COVID-19 Spread prediction
Based on Food Categories
using Data Science

Y.Mahima
Informatics Institute of Technology
University of Westminster
Colombo,Sri Lanka
yasasymahima@gmail.com

T.N.D.S. Ginige
Universal College Lanka
Monash University
Colombo,Sri Lanka
thepul@ucl.lk

ABSTRACT – Novel corona (COVID-19) disease is an infectious disease caused by a newly discovered coronavirus. Covid-19 was started from Wuhan China and present it is spread all over the world. Presently the USA is affected by the Covid-19 virus hardly. Present there are more than 4.0 million confirmed cases and more than 200 thousand happens because of the Covid-19 virus. [1]. However, there is no medicine found by the researchers yet for the virus. According to the WHO (World Health Organization), the COVID-19 virus is spreading by several methods. And for protection from the COVID-19 virus, their suggestion is community distance. Moreover, doctors and other medical instructors advise concentrating more on doing sanitizations often. Presently there are lots of researches doing for identifying reasons, make a medicine in this area. When it concentrates on the spread of the COVID-19 virus it cloud be able to identify that there is a difference between Asian and European countries. As an example from Asian countries, only China and Russia are affected by the COVID-19 virus majorly. However European countries like Italy, Spain has the highest death rates more than china. Mainly south Asian countries have a low confirm rate of the COVID-19 virus. [2] For this the main reason what here in this research identified is food used by each region. We know that people in European countries eat fast foods and they have used alcohol more than the people in Asian countries. [3]

Therefore this research was done to identify a pattern between the spread of the COVID-19 virus and the Food categories used by the people. So here in this research using data science. Here it does some data modeling part to visualize a pattern between the spread of the COVID-19 virus and food categories. Thereafter using a machine learning model here it implement a prediction model. Through this research, it could be able to understand that there is a clear pattern between the spread of the virus and the food factors. Moreover, this research will be a novel approach to evaluate the spread of the COVID-19 based on food categories.

I. INTRODUCTION

World Health Organization has declared the Novel Corona Virus disease in 2019 a pandemic. This was started on December 31 by a Cluster of patients found in the city of Wuhan and Hubei province in China. The Impact of the Coronavirus is bigger than the previous outbreaks like SARS and MERS. As of 15 May 2020, there are over 4 million cases have been identified and over 200 thousand deaths were reported globally in 188 countries. Moreover, nearly 1.5 million were recovered. By the below graph it could be able to clearly understand the spread of the virus from the beginning. [4] [5] [6]
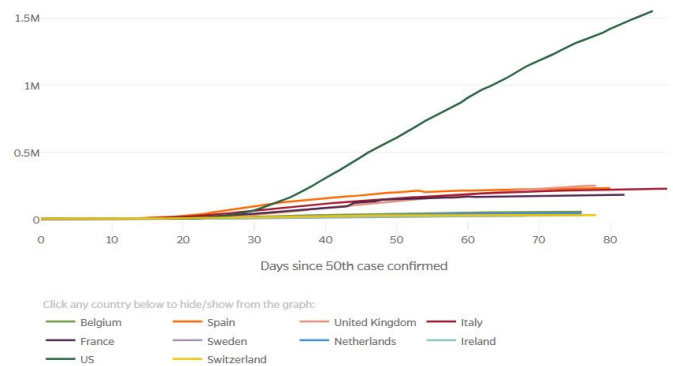

Figure 1: Distribution of the Virus

When it concentrates on the spread of the death rates of the COVID-19 virus there can understand that the spread of the virus is faster in European countries and the USA when comparing to the Asian countries. Coronavirus was started from china. But there were only nearly 82 thousand confirmed cases and 4.5 thousand deaths reported in China. [7] But in Italy, Spain Britain the confirmed cases and deaths are several times greater than China. From the below chart it can clearly understand that most of the European and Western countries affected by the virus in a very fast manner. [8]
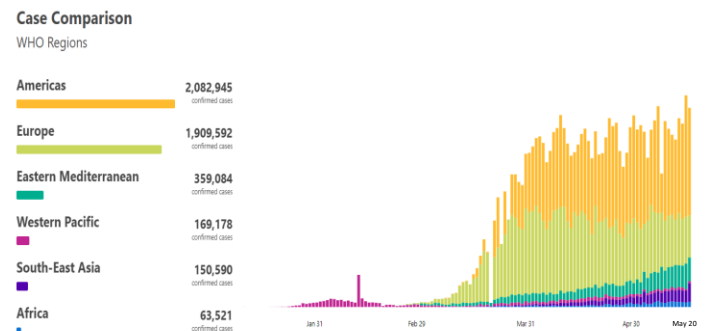

Figure 2: Region wise Distribution of the Virus

For the difference between the increment of the COVID-19 cases in western and Asian countries, there are several reasons identified by the previous researchers. The main thing they identified in lack of social distance in the daily activates of the people. The next thing is the weather. Most of Europe and western countries have cold weather. But Asian countries have hot weather. Hot weather is able to kill the COVID-19 virus in the environment.

However, from this research it will concentrate on the Food Categories which causes for the spread of the COVID-19 virus. When it compare the foods taken by the Asian and Western people there is a difference. Western people eat fast foods and foods include meats and using alcohol more when comparing to the Asian people. The below graph shows how meat product usage grows in each region according to the year. [3]
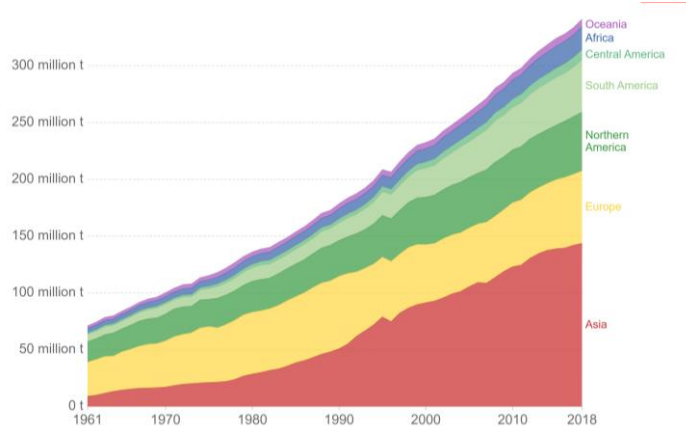

Figure 3: Meat Product Usage in Each Region

From those factors, there can assume that there is a probability that food categories also cause for the spread of the virus. Therefore from this research, it is able to recognize a pattern between food categories and the spread of the coronavirus. When it considers the food categories several food categories can be identified. According to the LCHF food categories, there are six main food categories. They are vegetables, plant fat, low sugar fruit meat, and seafood items, animal fat, and beverages. [9]

As mentioned about to predict how the COVID-19 virus spread with the Food categories here this research used a data science component. First to identify how the confirmed cases change according to the food categories in each country here this research used data modeling concepts. Then to identify a pattern and predict the spread of the virus according to food categories usage here this research used some algorithmics approaches and trained a model.

So this is a brief summarization of this research. Here this research concentrate on a less concentrated area on the COVID-19 virus previously. Finally, in the implementation process of this research here it got a 69.7% accurate data science model to predict COVID-19 spread prediction relevant to the usage of food categories. This research will take a novel approach to analyze the spread of the COVID-19 virus and it helps to self analyze and save from the COVID-19 outbreak. In the next chapter, it will discuss several competitive pieces of research done in this area.

II. LITERATURE REVIEW

Novel coronavirus became a world problem now. Therefore there are lots of research done in this area. There are researches to predict if a person has a virus or not, methods that virus spread, and predict the spread of the virus. As mentioned in the introduction chapter here it analyzes how the COVID-19 virus spread according to the usage of food categories. Moreover, to analyze here it used a data science approach. When it concentrates on previous researches done in areas like nutrition, foods related to the spread of the COVID-19 Virus, and reasons for the spread of the virus several types of research can find.

Several pieces of past researches were helpful to identify the food categories. Research by several researchers to analyze the current food production meets the global nutritional needs, they categorized foods and nutrition into six main categories. They are Sugar, Milk and Milk products, Proteins, Oils and Fats, Vegetable and Fruits, Grains. Below bar chart shows the categories which they used in their research. [10]
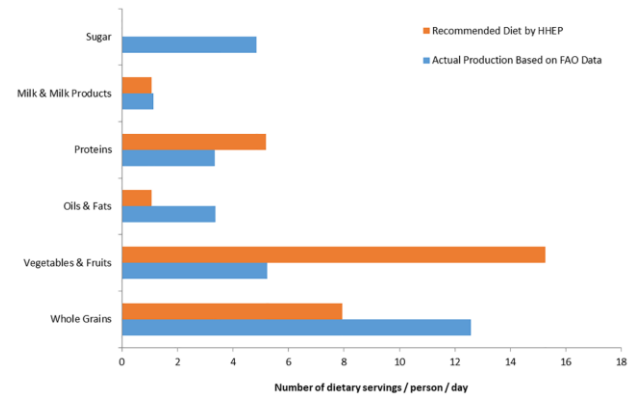

Figure 4: Food Categories

A Journal article by J.Poore and T.Nemcek named Reducing food's environmental impacts through producers and consumers categorized they categorized foods into nine main categories. They are Protein-rich products, Milks. , Starch-rich products, Oils, Vegetables, Fruits.,Sugars, Alcoholic, and Stimulants. [11].Therefore these are some recent researches done in categorizing food facts. From this research, it can get a clear idea about how the food categorization for the authors' research needs to be done.

There need to analyze how these food categories cause the immune system of the human body. A research by Sarmad Ghazi Al–Shawi gave a brief explanation for this. Here according to his study, it clearly described the importance of a healthy diet that contains all the nutrition for the immune system. Moreover here it described the impact of taking Alcholol for the system and the role of vegetable dairy products for the process of the immune system to make secure the body from the diseases. [12]. Research named "Nutritional Modulation of Immune Function: Analysis of Evidence, Mechanisms, and Clinical Relevance" described the correlations of each vitamin and other nutritions like fats, proteins with the process of the immune system. Here in this research it well defines the impact of the vitamins and irons which get from the vegetables, fats, and proteins. According to this research, vitamin C and E are helpful to the Immune process. [13].

In addition according to a blog Article by Physicians Committee for Responsible Medicine eating a low-fat, plant-based diet may help give the immune system a boost. Moreover, they mentioned Vegetarians have been shown to have more effective white blood cells when compared to nonvegetarians, due to a high intake of vitamins and low intake of fat. [14] .Moreover, a web article by Charles Patrick Davis who is a Ph.D. owner suggests 16 foods to boost the immune system. Here he suggest vegetarian food items more and some low-fat food items also. [15].So by the above past works, it can understand how each food category response to the process of the human immune system.

Charis M. Galanakis did a research named "The Food Systems in the Era of the Coronavirus Pandemic Crisis". Here in this research, he had concentrate on the bioactive ingredients of foods and herbs for the support of the human immune system against infections before discussing the possibility of COVID-19 transmission. Here this research had clearly discussed how the bioactive ingredients in the food support for the Human Immune system to protect from the coronavirus and further it had discussed sustainable food systems for the virus. Other than these topics this research focused on food security within the population lockdown which is not relevant to this research.
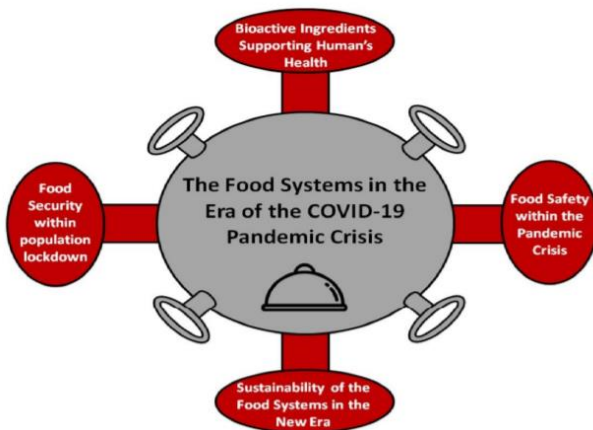

Figure 5: Food Systems in Covid-19 Era

Here in this research, it discussed how nutrients like Vitamin C, B-Carotine helps to the Immune system of the Body and they are helpful to secure from the COVID-19 virus. Moreover, it discussed how the herbs and Chinese medicines support for Immune system by using pieces of evidence like usage of those herbs in other outbreaks like SARS. [14]. The difference between this research and when comparing to our approach here it does not predict the spread of the COVID-19 virus with the food categories. Moreover here in this previously mentioned research, it is not a data science approach.

An article by the world health organization also described the relation between the human Immune system and the COVID-19 virus. According to that, the human Immune System makes antibodies that specifically bind to the virus. And it also makes T-cells that recognize and eliminate other cells infected with the virus. Moreover here it describes the process of the Human Immune system when a disease like the COVID-19 virus by taking evidence like SARS-CoV-2 inflection.

From these past research, there can understand that the human immune system does great work to secure the body from the COVID-19 virus. A study by Mitch Leslie proved that the human immune systems T cells react against the COVID-19 virus in 79% times. [15]. Based on that another study by Keck University of medicine and the University of Southern California suggests giving the proper immunosuppressant drug for the COVID-19 patients. Moreover here they mentioned a strong immune system is able to faster elimination of the virus and the infected cells. [16]. Therefore these past studies show the importance of this strong Immune system.

Another research done related to foods and the Covid-19 virus by two researchers named Farah Naja, Rena Hamadeh to analyze the food security in the world because of the COVID-19 virus. Here they were concentrated on how food security can be improved by nationally and globally. [15]. There is another research done by HLPE (High-Level Panel of Experts on Food Security and nutrition) Italy which is almost similar to the previous one. Here in this research, it will also concentre on Food security in Italy and the increment of the food demand. [16] However both of these researches do not concentrate on how food categories, food usage causes the spread of the COVID-19 Virus. But the area here in this research concentred by the author is how the distribution of the COVID-19 virus change by the usage of the food categories. So these are the researches done related to foods and the COVID-19 virus.

Several pieces of research done to predict the distribution of the COVID-19 virus and causes for the virus can be found. Several researchers have researched to predict the spread of the COVID-19 virus using machine learning named "Preparation analysis and prediction of the COVID-19". Here they were used time-series analysis to predict the spread of the COVID-19 virus. Moreover here they got a good accuracy rate for the model. [17].However, the difference between this and the proposed research by the author is here they used time series analysis but in the authors' research, it will analyze the food categories which cause the COVID-19 and predict the spread of the virus using that. Rather than predicting the spread here, this research will concentrate to identify how the food usage of the people causes the spread of the virus.

A research which used to analyze and predict COVID-19 distribution in African countries can be found. Here they used the SIER and MH-algorithm based model for the prediction. Moreover here in this research Epidemic controlling measures including risk level classification were proposed. When comparing to the authors' research here this research also analyzes the past data and does a time series analysis for the prediction. [18] . There is another research done by an Indian researcher Arti .M.K to modeling and the prediction of the COVID-19 spread in India. Here in this research, it used a Tree-based approach to predict the spread of the virus. Here in this research, it had also analyzed how the lock-down and isolation techniques cause to minimize the spread of the virus. [19].

So these are some related researches done to predict the spread of the COVID-19 virus and how food categories related to the spread of the virus. From these previous research approaches it

can understand that there are very few researches done to analyze how the people's food usage causes the Covid-19 virus and what types of foods which support the humans Immune system to secure from the COVID-19 virus.

Through these past researches, there can be identified that there can be a relationship between food categories, immune systems, and the spread of the COVID-19 virus. If a man has a strong immune system he can secure from the COVID-19 virus from the COVID-19 antibodies generated by the human Immune system. Moreover, those past studies clearly describe the impact of the food categories for the human Immune system. And present researchers goal for the COVID-19 also discovers a proper immunosuppressant drug. Therefore the concept here the authors highlighted is cant estimate the spread of the COVID-19 virus by the food categories usage by the people because according to these past researches mentioned foods are the main reason for a strong immune system. Therefore this research takes a novel approach to analyze and predict how the spread of the COVID-19 virus change according to the food categories. In the next chapter, it will clearly describe the proposed method and materials of the research.

III. METHODS AND MATERIALS

Here in this research, it proposed a system that analyzes the spread of the COVID-19 virus based on the food categories used by the people and predicts the spread of the virus by that. For the implementation as the main technique here, it is going to use a data science. Since there is an analysis section the implementation process of the research can divide into two main parts. They are
1. Data Modeling Process
2. Covid-19 Spread Prediction Model

Therefore the implementation process will discuss according to the above sections briefly.

For both data modeling and for the prediction here it used a country-wise data set which includes a COVID-19 virus spread and the use percentage of each food factor according to the population. BY using this here first analyze how the spread of the virus happens according to each food factor. For the implementation of this the programming language used in Python. Therefore to the data modeling here it used several libraries and packages like matplotlib and plotty. The results which get from the data modeling section will discuss in the next chapter

The most important part of the research is predicting the spread of the COVID-19 virus according to the food categories. For that here also used the same data set. The implementation process of the prediction has two main sections. They are
1. Data Pre-processing
2. Choosing the correct model and do the prediction.

In the preprocessing here it remove null values and encode the strings. Moreover for improving the accuracy of the model here it standardized the data. Moreover here it analyze the correlations
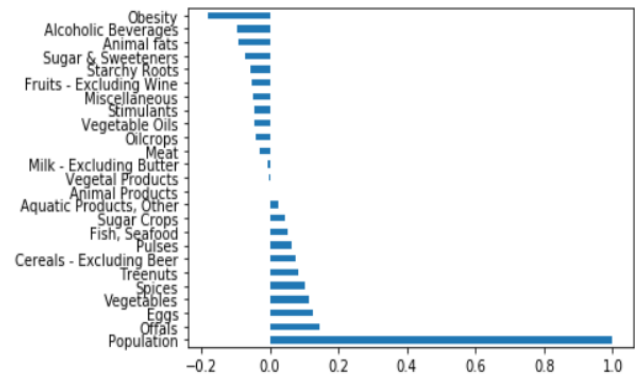

Figure 5: Correlations

By the above image, it can understand that the population is the most correlated attribute. Therefore here it remove the highest correlated variable (Population) from the data frame because the model mostly depends on that attribute. However, the population is one of the main reasons for the spread of the virus. But since this research analyzes how the spread is distributed according to the food categories usage authors removed population attribute.

In choosing the correct algorithmic approach process here it used the regression model approach because here it needs to predict a specific value. Therefore here it analyzes several regressions models from the Scikit learn libraries and choose the best accurate model for the prediction.

The above image shows the used algorithms to choose the best accurate model. So here the BaggingRegresseor model gives the highest accuracy. Therefore the predictions here this research use the BaggingRegresssor model. The results of each model will discuss in the next chapter further.

```
algorithms = []
algorithms.append(('LinearRegression', LinearRegression()))SSS
algorithms.append(('BaggingRegressor', BaggingRegressor()))
algorithms.append(('RandomForest', RandomForestRegressor()))
algorithms.append(('KNeighbours', KNeighborsRegressor()))
algorithms.append(('Boosting',XGBClassifier()))
```

Figure 6: Used Algorithms

This is the implementation process of the research to analyze and predict the impact of food categories to the distribution of the COVID-19 virus and predict the spread based on that, The whole process of the implementation of the research could be clearly described by the below flow diagram briefly.

Figure 8: Flow Diagram of the research.

Therefore this is the main process of the research implementation. In the next chapter, it will clearly describe the chosen data set, the result of the data modeling part, and the accuracy of the trained models for the spread prediction based on food categories used.

IV. DATA AND RESULT

Since this research is a supervised learning data science project here it used a data set that includes relevant data for the implementation process. Here the authors choose a data set which includes country-wise confirmed and death cases of the Covid-19 Virus and the percentage of the usage of each food category according to the population of the country from a well-known machine learning data set providing web page Kaggle. Moreover, the chosen data set is updating daily by the owner of the dataset.

When it concentrates on the elements of the dataset there are 170 countries data available. Moreover, there are nearly 31 attributes in the dataset. Below Image shows those attributes in the dataset.

```
data.columns

Index(['Country', 'Alcoholic Beverages', 'Animal fats', 'Animal Products',
       'Aquatic Products, Other', 'Cereals - Excluding Beer', 'Eggs',
       'Fish, Seafood', 'Fruits - Excluding Wine', 'Meat',
       'Milk - Excluding Butter', 'Miscellaneous', 'Offals', 'Oilcrops',
       'Pulses', 'Spices', 'Starchy Roots', 'Stimulants', 'Sugar & Sweeteners',
       'Sugar Crops', 'Treenuts', 'Vegetable Oils', 'Vegetables',
       'Vegetal Products', 'Obesity', 'Undernourished', 'Confirmed', 'Deaths',
       'Recovered', 'Active', 'Population'],
      dtype='object')
```

Figure 9: Attributes in the dataset

As mentioned in the Implementation chapter there needs to pre-process this dataset because there are some attributes which highly correlated to the prediction and there are some null value columns.

The process of the implementation was done according to 2 sub-areas. They are

1. Data Modeling
2. Prediction Model

When considering the result of the data modeling first here it analyzes the distribution of each attribute of the dataset using matplotlib hist method.
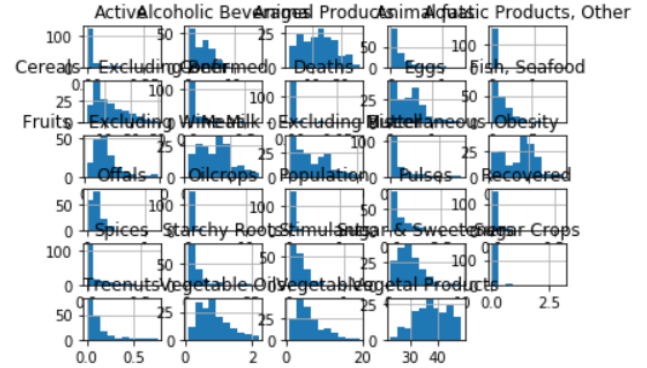

Figure 10: Distribution of the Attributes

The above image shows the result that the authors get for the distribution analysis of each attribute. Thereafter here it analyzes how the number of confirmed cases of the COVID-19 changes according to each food category. In these plots, the x-axis represents the supply percentage of each food category annually according to the population and the y-axis represents the percentage of confirmed cases according to the population. So here, authors get meat usage of meat and eggs is highly correlated with the spread of the virus.
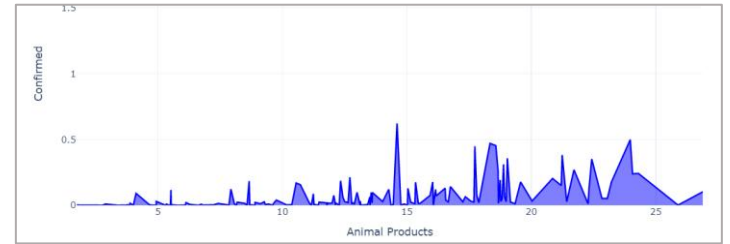

Figure 10: Spread of the Virus with animal products usage

The above image shows how the spread of the virus happens according to the usage of meat products. Moreover here authors got that vegetable / non-veg products correlated for the spread of the virus in a very low manner when comparing to the effect of animal products.
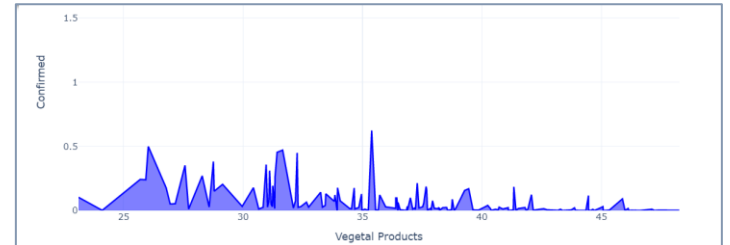

Figure 11: Spread of the Virus and Vegetable Products usage

By the above two graphs it can clearly understand that more animal product usage effect highly for the spread of the virus and less usage of the vegetable product is also effect highly for the spread.
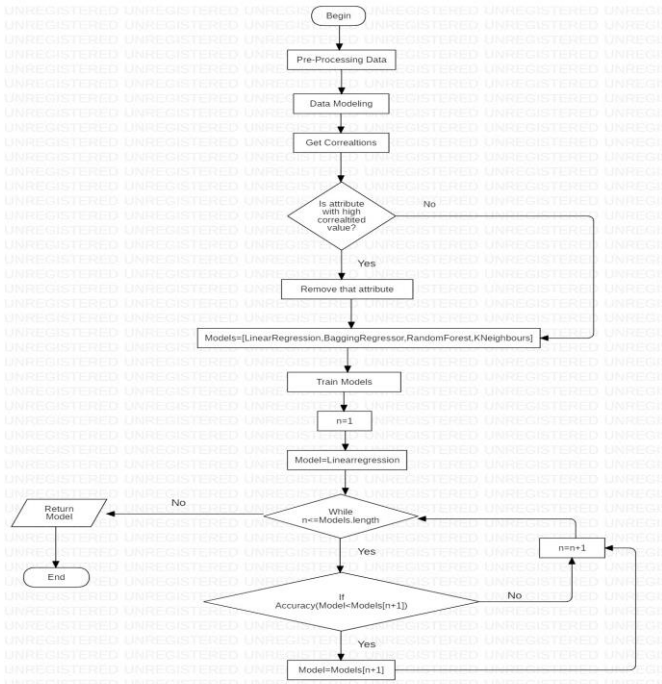
Moreover, here the authors analyzed how each food category affects the death rates of the COVID-19 virus.
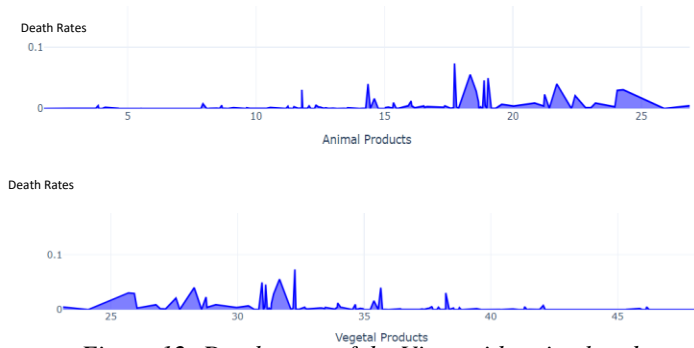

Figure 12: Death rates of the Virus with animal and vegetable product usage

Here also the x-axis of the plot represents the percentage of annual Animal product supply and the y-axis represent the number of confirmed cases according to the population. According to the above image, it can clearly understand how the usage of Animal products affects the death rates of the virus. Moreover from these plots, it can clearly understand that high usage of vegetable products controls the spread of the virus. So these are some main results got by the user in the data modeling process.

In the prediction model here it predicts the spread of the virus according to each food category. To that, as mentioned in the Implementation chapter authors used regression algorithms. Here in the implementation process it analyzes the accuracy of the several regression models and uses the best model.
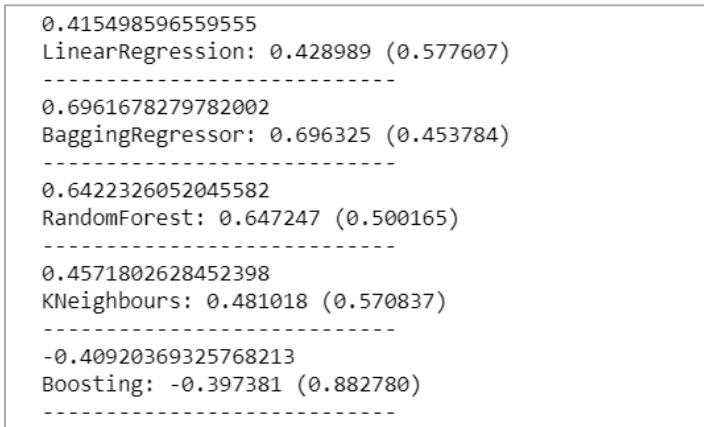
```
0.415498596559555
LinearRegression: 0.428989 (0.577607)
---------------------------
0.6961678279782002
BaggingRegressor: 0.696325 (0.453784)
---------------------------
0.6422326052045582
RandomForest: 0.647247 (0.500165)
---------------------------
0.4571802628452398
KNeighbours: 0.481018 (0.570837)
---------------------------
-0.40920369325768213
Boosting: -0.397381 (0.882780)
---------------------------
```

Figure 13: Accuracy of each model

The above image shows the variance and the accuracy that the authors got for each regression model. Therefore here the Baggingregressor model has the highest accuracy rate which is 69.61%. Therefor for the prediction process authors choose this model.

These are the results that got in each data science component of the research. Since the data set has only 170 countries data identifying a patter between each class is a challenging task. In the next chapters, it will discuss the limitations and the future enhancement of the research.

V. DISCUSSION

The main purpose of this research is to analyze how the spread of the novel coronavirus (COVID-19) changes with the usage of good categories and predict the spread of the virus from that. The research idea was gained by analyzing past research done to analyze how the human Immune system response for the Covid-19 virus. From that authors entered to an approach that how the Food categories cause for the Immune system and how those food categories causes for the spread of the Covid-19 virus. Since there are several pieces of research done how the food categories cause for the Immune system authors further analyzed and predict a prediction model which predict the spread of the Covid-19 virus with the usage of those food categories because according to the past researches it can clearly understand that food categories are highly responsible for the process of the human Immune system.

Thereafter authors got data components that includes how each food categories used in country wise and the number of COVID-19 virus confirmed cases. S mentioned in the implementation chapter thereafter authors implemented a data modeling and prediction components. During the implementation process, the authors didn't face any social or legal issues. Because the data which used are publically available on the Kaggle web page.

However, since this research is based on the Food categories and spread of the Covid-19 other key factors which cause the spread like social distance and the population was not concentrated by the authors. In the next chapter, it will discuss the future enhancements and the limitation of this research.

VI. CONCLUSION

This research was done to analyze and predict the spread of the COVID-19 virus based on the usage of food categories. To identify a pattern to predict the spread first authors have to do the data modeling. Thereafter authors chose the regression approach for the prediction process. However, in the implementation process, there were several limitations can be understood. They are

1. There are only 170 countries data available. Because of that, it's hard to increase the accuracy of the model.
2. The data will be updated daily. Therefore it is good that train the model for most relevant data.
3. The key reason for the spread of the COVID-19 virus is social distance. However here the authors ignore that.

So these are the main limitation identified in this research. So next it will discuss the future enhancement of the research.

1. Gather more data and accurate the model.
2. To accurate, the model authors suggest implementing an LSTM Neural network model.

Therefore, these are the future enhancement that could be done for this research. So this is the way how the research was done by the authors. Authors hope that this research will give a brief explanation about how the food categories cause the spread of the virus and the prediction model will be a novel approach to keep humans safe.

REFERENCES

[1]   WorlOmeter, "COVID-19 CORONAVIRUS PANDEMIC," WorlOmeter, 20 5 2020. [Online]. Available: https://www.worldometers.info/coronavirus/. [Accessed 20 5 2020].

[2]   BBC, "Coronavirus: Five of the countries most at risk from famine in 2020," 22 4 2020. [Online]. Available: https://www.bbc.com/news/world-52379956. [Accessed 16 5 2020].

[3]   BBC, "Which countries eat the most meat?," 4 2 2019. [Online]. Available: https://www.bbc.com/news/health-47057341. [Accessed 16 5 2019].

[4]   Physiopedia, "Coronavirus Disease (COVID-19)," Physiopedia, 2020. [Online]. Available: https://physio-pedia.com/Coronavirus_Disease_(COVID-19). [Accessed 16 5 2020].

[5]   T. Nace, "Coronavirus Map: How To Track Coronavirus Spread Across The Globe," 16 3 2020. [Online]. Available: https://www.forbes.com/sites/trevornace/2020/03/16/coronavirus-map-how-to-track-coronavirus-spread-across-the-globe/#34308763f980. [Accessed 16 5 2020].

[6]   Johns Hopkins University, "New Cases of COVID-19 In World Countries," Johns Hopkins University, 20 5 2020. [Online]. Available: https://coronavirus.jhu.edu/data/new-cases. [Accessed 20 5 2020].

[7]   Worldometer, "Coronavirus Cases China," Worldometer, 16 5 2020. [Online]. Available: https://www.worldometers.info/coronavirus/country/china/. [Accessed 16 5 2020].

[8]   WHO, "WHO Coronavirus Disease (COVID-19) Dashboard," WHO, 19 5 2020. [Online]. Available: https://covid19.who.int/?gclid=CjwKCAjwh472BRAGEiwAvHVfGnOF1LYfHlUqHiJYHqCwnrlKcYwwc1zfTD5eF4OegeLX8M9vjoDdzhoCfVwQAvD_BwE. [Accessed 20 5 2020].

[9]   B E T T E R B Y D E S I G N, " LCHHF Foods – food categories," Wordpress, [Online]. Available: https://www.bbdnutrition.com/food-for-thought/lchhf-foods-food-categories/. [Accessed 21 5 2020].

[10]   K. B. KC, "When too much isn't enough: Does current food production meet global nutritional needs?," doi.org, 2018.

[11]   2. V. O. P. N. J. Poore1, "Reducing food's environmental impacts through producers and consumers," Science, vol. 360, no. 6392, pp. 987-992, 2018.

[12]   S. G. Al-Shawi, "The Effect of Nutrition on Immune System Review Paper," Researchgate, 2019.

[13]   D. Wu, "Nutritional Modulation of Immune Function: Analysis of Evidence, Mechanisms, and Clinical Relevance," Frontiers Media, Bethesda MD, 2019.

[14]   C. M. Galanakis, "The Food Systems in the Era of the Coronavirus," MDPI, Switzerland, 2020.

[15]   R. H. Farah Naja, "Nutrition amid the COVID-19 pandemic: a multi-level framework," Springer, 2020.

[16]   HLPE, "Impact of COVID-19 on Food Security and Nutrition," Rome, 2020.

[17]   C. Z. PengaYiming, "Propagation analysis and prediction of the COVID-19," Infectious Disease Modelling, vol. 5, pp. 282-292, 2020.

[18]   A. l. o. o. panelZebinZhao, "Prediction of the COVID-19 spread in African countries and implications for prevention and control: A case study in South Africa, Egypt, Algeria, Nigeria, Senegal and Kenya," Science of The Total Environment, vol. 729, 2020.

[19]   A. M.K., "Modeling and Predictions for COVID 19 Spread in India," India, 2020.

[20]   B. J. page, "The covid-19 pandemic in two animated charts," 27 3 2020. [Online]. Available: https://www.technologyreview.com/2020/03/27/950263/the-covid-19-pandemic-in-two-animated-charts/. [Accessed 16 5 2020].

[21]   N. Routley, "This interactive chart shows countries flattening their COVID-19 curves," World Economic Forum, 13 5 2020. [Online]. Available: https://www.weforum.org/agenda/2020/03/covid19-coronavirus-countries-infection-trajectory/. [Accessed 19 5 2020].

[22]   D. Matthews, "9 CHARTS THAT EXPLAIN THE CORONAVIRUS PANDEMIC," Sri Lankan Brief, 15 5 2020. [Online]. Available: https://srilankabrief.org/2020/03/9-charts-that-explain-the-coronavirus-pandemic/. [Accessed 19 5 2019].