Analysis of the impact of the COVID-19 pandemic on the global

Haotian Chen[1]

(Dated: May 31, 2023)

Abstract: The novel coronavirus has swept the world, destroyed the global supply chain, destroyed the economic stability, and affected all aspects of people's lives. It is particularly important to study the impact of the novel coronavirus epidemic on the global economy and judge the infection of the novel coronavirus.

Firstly, the relevant test data are briefly introduced, and the Pearson correlation coefficient is used to calculate and analyze the correlation between the data, and it is judged that there is a strong linear correlation between GDP and HDI, total cases and total deaths. GDP has a weak correlation with STI, total cases and population, total deaths and population. Based on this, we plot the GDP as a function of HDI and STI to visually observe the correlation with each other. It is found that GDP fluctuates up and down with the increase of HDI and STI. According to the images of STI changes over time, some countries never paid attention to COVID-19 from the outbreak of COVID-19 to the end of 2020, and most countries' government policies were first strict and then slowed down over time. Then, a time series model is established, and the rolling mean is used to smooth the short-term fluctuations in the time series data and highlight the long-term trend. It is concluded that the mortality rate increases exponentially around April 2020 and gradually slows down in June. It may decline in November, and after October, the epidemic in countries around the world may be basically under stable control to a certain extent. STI also increased exponentially in March 2020, remained stable in April and showed a small downward trend (policy slowdown), and is expected to decline significantly after October.

We then analyzed the impact of COVID-19 on a single country and plotted the changes of each indicator across countries around the world over time, mainly selecting China, the United States and Cambodia. It can be concluded that the COVID-19 epidemic first broke out in China and spread rapidly, but it was controlled in March in time. However, the epidemic control and management in the United States are not good. Cambodia's death toll remained at zero.

In view of how to determine whether you are infected with the new coronavirus, I established decision tree, random forest, adaboost, naive Bayes and other models, and made comparative analysis, and finally concluded that the decision tree had the best classification effect for the data set selected in this paper.

For GDP, global happiness index and other indicators, I used kmeans algorithm to cluster countries, and used elbow rule to select the appropriate number of clusters. Grouping the world's countries

into five groups, the first group has a happier population, freer population, more developed economy... And use python to visualize the results. Then, the time series model is continued to be used to predict the GDP of China and the United States by first taking the logarithm and the difference to make the series able to reject the null hypothesis. Projections show that China's GDP will continue to rise in 2021, but the US will decline. The pandemic's impact on the U.S. could be more lasting.

Keywords: COVID-19; Economic globalization; The digital economy; Stringency Index; GDP; Decision tree; adaboost; Kmeans; Time series;

1 Introduction

The COVID-19 pandemic has disrupted global supply chains and destabilized economies. Global supply chain is an important force supporting economic globalization. The vast and complex network of supply chains linking countless manufacturing and service companies across the globe has made the world economy intertwined into an organic, interdependent and integrated whole, but at the same time it has increased the vulnerability of the global economy. In a highly interdependent global economic system, a temporary production shutdown or trade restriction in any economy, especially one that is a key link in the global supply chain, will have a significant external impact on other economies. The COVID-19 pandemic is a global health crisis, and in the same way it hurts the fabric of the global economy built through supply chains. After the outbreak of the epidemic, from cars to smartphones, from shopping to travel, from aviation services, financial services to technology services... The international production of countless goods and services around the world has been forced to slow down or even suspend, the population and labor force have shrunk, and the cost of producing goods has increased dramatically... The extreme fragility of global supply chains has been exposed, fueling a rise in panic and new accusations of economic globalisation.

2 Analysis and Inquiry

2.1 The impact of COVID-19 on countries

First of all, we need to read and sort out the data. Next, we will analyze the impact of COVID-19 on the country from the following aspects: Total cases, Total Death, Population, HDI, STI, GDP...

2.1.1 Total Cases and Total Deaths

In the data, the corresponding meanings are total cases and total deaths. According to the sorted data, it can be seen intuitively that these two data are increasing with the growth of time. And the death toll is always less than the number of cases, this means that infect the novel coronavirus does not represent the death, and we can also get through these two data, mortality of patients with novel coronavirus, which can analyze the obtained novel coronavirus of cure rate, each country cure rate with the national advanced level, GDP has certain correlation relationship.

	Country_code	Country	Total_Population	Total_Cases	Total_deaths	Stringency_Index	GDP	Human_Index
27	BRA	Brazil	19.174732	425704517.0	14340567.0	3.136028	9.828942	0.759000
90	IND	India	21.045353	407771615.0	7247327.0	3.610552	9.189712	0.640000
157	RUS	Russia	18.798668	132888951.0	2131571.0	3.380088	7.358760	0.816000
150	PER	Peru	17.311165	74882695.0	3020038.0	3.430126	11.356685	0.599490
125	MEX	Mexico	18.674802	74347548.0	7295850.0	3.019289	10.544307	0.774000
178	ESP	Spain	17.660427	73717676.0	5510624.0	3.393922	9.930685	0.887969
175	ZAF	South Africa	17.898266	63027659.0	1357682.0	3.364333	11.116819	0.608653
42	COL	Colombia	17.745037	60543682.0	1936134.0	3.357923	9.650207	0.581847
92	IRN	Iran	18.246243	52421884.0	2914070.0	3.207064	8.129609	0.798000
40	CHL	Chile	16.766047	51268034.0	1283880.0	2.989194	7.254100	0.656622

Figure 1: Total Cases

	Country_code	Country	Total_Population	Total_Cases	Total_deaths	Stringency_Index	GDP	Human_Index
27	BRA	Brazil	19.174732	425704517.0	14340567.0	3.136028	9.828942	0.759000
125	MEX	Mexico	18.674802	74347548.0	7295850.0	3.019289	10.544307	0.774000
90	IND	India	21.045353	407771615.0	7247327.0	3.610552	9.189712	0.640000
97	ITA	Italy	17.917523	50752853.0	6664225.0	3.629838	6.623784	0.880000
68	FRA	France	17.994097	50084335.0	5633444.0	3.385794	9.517375	0.901000
178	ESP	Spain	17.660427	73717676.0	5510624.0	3.393922	9.930685	0.887969
150	PER	Peru	17.311165	74882695.0	3020038.0	3.430126	11.356685	0.599490
92	IRN	Iran	18.246243	52421884.0	2914070.0	3.207064	8.129609	0.798000
157	RUS	Russia	18.798668	132888951.0	2131571.0	3.380088	7.358760	0.816000
42	COL	Colombia	17.745037	60543682.0	1936134.0	3.357923	9.650207	0.581847

Figure 2: Total Deaths

To some extent, the more cases there are, the more deaths there are. However, comparing the two tables, we can see that the number of cases in some countries is not in the top ten, but the number of deaths is in the top ten, indicating that some countries have a high death rate. It is reasonable to speculate that the medical level of this country is not developed, or the economic level is not high.

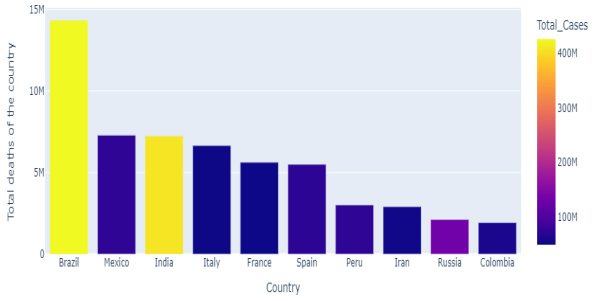


Figure 3: Mixed cases and deaths.

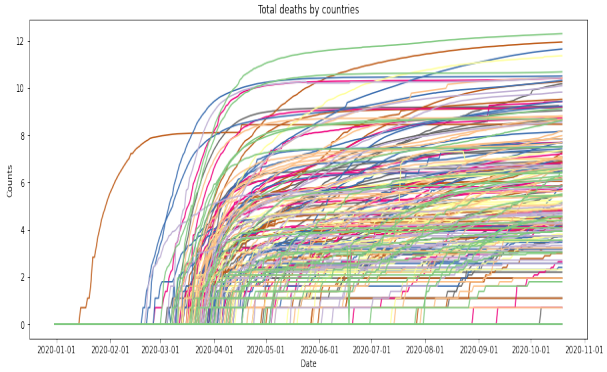


Figure 4: deaths toll over time

From figure 3 we can see intuitively that Brazil has the highest number of cases and deaths. Mexico has the second highest number of cases, followed by: India, Italy, France, Spain, Peru, Iran, Russia, Colombia, and yellow color indicates a higher death toll. From figure 4 We can see that with the spread of the novel coronavirus, the death toll increases day by day, and the growth is rapid. After a certain time, the death toll reaches a stable state, which can also reflects the transmission law of the novel coronavirus. It is also speculated that the outbreak period of the novel coronavirus is concentrated in February and March.

2.1.2 Population

Obviously, this data represents the population of each country, and also represents the productivity level of the country to some extent.

From figure 5, We can know the ranking of population. The top ten are China, India, Indonesia,

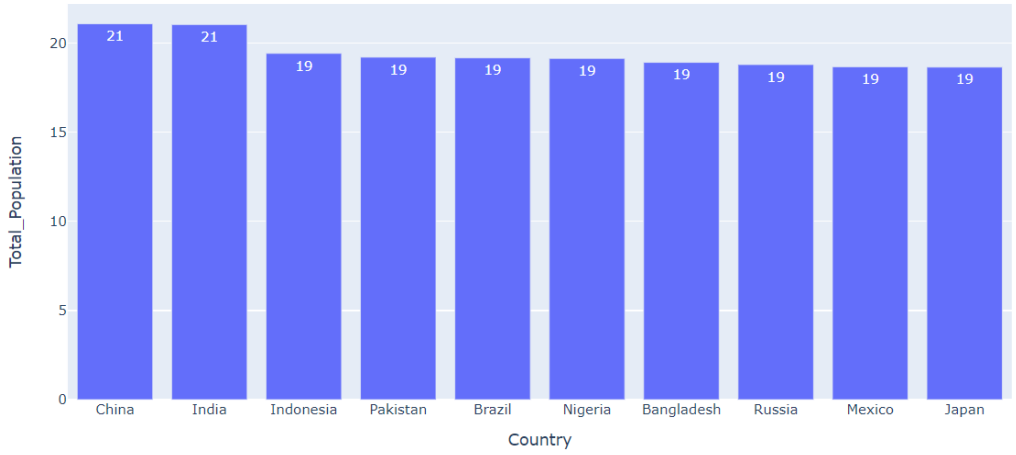


Figure 5: Population

Pakistan, Brazil, Nigeria, Bangladesh, Russia, Mexico and Japan.To some extent, the data also reflect the country's productivity level.

2.1.3 GDP

Gross domestic product (GDP) is the final result of production activities of all permanent units in a country (or region) in a certain period of time. This value is the core index to measure a country's national economy, A country's GDP or gross domestic product is calculated by taking into account the monetary worth of a nation's goods and services after a certain period of time, usually one year.It's a measure of economic activity.

How Does GDP Per Capita Work? Essentially, GDP per capita acts as a metric for determining a country's economic output per each person living there. Often times, rich nations with smaller populations tend to have higher per capita GDP. Once you do the math, the wealth is spread among fewer people, which raises a country's GDP.The fact that the GDP per capita divides a country's economic output by its total population makes it a good measurement of a country's standard of living, especially since it tells you how prosperous a country feels to each of its citizens.Through relevant data, we can also indirectly analyze the impact of the COVID-19 outbreak on the national economy.

From figure 6, We can see intuitively that the top 25 countries in GDP to some extent reflect the relatively high economic level of these countries.

2.1.4 HDI

Human Development Index, it was created to emphasize that people and their capabilities should be the ultimate criteria for assessing the development of a country, not economic growth alone. The HDI can also be used to question national policy choices, asking how two countries with the same level of GNI per capital can end up with different human development outcomes. These contrasts can stimulate debate about government policy priorities. The Human Development Index (HDI) is

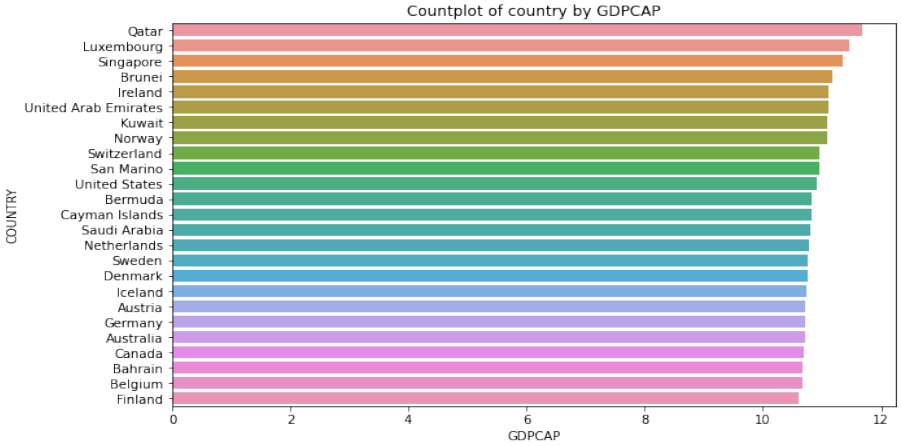


Figure 6: Top 25 countries of GDP

a summary measure of average achievement in key dimensions of human development: a long and healthy life, being knowledgeable and have a decent standard of living. The HDI is the geometric mean of normalized indices for each of the three dimensions.

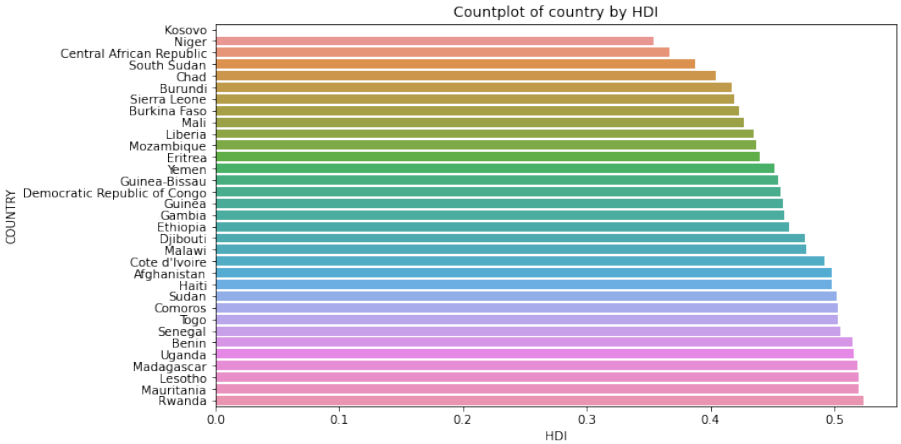


Figure 7: Top 33 countries of HDI

The health dimension is assessed by life expectancy at birth, the education dimension is measured by mean of years of schooling for adults aged 25 years and more and expected years of schooling for children of school entering age. The standard of living dimension is measured by gross national income per capital. The HDI uses the logarithm of income, to reflect the diminishing importance of income with increasing GNI. The scores for the three HDI dimension indices are then aggregated into a composite index using geometric mean.

2.1.5 STI

The Oxford Coronavirus Government Response Tracker project calculate a Stringency Index, a composite measure of nine of the response metrics. The nine metrics used to calculate the Stringency

Index are: school closures; workplace closures; cancellation of public events; restrictions on public gatherings; closures of public transport; stay-at-home requirements; public information campaigns; restrictions on internal movements; and international travel controls. The index on any given day is calculated as the mean score of the nine metrics, each taking a value between 0 and 100.

A higher score indicates a stricter response (i.e. 100 = strictest response). If policies vary at the subnational level, the index is shown as the response level of the strictest sub-region.

2.1.6 Correlation between data

Pearson's correlation coefficient, First, Pearson correlation coefficient was calculated to observe the relationship between the variables.

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E\left[(X-\mu_X)(Y-\mu_Y)\right]}{\sigma_X \sigma_Y} \tag{1}$$

Pearson's correlation coefficient ranges from -1 to 1. A coefficient value of 1 means that X and Y are well described by the equation of the line, that all the data points fall nicely on the same line, and that Y increases as X increases. A coefficient value of -1 means that all the data points fall on the line and Y decreases as X increases. A coefficient value of 0 means that there is no linear relationship between the two variables.

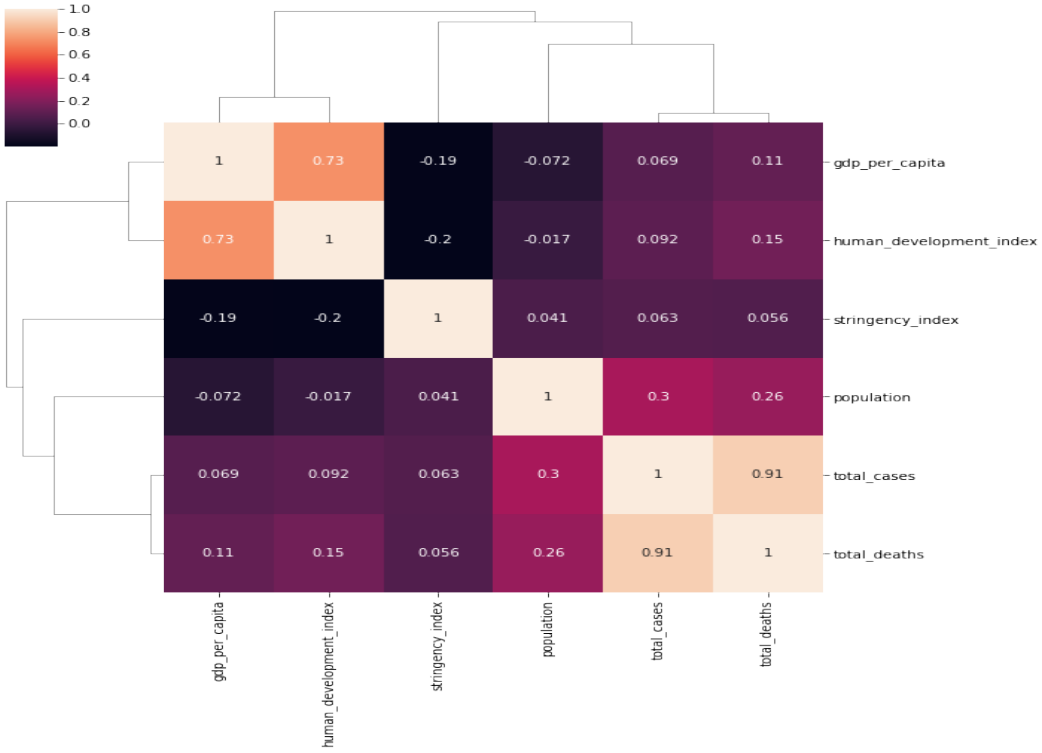


Figure 8: Correlation coefficient diagram

From figure 8, It can be seen that there is a strong linear correlation between GDP and HDI, total cases and total deaths. And they are all positively correlated. There was a weak correlation between

total deaths and population, GDP and STI, total cases and population.

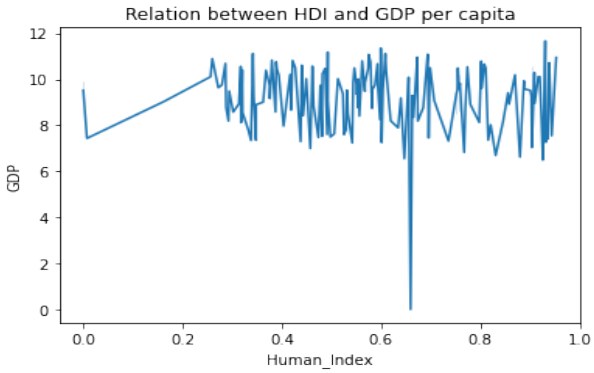


Figure 9: HDI-GDP

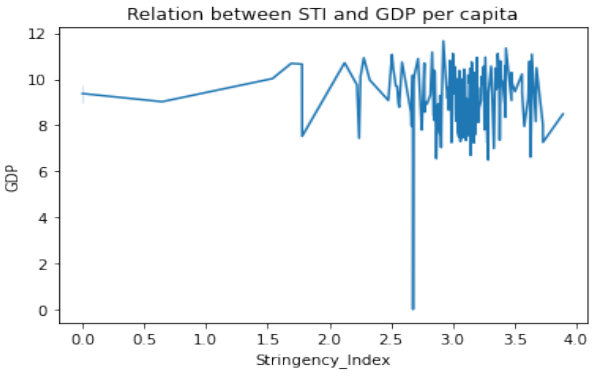


Figure 10: STI-GDP

The correlation between cases and deaths is described above and will not be analyzed here. Since HDI and GDP, STI and GDP have a certain correlation, we observe the corresponding images, and we can see that GDP fluctuates up and down with the increase of HDI and STI, which indicates that HDI and STI affect GDP to a certain extent. However, when HDI is around 0.65, and STI is around 2.6, GDP drops sharply. Which can only show that a country's GDP is very low, HDI and STI are also very low, and people in this country's happiness level may be not high.

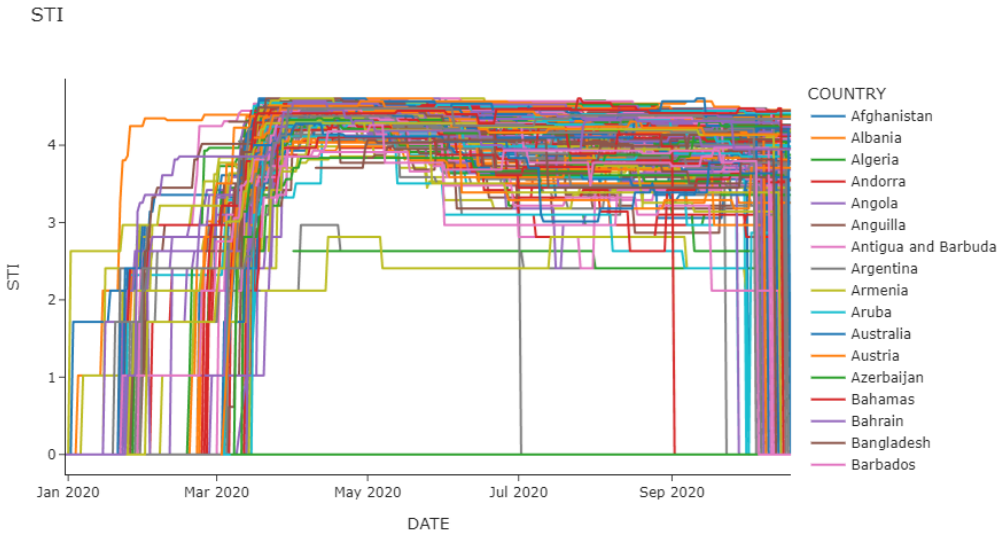


Figure 11: STI-DATE

From figure 11, we can see STI index increased first and then the steady for a period of time and then the final reduction, treat the novel coronavirus that can infer the countries policies are tough first, for example: reduce people go out, shut down in public places, limited public transportation and so on, and then for a while, when the spread of the epidemic is slowly, gradually easing the policies. Of

course, some countries have never taken COVID-19 seriously.

2.1.7 Time series prediction

Inspecting time series and rolling mean, we use rolling means (or moving averages) to smooth out short-term fluctuations in time series data and highlight long-term trends. From figure 12 and 13, deaths and STI change over the time, Time series predictions for STI and deaths were performed. This is a kind of regression prediction method, which belongs to quantitative prediction. Its basic principle is: on the one hand, the continuity of the development of things is recognized, and the past time series data is used for statistical analysis to infer the development trend of things; On the other hand, the randomness due to the influence of accidental factors is fully taken into account. In order to eliminate the influence of random fluctuations, historical data is used for statistical analysis, and the data is properly processed for trend prediction. Looking at the chart, we can see that the time series may not be a stationary series. In order to verify the integrity, we can also add Dickey-Fowler test to detect the integrity of time series. Dickey-fowler test is to test whether an autoregressive model has a unit root, so as to help judge whether the series is stationary. As can be seen, the rate of deaths increase exponentially around April and slowed gradually in June. It may fall in November, and the rolling average is on an upward trend. However, the overall trend is stable, indicating that the COVID-19 epidemic in various countries around the world may be basically under stable control after October to some extent. STI also showed exponential growth in March, remained stable in April, and showed a slight downward trend (policy slowdown), and is expected to decline sharply after October. To some extent, this reflects that governments around the world have relaxed restrictions on people's movements, and also reflects that the COVID-19 epidemic has been effectively controlled.

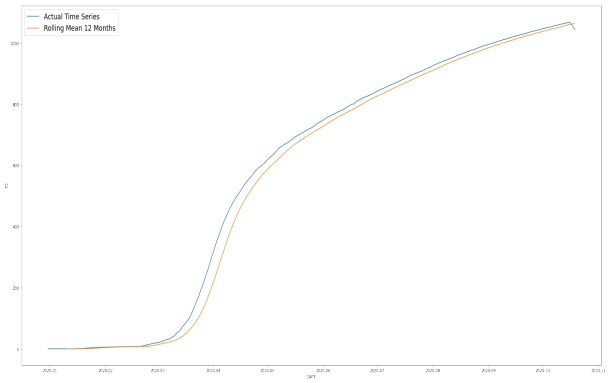


Figure 12: Time series forecasts of deaths

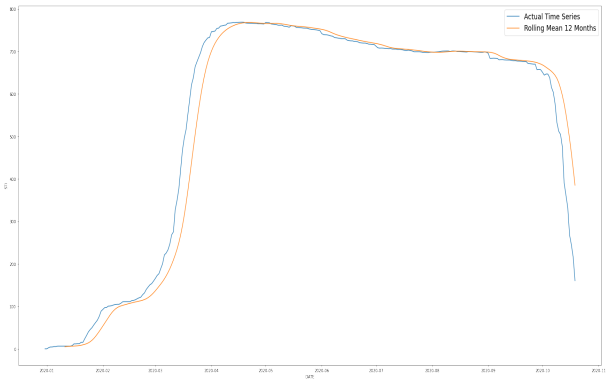


Figure 13: Time series forecasts of STI

2.2 Single country impact

This section mainly analyzes the impact of the COVID-19 pandemic on the world by comparing data from different countries.

2.2.1 China,US,Cambodia

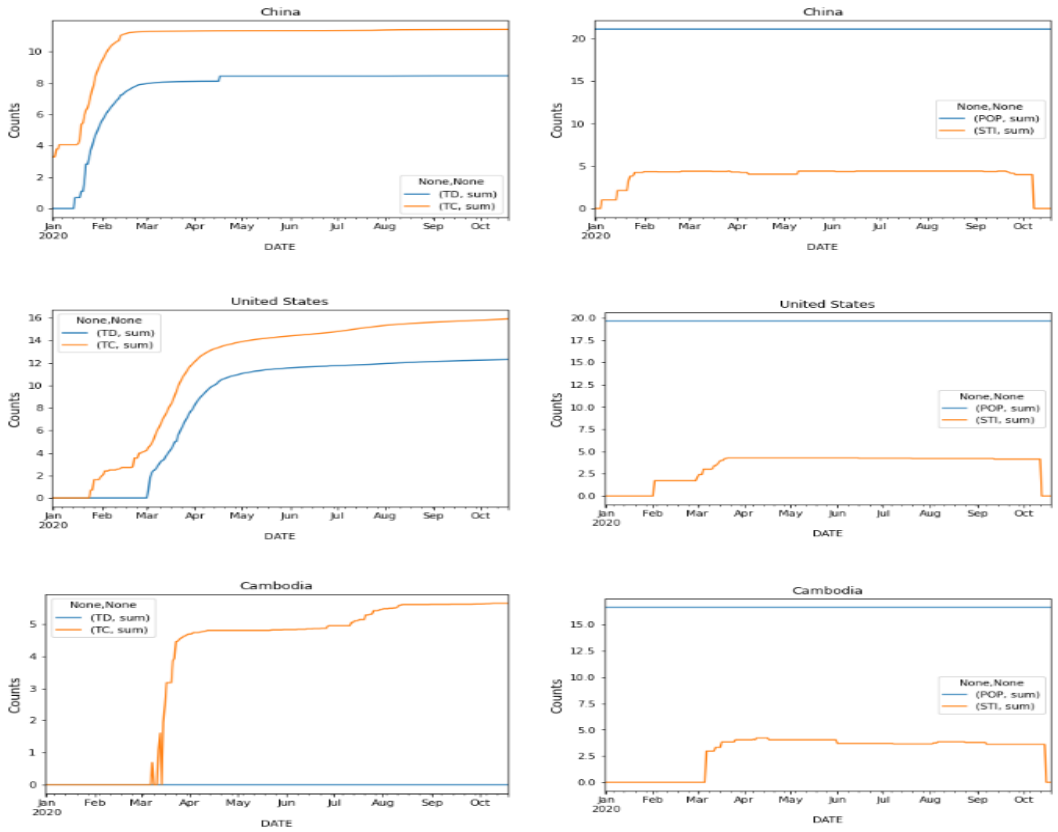


Figure 14: China,US,Cambodia

In the following, we will conduct more specific data observation and analysis for some typical countries, such as China, the United States, Cambodia, and other countries. To better study the global impact of the pandemic. And the analysis of GDP will be carried out in detail below.

Through comparison, we can see that the impact of COVID-19 on different countries is different. We can clearly see that the outbreak of COVID-19 in China started early in January, and the epidemic spread rapidly due to the large population in China, and was controlled around March. China's policies are also being implemented quickly compared to those of the United States and Cambodia. In the United States, the outbreak started around the end of January, was equally stable growing but still growing around March, and the implementation of the U.S. policy was completed around April. In Cambodia, the outbreak began at the beginning of March, but the death toll remained at zero.

2.2.2 countries at dates

From figure 15, As you can see, both the number of deaths and the STI were 0 in December 2019, indicating that the pandemic had not yet swept the world. The global GDP index is not very high, and some countries in Europe, North America and Oceania have relatively high GDP levels. By comparison, we can find that the HDI index of a country has a certain relationship with GDP. When the GDP level

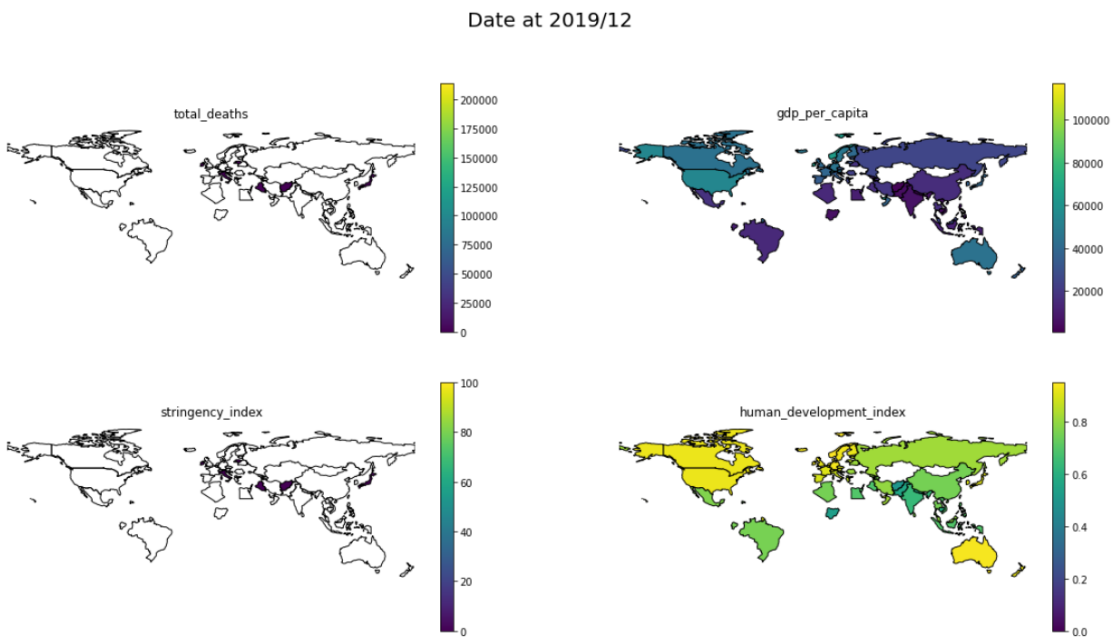


Figure 15: Worlds date at 2019/12

of a country is high, the HDI index is relatively high. As can be seen from the figure, the HDI index is also higher in some countries in North America, Europe and Oceania. The number of deaths and the STI index begin to change in countries around the world when the date comes to February 2020, and gradually expand to include a global range and darker color over time. The data from February 2020 to October 2020 are detailed in the appendix.

3 Classification diagnosis of COVID-19

Classification refers to learning statistical regularity from labeled data, that is, finding a mapping function to map input and output variables, essentially making predictions on the input data and qualitatively outputting the results. We will use decision tree, random forest, adaboost, naive Bayes and other models to judge whether the patient is infected with the new coronavirus, analyze the results, and compare the quality of the model.

The data processing and integration process will not be covered here. We split the data into test and training sets with a ratio of 8:2. And the COVID-19 test results are divided into four categories, namely: missing; No infection; The unknown; And have been infected. And replaced by 0,1,2,3 in the way of labels.

3.1 Decision tree

In machine learning, a decision tree is a predictive model. It represents a mapping between object properties and object values. Each node in the tree represents some object, each branching path

represents some possible attribute value of, and each leaf node corresponds to the value of the object represented by the path traversed from the root node to that leaf node.

3.1.1 Model training and solving

$$\mathrm{GINI}(D, A) = \frac{D_1}{D}\,\mathrm{GINI}\,(D_1) + \frac{D_2}{D}\,\mathrm{GINI}\,(D_2) \tag{2}$$

By training the model, the decision nodes, leaf nodes, and the depth of the decision tree that the model adapts to the data are judged. The classical algorithms of decision tree are ID3, C4.5, CART, etc. We use CART algorithm, which adopts Gini coefficient as the index of attribute selection. Gini coefficient reflects the uncertainty of samples. When Gini coefficient is smaller, it means that the difference between samples is small and the uncertainty is low. We can see that in the classification results, the model of decision tree has an accuracy of 0.783. When the original data is missing, the decision tree correctly judges that there are 564908 missing samples. When the original data is infected, the decision tree model correctly judges that there are 592355 infected samples... The sample size for all correct judgments is 564908+592355+147257+11254=1315774. When the original data is missing, the decision tree makes a wrong judgment, with 59703 samples judged to be infected, 8179 samples judged to be unknown, and 4202 samples judged to be not infected. We observe that the decision tree model misclassifies the data as missing and the data as not infected for a larger sample size. However, in fact, the probability of misclassifying the data as infected is the highest, because we calculate that 0.51 of the total infected data are correctly classified. The next highest probability is unknown data, followed by uninfected data, and missing data actually has the highest classification accuracy.

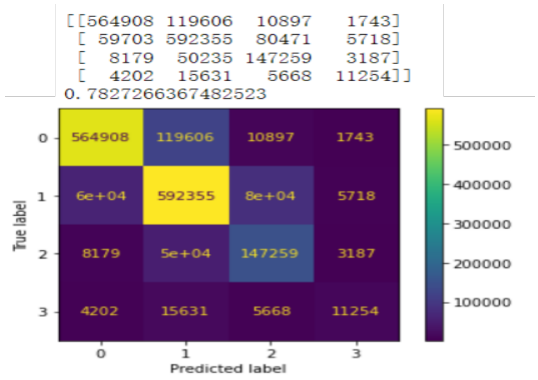


Figure 16: Decision tree result

3.2 Random forest

A random forest is a classifier that contains multiple decision trees whose output class is determined by the mode of the class output by the individual trees. Can handle a large number of input variables. For imbalanced categorical data, it can balance the error.

3.2.1 Model training and solving

A random forest is essentially an algorithm consisting of multiple decision trees that randomly sample a subset of the data with replacement. And a certain number of the most features are randomly selected as the input features of the decision tree. Finally, the class with the most votes is selected as the result.

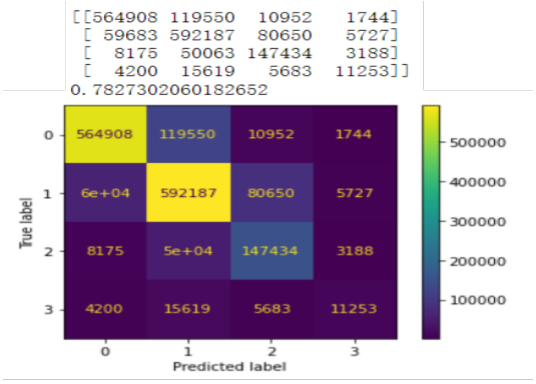


Figure 17: Random forest result

Observing the results We can see that the random forest has an accuracy of 0.783. The number of correctly classified samples is 564908+592187+147434+11253=1315782.

3.3 Adaboost

Adaboost is an iterative algorithm whose main idea is to train different classifiers (weak classifiers) for the same training set, and then combine these weak classifiers to form a stronger final classifier (strong classifier). The first weak classifier is obtained by learning N training samples, and the misclassified samples and other new data together form a new N training samples, and the third weak classifier is obtained by learning on this sample. Then, the previous two misclassified samples and other new samples are added to form another N new training samples, and the third weak classifier is obtained by learning this sample. Finally, the improved strong classifier is obtained, that is, the class of a data is determined by the weight of each classifier.

3.3.1 Model training and solving

In Adaboost model, there are two algorithms: SAMME and SAMME.R. SAMME algorithm uses the classification effect of the sample set as the weight of the weak learner, and SAMME.R is the predicted probability of the sample set classification as the weight of the weak classifier. We will use the SAMME.R algorithm directly for training, but we should be careful to limit the use of classifiers that support probabilistic predictions.

From Figure 18, we can see that the model of the adaboost has an accuracy of 0.757. When the original data sample was missing, the number of correctly classified samples was 559,706; when the original data sample was infected, the number of correctly classified samples was 603,042; when the

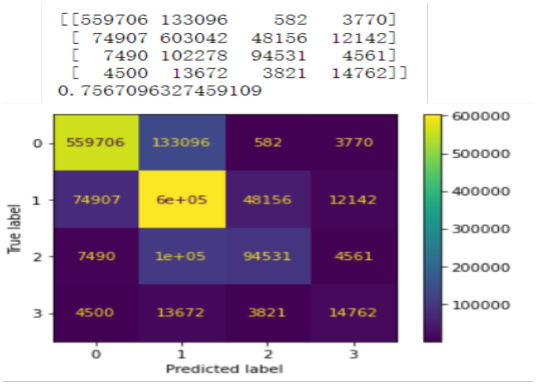


Figure 18: Adaboost result

original data sample was unknown, the number of correctly classified samples was 94,531; when the original data sample was not infected, the number of correctly classified samples was 14762. The total number of correctly classified samples is 559706+603042+94531+14762=1272041. We can see that the biggest misprediction is for people who are already infected, with almost half of the sample size misclassified. Let's not look down on it, it would be a big mistake in medicine.

3.4 other models

I also used linear discriminant analysis, logistic regression, and naive Bayes, but I won't go into each model in detail. Only a brief introduction is given here. Linear Discriminant analysis (LDA) is a supervised data dimensionality reduction method. In the process of data dimension reduction, the information provided by the category label of the data is used for judgment and analysis. The dimensionality of the labeled data is reduced and projected to the low-dimensional space while preserving as much information of the data samples as possible. After projection, the samples of the same class are as close as possible, and the samples of different classes are as far as possible. Logistic regression is essentially linear with a sigmoid function on top of the feature-to-outcome mapping, where the features are linearly summed and the prediction is made using the hypothesis function g(z), which maps the continuous values to 0 and 1. Bayesian method is based on the Bayesian principle, using the knowledge of probability and statistics to classify the sample data set. It first assumes that the features are independent of each other, and then learns a joint probability distribution from the input to the output through the given training set, assuming that the features are independent. Based on the learned model, the input X is used to find the output Y that maximizes the posterior probability.

3.5 Comparison of models

We will compare the models of linear discriminant analysis, logistic regression, Naive Bayes, decision tree, random forest, and adaboost. The analysis mainly focuses on the running speed of the training model and the accuracy of the results.

3.5.1 Run rate

In terms of running time, the naive Bayes model is the fastest, followed by the linear discriminant model, the decision tree model and the adaboost model. The logistic regression model and the random forest model run at a slower rate. From querying the data and training the model, we can see that the naive Bayes model is very suitable for large datasets and high-dimensional data, but the accuracy is usually lower than the linear model. Decision trees are also fast because they do not require scaling of the data. Random forest is more robust, but it is less suitable for high-dimensional sparse matrices.

3.5.2 Accuracy score

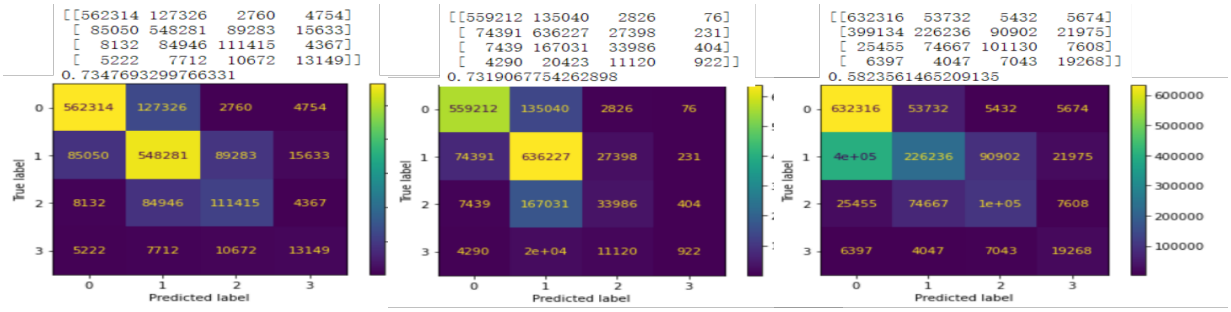


Figure 19: Linear discriminant result

Figure 20: Logistic regression result

Figure 21: Naive Bayes result

Intuitively, we can see that the linear discriminant model has a score of 0.734, the logistic regression model has a score of 0.732, and the naive Bayes model has a score of 0.582. We can observe that in all these models, it is easy to misclassify missing data as uninfected, easy to misclassify uninfected data as missing data and unknown data, and easy to misclassify unknown data as uninfected data. In terms of accuracy score, the random forest model has the highest accuracy score, and the naive Bayes model has the lowest accuracy score, only 0.582.

But, in fact, we shouldn't think that way. We should pay more attention to the accuracy of the models on uninfected and infected data. The classification accuracy of each class is calculated by dividing the number of correctly classified samples in each class by the total number of samples in that class. In this way, the linear discriminant model and the naive Bayes model are excluded first. The two models were calculated to misclassify the infection data up to 65 percent of the time. This means that 65 out of 100 people infected with the coronavirus will be considered either uninfected or undetermined. What a scary thing. The adaboost model achieves an accuracy of 0.42 on the infection data, which is still not good. The accuracy of logistic regression model, decision tree model and random forest model for the correct classification of uninfected data is as high as 0.87 to 0.89. The accuracy of correct classification of infection data is between 0.51 and 0.56. From the running speed, the classification accuracy of the model and other angles. A decision tree should be selected as the classification model for this dataset to determine whether a patient is infected with COVID-19.

3.5.3 Improvement and enhancement

In order to improve the accuracy of classification, the method of merging the sample size of missing data and unknown data can be used, so that the model is trained to classify three types. Then the accuracy of model classification is improved. We can also speed up the model by reducing the amount of sample data used to train the model. However, this may also cause the problem of reduced classification accuracy of the model. No attempt will be made in this paper.

4 Clustering and prediction

Since the GDP values for the individual countries in the previous dataset were recorded on a monthly basis, they remained constant between the end of 2019 and 2020. For better analysis and prediction, we aggregate new GDP data from 1960 to 2020 for countries around the world. We will use k-means clustering to cluster and analyze the data, and the elbow rule to determine the best k-means classification point. Then we use time series forecasting to forecast China's GDP data, analyze the results, and verify the accuracy of the model.

4.1 Clustering

The process of grouping a collection of physical or abstract objects into multiple classes consisting of similar objects is called clustering. A cluster generated by clustering is a set of data objects that are similar to each other and different from the objects in the same cluster.

4.1.1 K-means clustering

K-means clustering algorithm is an iterative clustering analysis algorithm. Its steps are to divide the data into k groups in advance, then randomly select k objects as the initial cluster centers, and then calculate the distance between each object and each seed cluster center. Assign each object to its nearest cluster center. The cluster centers and the objects assigned to them represent a cluster. Each time a sample is assigned, the cluster center is recalculated based on the existing objects in the cluster. This process is repeated until some termination criterion is met. The termination criteria can be that no (or minimum number of) objects are reassigned to different clusters, no (or minimum number of) cluster centers change again, and the sum of squared errors is locally minimum.

4.1.2 Model Checking and Training

Because the K-means clustering algorithm needs to calculate the distance of the samples, we need to standardize the data before building the model. The commonly used normalization methods include min-max normalization and z-score normalization. We directly adopt the z-score normalization method. We then use the elbow rule to find the optimal number of types to cluster this data. The result is 5. We start training the model and performing clustering. The clustering results are shown in Figure 22.

	cluster_0	cluster_1	cluster_2	cluster_3	cluster_4
Happiness Score	0.674292	1.398553	-1.127547	-0.376011	-0.803450
Lower Confidence Interval	0.665731	1.405922	-1.127978	-0.373870	-0.792671
Upper Confidence Interval	0.682235	1.389647	-1.125933	-0.377780	-0.813498
Economy (GDP per Capita)	0.492601	1.227856	-1.316820	0.198655	-0.786771
Family	0.581734	1.033011	-1.172858	-0.293988	-0.149191
Health (Life Expectancy)	0.470285	1.103235	-1.427762	0.260222	-0.470099
Freedom	0.239866	1.241521	-0.680486	-0.718942	0.364065
Trust (Government Corruption)	-0.248479	1.684637	-0.258970	-0.541506	0.005748
Generosity	-0.365158	1.037772	-0.115695	-0.695207	1.161080
Dystopia Residual	0.636468	0.101565	0.072789	-0.432934	-1.206164

Figure 22: Clustering result

From the table, we can see that countries with high happiness index, high GDP level, relatively healthy, free and high trust degree are divided into one group (cluster-1), followed by countries with average happiness index and middle GDP level (cluster-2)...Next, let's look at the distribution of happiness, GDP and other categories across the world.

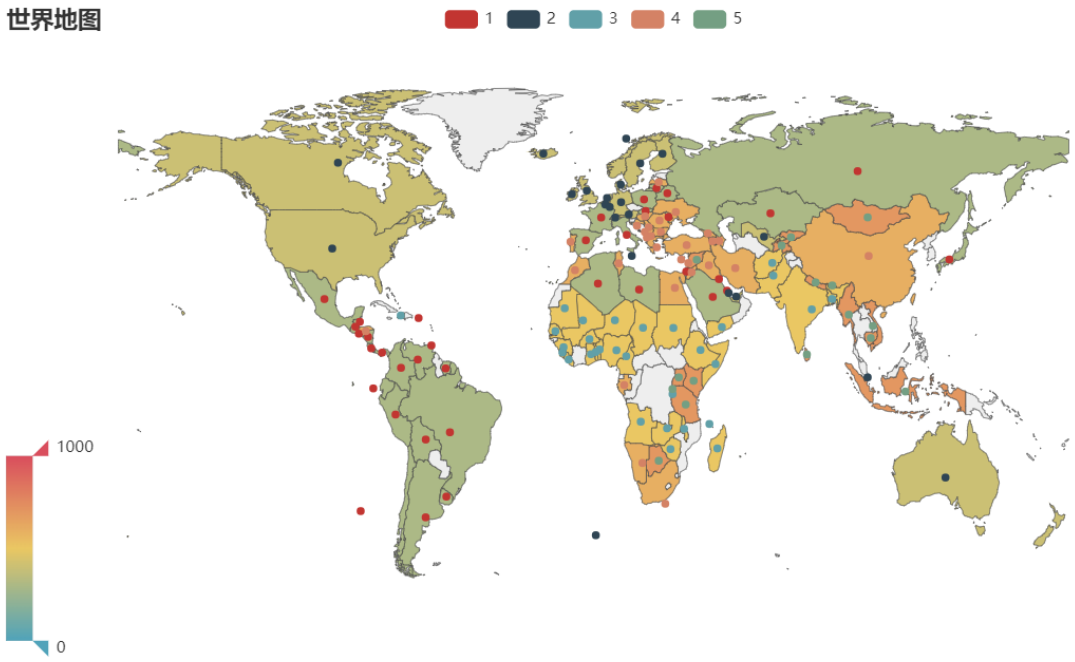


Figure 23: World map distribution

From Figure 23, We can intuitively see the distribution of the five clustering results around the world. Through observation, we find that some countries in North America, Europe and Oceania have the highest GDP level in the world, and their people's happiness index is relatively high, and their freedom and trust are also high. South America and some Asian countries followed. This clustering

result coincides with our previous analysis. The accuracy of K-means model is verified to some extent. And its applicability on this dataset.

4.2 Prediction

In the following, we will use the time series forecasting model to predict the GDP of each country, and then judge the impact of the COVID-19 epidemic on the GDP of China and the United States, and compare the results.

4.2.1 Model building and training

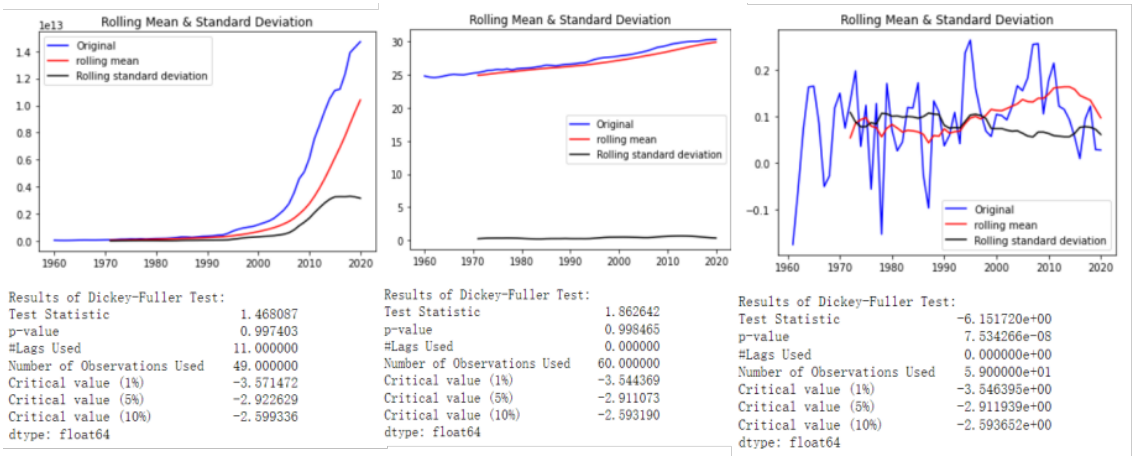


Figure 24: Original,log,diff

First, we process the data as time series data, and then test the stationarity of the data, we use the unit root test. Both the original series and the log series are tested, and the p-values are too large to reject the null hypothesis. So let's do a difference and check. At this point, the value of the p-value is found to be less than 0.05, and the null hypothesis can be rejected and the series is considered stationary. Next, we set the order and fit the model. We find the optimal order (1,0) by brute-force method. The model is then fitted. Then we carry out residual white noise test. If the residual is white noise sequence, it means that the useful information in the time series has been extracted, and the rest is all random disturbance. We can consider the residual sequence as normal distribution through the qq plot of the white noise sequence. Then we test the correlation of the residual series, using the Durbin-Watson test (D-W test), we can get that the DW values for China and the United States are 2.10 and 1.69, respectively. Finally, we forecast the GDP of China and the United States for 5 years after 2020 based on the smoothed series.

4.2.2 Model checking and comparative analysis

We calculate the prediction residuals and find that most of them are within 5% in absolute terms, with the errors getting smaller as the years go by. This indicates that the model has some accuracy.

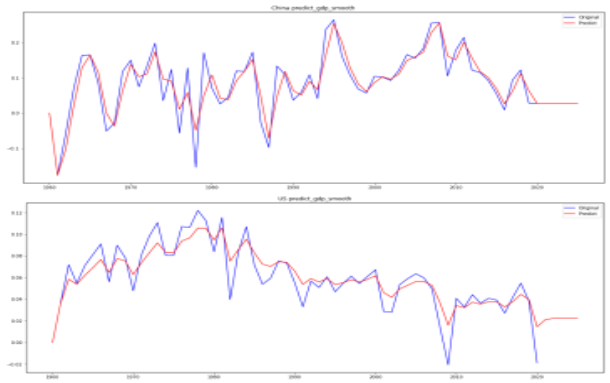


Figure 25: Time series prediction after smoothing

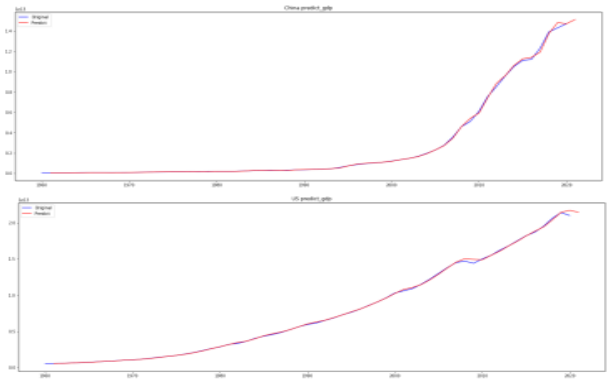


Figure 26: Original time series prediction

From Figure 25, If we look at the smoothed time series forecasts for China and the US, we can see that the fit is OK. We do the reduction and plot the original time series forecasts for comparison. The reduction can only forecast GDP figures for the year after 2020. From Figure 26, Comparing the real data and predicted data of China and the United States, it can be seen that the fitting situation is good. Through the forecast, we can see that although the COVID-19 epidemic broke out first in China, China's handling efficiency and speed are very fast, and the forecast results show that China's GDP will still maintain an upward trend in the next year. It can be seen that China has grown into a mature power capable of shouldering important responsibilities in the world. The GDP time series forecast for the US, on the other hand, shows a decline in GDP in the coming year after 2020. As a world power, the United States has yet to improve its ability to deal with global health security and contain the economic impact of COVID-19 in a timely manner.

5 Conclusion

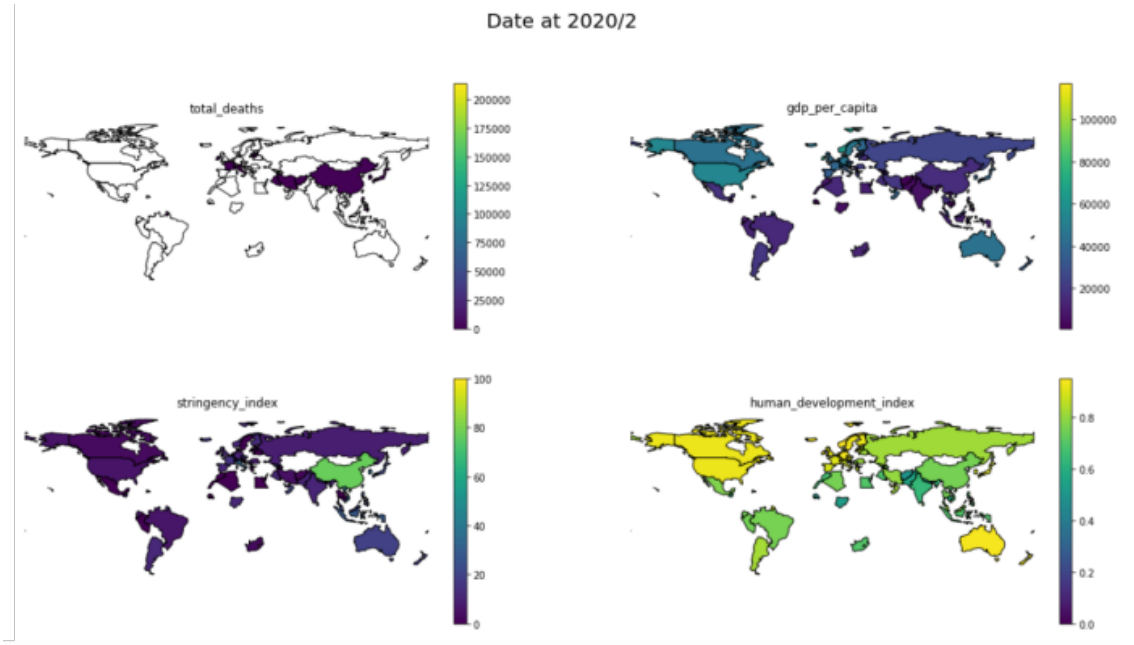
I just want to quote a sentence in the conclusion: But the pestis forced them to live a static life, confining their activities to some dreary part of the city, and leaving them to seek solace in their thoughts day after day –Albert Camus (*pestis)*

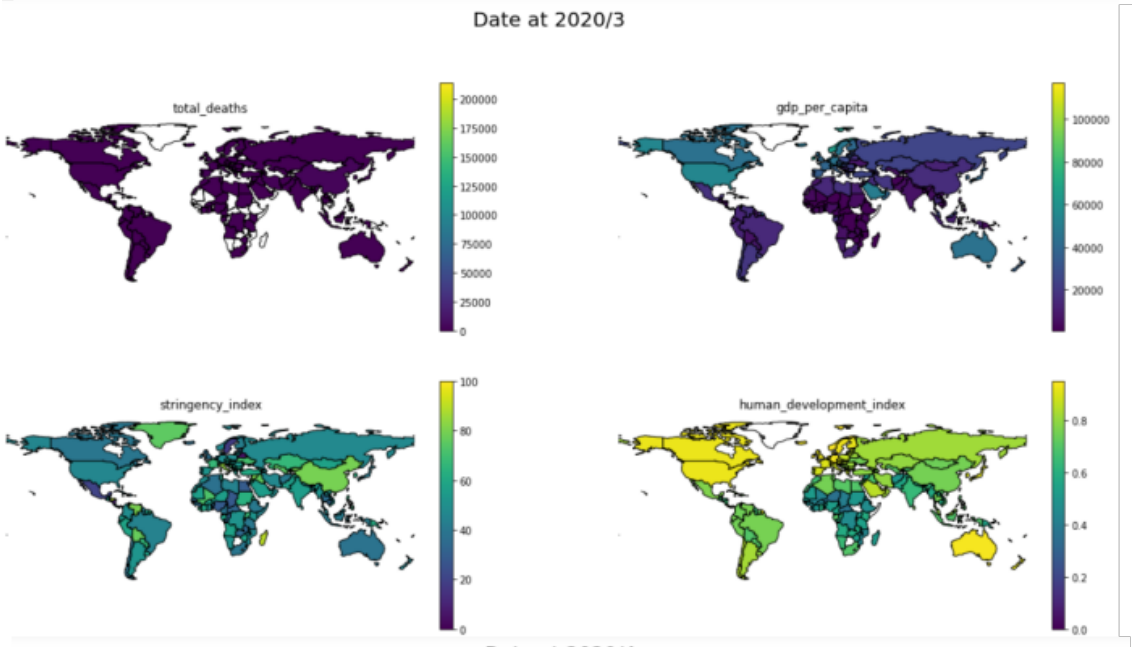
6 Appendix

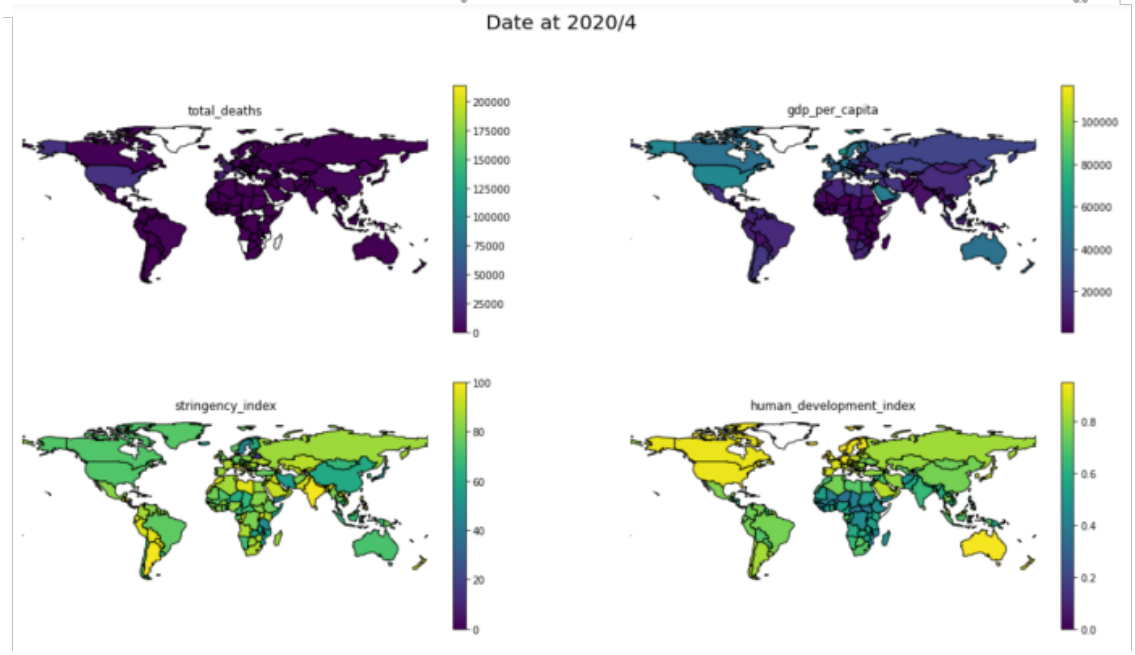
6.1 Source of dataset and countrys‘ indexs at date

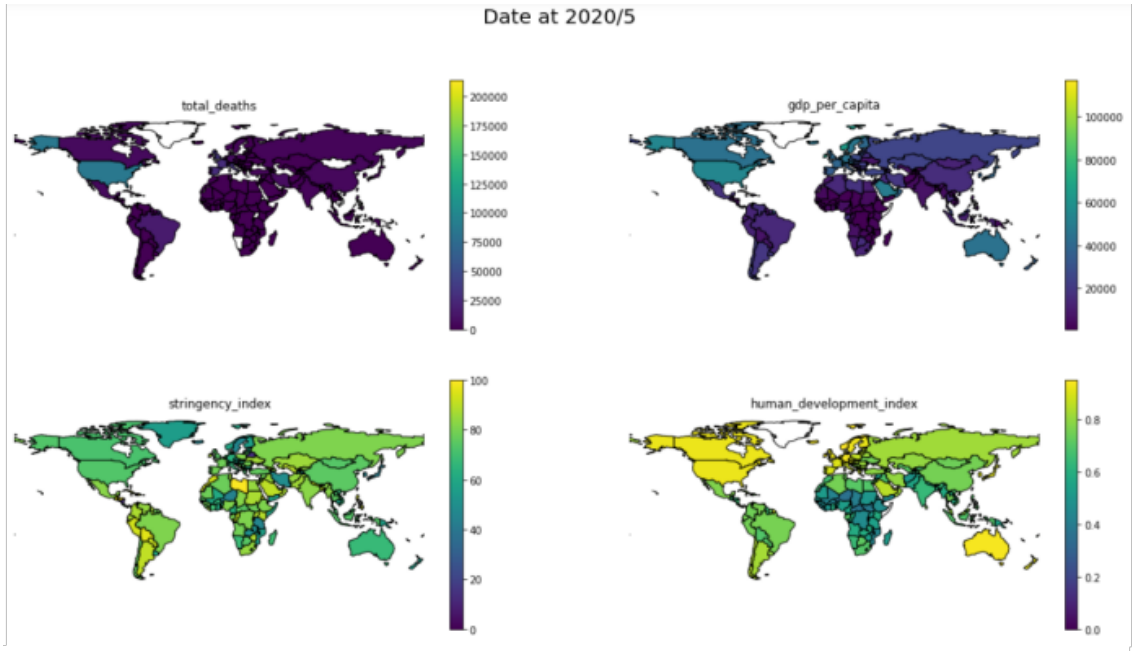
1.Vitenu-Sackey, Prince Asare (2020), "The Impact of Covid-19 Pandemic on the Global Economy: Emphasis on Poverty Alleviation and Economic Growth", Mendeley Data, V1, doi: 10.17632/b2wvnbnpj9.1

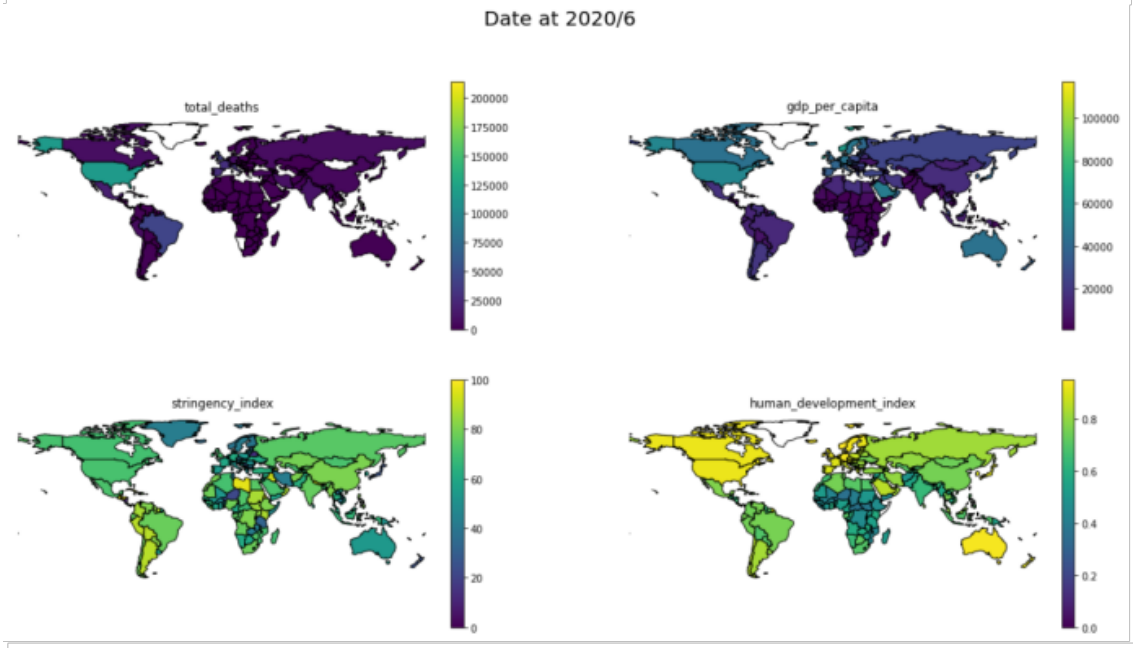
2.https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data/vbim-akqf

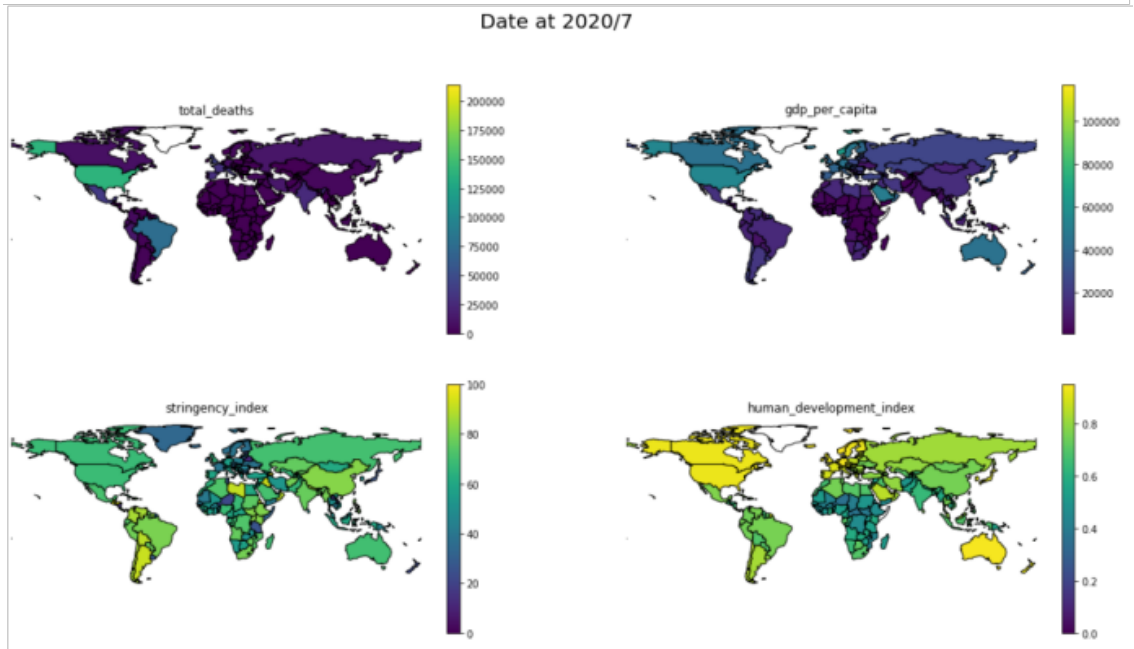

Date at 2020/2
total_deaths
gdp_per_capita
stringency_index
human_development_index

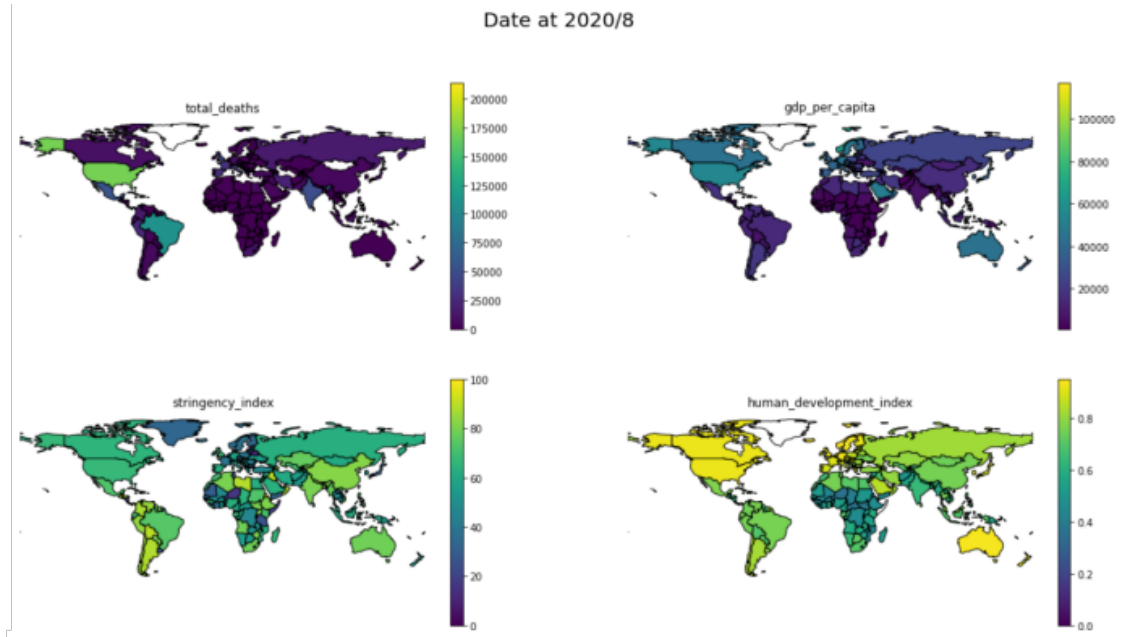

Date at 2020/3
total_deaths
gdp_per_capita
stringency_index
human_development_index

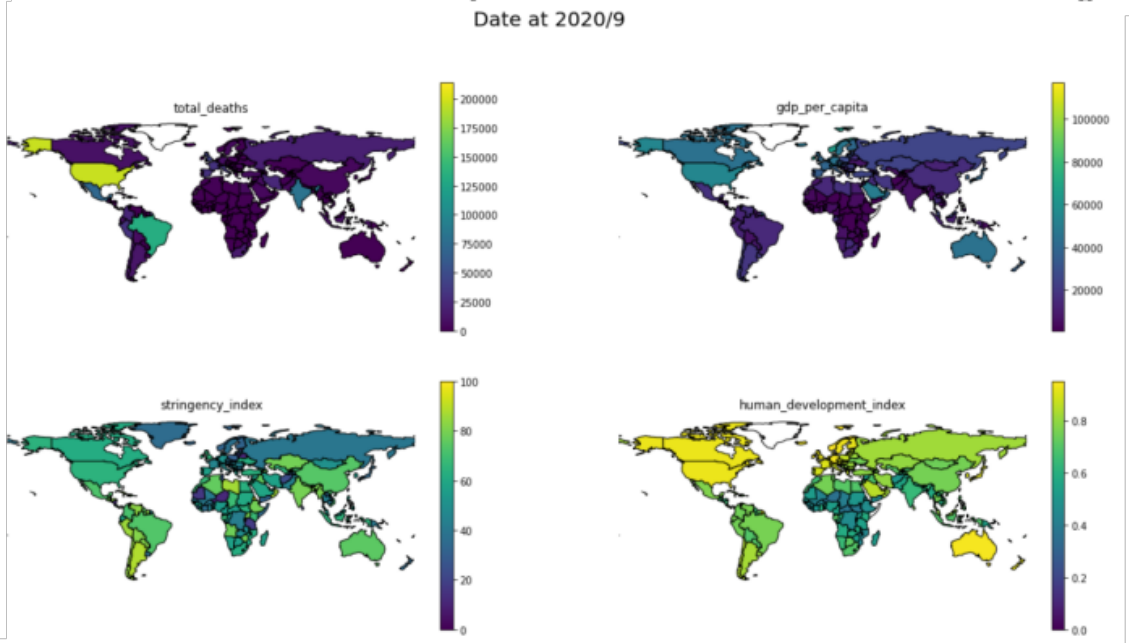

Date at 2020/4
total_deaths
gdp_per_capita
stringency_index
human_development_index

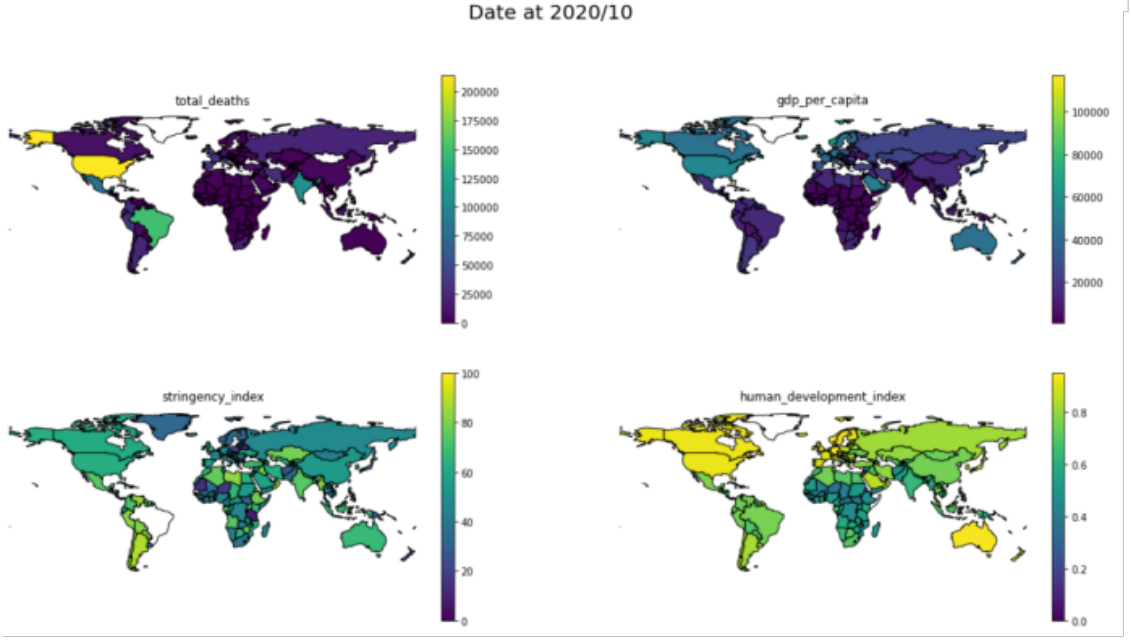

Date at 2020/5
total_deaths
gdp_per_capita
stringency_index
human_development_index


Date at 2020/6
total_deaths
gdp_per_capita
stringency_index
human_development_index


Date at 2020/7
total_deaths
gdp_per_capita
stringency_index
human_development_index


Date at 2020/8
total_deaths
gdp_per_capita
stringency_index
human_development_index


Date at 2020/9
total_deaths
gdp_per_capita
stringency_index
human_development_index


Date at 2020/10
total_deaths
gdp_per_capita
stringency_index
human_development_index

References

[1] Jin Chenxia,Li Fachao,Ma Shijie Wang Ying.(2022).Sampling scheme-based classification rule mining method using decision tree in big data environment. Knowledge-Based Systems. doi:10.1016/J.KNOSYS.2022.108522.

[2] Shanbehzadeh Mostafa,Nopour Raoof Kazemi Arpanahi Hadi.(2022).Using decision tree algorithms for estimating ICU admission of COVID-19 patients. Informatics in Medicine Unlocked(prepublish). doi:10.1016/J.IMU.2022.100919.

[3] Heldner Mirjam R.,Chalfine Caroline,Houot Marion,Umarova Roza M.,Rosner Jan,Lippert Julian... Rosso Charlotte.(2022).Cognitive Status Predicts Return to Functional Independence After Minor Stroke: A Decision Tree Analysis . Frontiers in Neurology. doi:10.3389/FNEUR.2022.833020.

[4] Jiang Yan.(2022).Using Decision Tree Classification and AdaBoost Classification to Build the Abnormal Data Monitoring System of Financial Accounting in Colleges and Universities. Computational Intelligence and Neuroscience. doi:10.1155/2022/1467195.

[5] Gao Xue Yao,Li Kai Peng,Zhang Chun Xiang Yu Bo.(2021).3D Model Classification Based on Bayesian Classifier with AdaBoost. Discrete Dynamics in Nature and Society. doi:10.1155/2021/2154762.

[6] S. Suganya,T. Meyyappan S. Santhosh Kumar.(2020).Performance Analysis of KMeans and KMediods Algorithms in Air Pollution Prediction. International Journal of Recent Technology and Engineering (IJRTE)(5).

[7] G.G Gokilam K Saravanan.(2017).Improved Optimization centroid in modified Kmeans cluster. International Journal of Engineering and Technology(2). doi:10.21817/ijet/2017/v9i2/170902224.

[8] Franses Philip Hans.(2021).Inclusion of older annual data into time series models for recent quarterly data. Applied Economics Letters(19). doi:10.1080/13504851.2020.1866152.

[9] Palladino R,Affinito G Triassi M.(2021).The impact of COVID-19 on School of Medicine students' performance: an interrupted time series study. European Journal of Public Health(Supplement3). doi:10.1093/EURPUB/CKAB165.103.

[10] Sekadakis Marios,Katrakazas Christos,Michelaraki Eva,Kehagia Fotini Yannis George.(2021).Analysis of the impact of COVID-19 on collisions, fatalities and injuries using time series forecasting: The case of Greece. Accident Analysis and Prevention. doi:10.1016/J.AAP.2021.106391.