was 49, and the class imbalance problem was also severe(level 1: 24, level 3: 5). We proceeded oversampling using the SMOTE algorithm because the number of samples were small, so it is assumed that the prediction model is overtrained.

Various efforts have been made to increase the number of samples and to solve these problems. We tried to control the number and range of levels, but still there was no progress. We also wanted to try to increase the data by collecting data of each state, instead of country, or by collecting data several times with a time interval for a single country. However, we judged that these methods were not effective, because the death rate is time independent.

Through this study, we concluded that cancer, median age, and over 65 population is related to the COVID-19 death rate. Influence of two food sources, animal fats and vegetable oils are also included in above three variables. By observing these variables, we noticed that all these variables are related to countries' wealth. Especially in Europe, the consumption level of animal fat and vegetable oils are high, DALYs of cancer is higher than other diseases since other diseases are easy to treat. Also the over 65 population and median age of countries in Europe are high. This comes to the conclusion that the factors we've selected are factors targeting Europe, which is suffering heavily from COVID-19. We wanted to infer the factors related to COVID-19 death rate, but actually the factors we specified were only the characteristics of countries which are suffering from COVID-19.

While thinking about the cause of this situation, we concluded that the problematic part was to underestimate the relation of COVID-19 confirmed cases and death cases. We first thought that when some people got COVID-19, the variables that determine whether a person survives are related to health data, and have a small relation with the confirmed rate. However, while considering the total percentage of death, it should be related somehow to the confirmed rate. So geographical features, number of cross-country movements and a massive number of variables should be included while calculating the COVID-19 death rate. The data we used had about 40 variables, and it was not sufficient enough.

Moreover, during this study, we noticed that sampling and pre-processing steps are crucial. Depending on the type of data, we should choose appropriate methods for processing, and analysis. Also the lack of sample numbers was the biggest problem in our study, so we learned that preparing a sufficient number of samples is significant during data processing. The conclusion we made was insufficient, but we could test various methods we've learned using R such as anova, decision tree, and randomforest, which was a meaningful experience.

6. R&R

김경수	Theoretical background investigation, such as the Decision tree-based method, parametric tests, etc. Conducting the analysis related to the Prediction model.
이채영	Investigate the data set and the related background theory. Investigate required R code package and proceed ANOVA & post hoc analysis.
전민식	Investigate R code base. Feedback about data, and codes. Proceed ANOVA & post hoc analysis, and further analysis.

7. References

[1] Lee, Sangseok, and Dong Kyu Lee. "What is the proper way to apply the multiple comparison test?." *Korean journal of anesthesiology* 71.5 (2018): 353.

[2] Agrawal, Astha, Herna L. Viktor, and Eric Paquet. "SCUT: Multi-class imbalanced data classification using SMOTE and cluster-based undersampling." 2015 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K). Vol. 1. IEEE, 2015.

[3] https://en.wikipedia.org/wiki/Gini_coefficient

[4] Gareth, James. An introduction to statistical learning: with applications in R. Springer Verlag, 2010.

[5] https://www.kaggle.com/nxpnsv/country-health-indicators

[6] https://ourworldindata.org/coronavirus

[7] https://data.worldbank.org/indicator/SP.POP.65UP.TO.ZS?end=2018&start=2018&view=map