

Analysis And Prediction of COVID-19 Based on Machine Learning

Ye Xia*, Peiyu Zhu and Zhe Zhou

Beijing Royal school, Beijing, China

* Corresponding Author Email: xiaye@brs.edu.cn

Abstract. Artificial intelligence (AI) technology is in high demand today for wireless infrastructure integration, real-time data collecting, and end-user device processing. The most innovative answer at hand right now is the use of AI to recognize and forecast widespread epidemics. The pandemic, which started with COVID-19, has had a terrible impact on world society and placed a significant load on the most developed healthcare systems globally. According to the European Centre for Disease Prevention and Control (ECDC), as of May 11, 2020, there had been 282,244 fatalities and more than 4,063,525 confirmed cases. However, given the current exponential and rapid development in the number of patients, it is essential to make use of AI technology to quickly and accurately estimate the prognosis of infected individuals.

Keywords: description, Outlier handling, Missing value handling, Evaluation index calculation.

1. Introduction

In order to anticipate the number of instances, this research presents a prediction model based on random forest that combines numerous feature data from COVID-19 patients in each area. There are five sections to the complete article. Prior to pre-processing, it examined the COVID-19 data set, looked at the different feature data types, as well as the distribution of missing values and outliers. The data were then put through a descriptive analysis, many models based on indiscriminate forest were built, and finally the models were evaluated and compared. On the used data set, this model has an accuracy rate of greater than 99.9%. Data research has shown that country differences exist in the prediction power of instances.

2. Data Pre-processing

2.1. Data description

Our World in Data, an open data platform, provided the research data for this article. (<https://github.com/owid/covid-19-data/tree/master/public/data>) has 206150 entries in 67 fields. Each record includes pertinent data for the total number of cases, new cases, and other factors in the following nation at a certain moment. Based on the example of time series prediction model of three machines learning algorithms, our goal is to use 20 of these fields to forecast the entire number of instances. Using the jupyter notebook and Python 3.8, the full analysis and modeling process is completed.

2.2. Outlier handling

Anomaly identification, sometimes referred to as outlier detection, is the action of locating unexpected occurrences, observations, or objects that deviate noticeably from average values. Unsupervised anomaly detection, which is founded on two fundamental presumptions, is a method used by data scientists to generally apply any sort of anomaly detection to unlabeled data. Therefore, from a commercial standpoint, it is crucial to spot real abnormalities rather than false positives or cluttered data.

Unsupervised, semi-supervised, and supervised anomaly detection systems are the three basic categories. The right anomaly detection technique essentially depends on the tags that are present in

the data collection. In this study, an unsupervised outlier management approach is used to identify outliers by examining the data's box-plot distribution. Only part of the continuous variables are displayed above, while the data for the box-plot are presented below. Descriptive statistics are shown in Table 2-2:

Table 2-1 Descriptive statistics for continuous variables

	Total_Cases	New_Cases	Total_Deaths
count	3.684000e+03	3677.000000	3618.000000
mean	6.203561e+06	16904.998368	79587.428137
std	1.207615e+07	45857.089216	152481.839034
min	1.000000e+00	0.000000	1.000000
25%	9.454875e+04	114.000000	4634.000000
50%	6.039745e+05	1659.000000	9904.500000
75%	3.696396e+06	12546.000000	53647.500000
max	4.408704e+07	414188.000000	526530.000000

The above table contains some statistical details about the continuous variables, such as the count, mean information, standard deviation, minimum and maximum values, and the three quartiles.

2.3. Missing value handling

Based on the data itself, this study determines how to handle missing values. Missing values are removed from the data-set and replaced with 0 in order to shorten training time and increase accuracy.

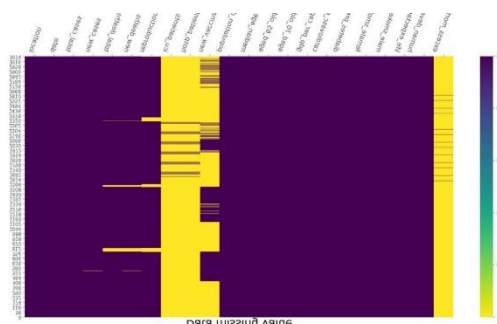


Figure 2-1 Feature missing in the dataset

The heat map shows that only the characteristics total deaths, new deaths, reproduction rate, icu patients, hosp patients, new vaccinations, and excess mortality have missing values among all the features. In terms of these attributes, this paper has substituted a value of 0 for them.

3. Descriptive Statistical Analysis

An investigation strategy called elucidating investigation is primarily one that breaks down the material acquired and identifies several quantitative highlights that depict objective anomalies. It includes information discrete examination, information recurrence conveyance inquiry, information focal propensity investigation, and so forth. Additional analysis of the data is required due to the graphic research.

3.1. Regional data distribution

Box-plot analysis of the total number of COVID-19 cases in different countries, the results are shown in Figure 3-1:

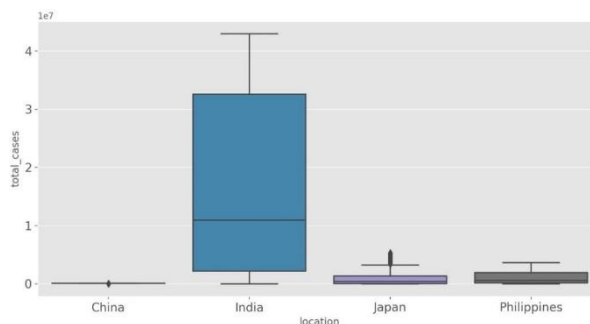


Figure 3-1 Total number of COVID-19 cases analyzed in different countries

Next, look at the distribution of the number of new cases in different countries, using the box-plot chart for statistical description, as shown in Figure 3-1:

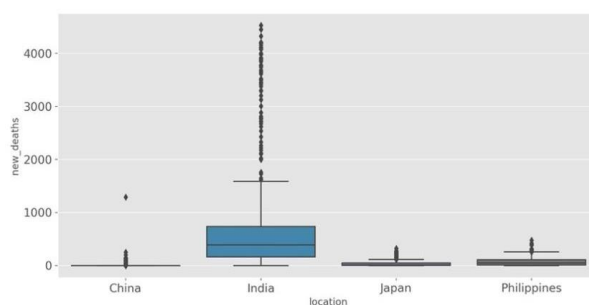


Figure 3-2 Distribution of the number of new cases in different countries

From the two figures above, it is clear that the number of cases in India is extremely widespread, indicating that it varies widely. Correspondingly, in the distribution of new cases, it can also be seen that there is a period of sharp increase in COVID-19 cases in India.

3.2. Time series data distribution

To further study the data distribution on the time series, the time data is created next. Time series analysis is a statistical method for processing trend analysis and time series data. It adheres to periodic time intervals and is measured or gathered at predetermined times. In other words, a time series is a collection of data points arranged chronologically, and time series analysis is the method used to comprehend that information. In this paper, first of all, the annual and quarterly case statistics for China, India, Japan, and the Philippines are performed, as shown in Figures 3-3 and 3-4.

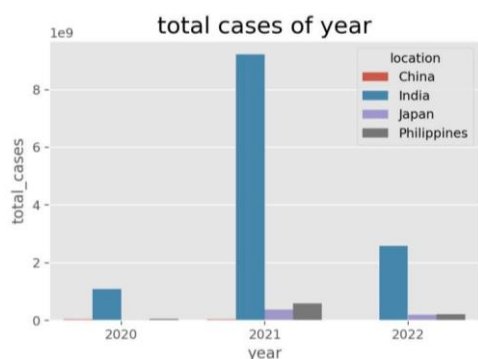


Figure 3-3 Total number of cases by year

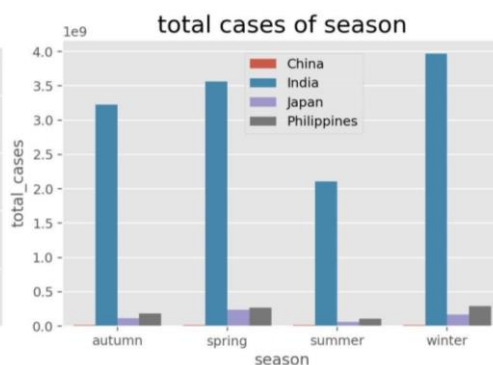


Figure 3-4 Total number of cases by quarter

The outbreak in 2021 will be the worst, as seen by the two numbers above, and it will be most prevalent in the spring and winter months. The overall number of instances in the four nations is then the subject of daily and monthly investigation.

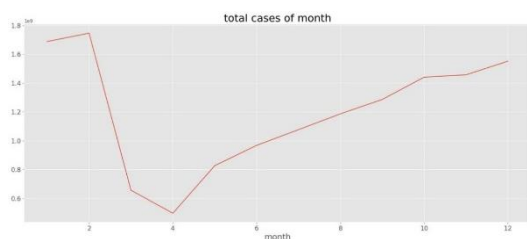


Figure 3-5 Total number of cases by month

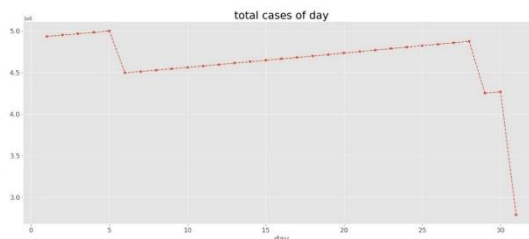


Figure 3-6 Total number of cases by day

The four figures above can be analyzed to determine that the epidemic incidence was high in 2021. The epidemic situation is more severe in the spring and winter seasons, peaking at the beginning of the year, declining to its lowest point around April after control, and then gradually increasing later. Every month, the epidemic situation gets slightly better until it is at its lowest point at the end of the month. Additionally.

3.3. Feature correlation analysis

In the feature selection stage of data pre-processing, data and feature correlation is regarded as a crucial step, particularly when the continuous data type of the feature. Understanding the connection between various variables and properties in a dataset is possible through data correlation. The reason why one or more traits depend on another or more attributes, as well as the link between one or more attributes and other attributes, may be discovered via correlation.

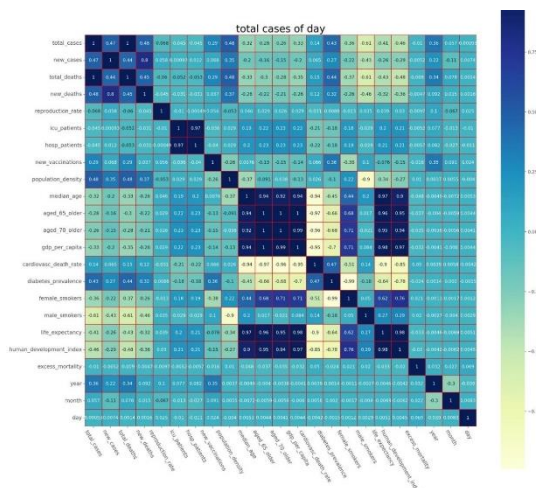


Figure 3-7 Distribution of total cases per day in a week

As shown in the figure above, variables in the black blocks have a strong positive correlation, and variables in the yellow blocks have a strong negative correlation.

3.4. Feature descriptive research

A total of 20 features have been studied in this dataset, and the following is a detailed exploration and analysis of each feature studied.

Total number of confirmed cases of COVID-19, including the reported probable cases. Confirmed cases of COVID-19, including reported possible cases. In rare cases, if the source reports a negative daily change due to data correction, the index is set to NA.

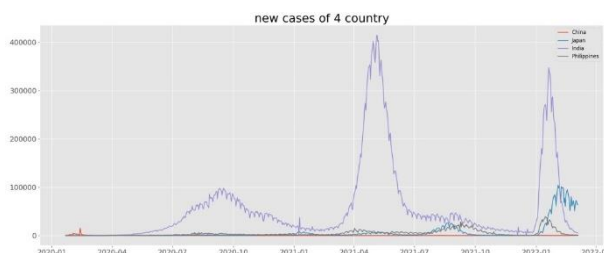


Figure 3-8 Distribution of the total number of new cases in the 4 countries over time

Total number of deaths caused by COVID-19, including reported possible deaths.

The number of new deaths due to COVID-19, including reported possible deaths. In rare cases, if there is a daily negative change due to data correction, the index is set to NA.

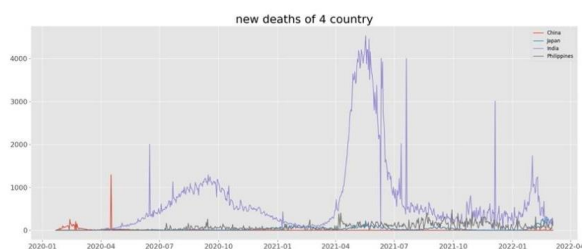


Figure 3-9 Distribution of the number of new deaths in the 4 countries over time

Real-time estimation of the effective Reproduction rate (R) of COVID-19.

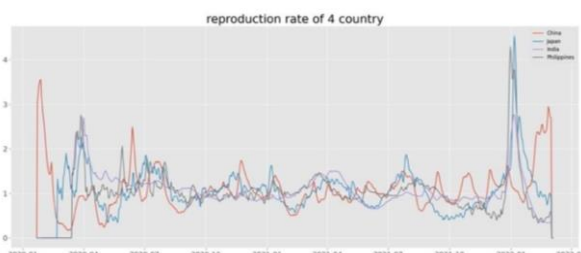


Figure 3-10 Distribution of effective reproduction rates for COVID-19 in the 4 countries over

Number of people with COVID-19 in the Intensive Care Unit (ICU) on a given day.

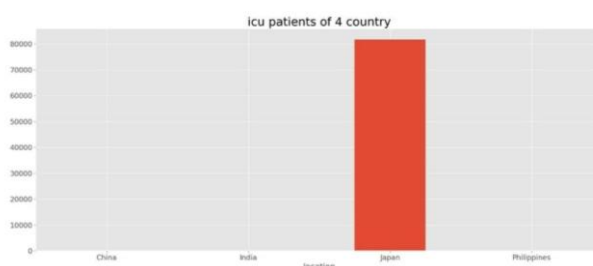


Figure 3-11 Number of COVID-19 patients in ICUs in the 4 countries

Number of COVID-19 patients hospitalized on a given day.

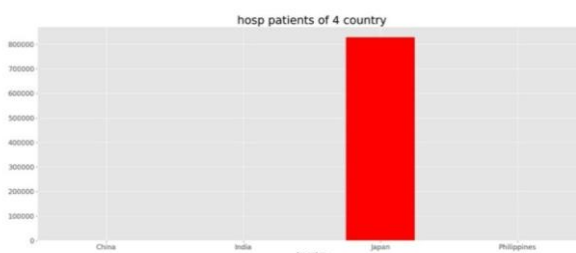


Figure 3-12 Number of hospitalized COVID-19 patients in the 4 countries

Newly administered COVID-19 vaccination dose (counted by consecutive days only)

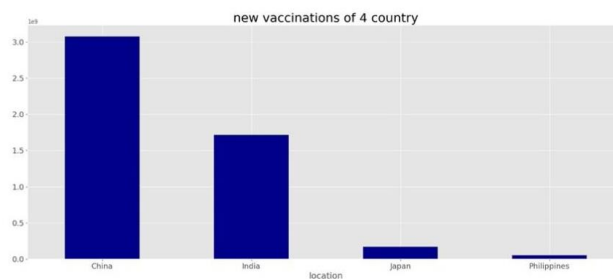


Figure 3-13 The number of newly vaccinated vaccines in the 4 countries

The number of people divided by land area, measured in square kilometres and based on the data of the most recent year.

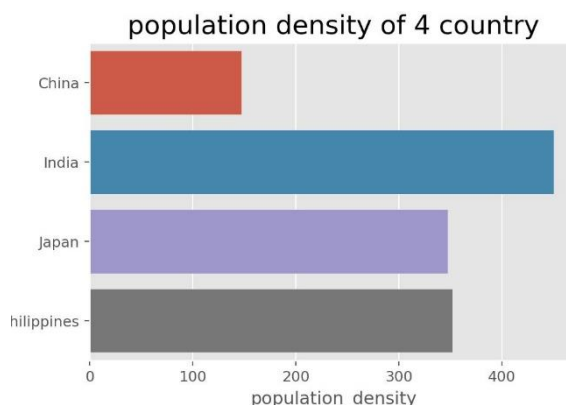


Figure 3-14 Population density of the 4 countries

Median population age, United Nations projections for 2020.

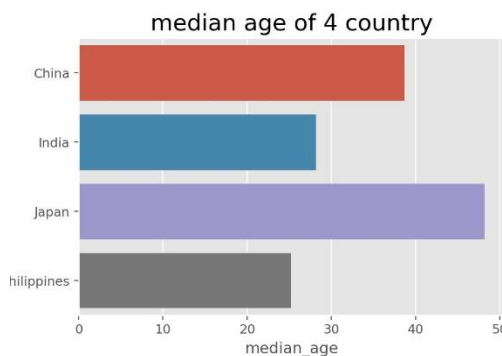


Figure 3-15 Median age of the population of the 4 countries

The proportion of the population aged 65 years and over, based on the data of the most recent year.

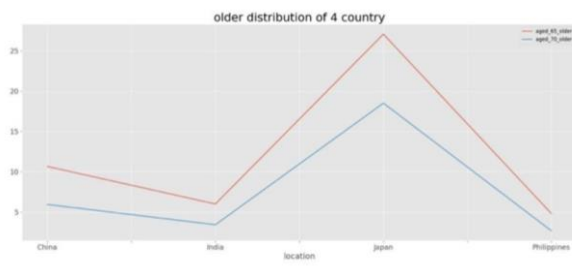


Figure 3-16 Proportion of the population aged 65 and over in the 4 countries

GDP at purchase power parity (unchanged for 2011 in the international dollar), based on the data of the most recent year.

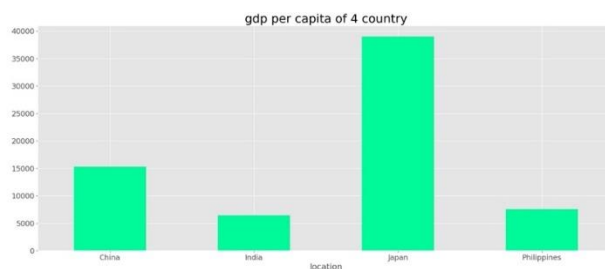


Figure 3-17 GDP of the 4 countries

Cardiovascular mortality rate in 2017 (deaths per 100,000 people per year)

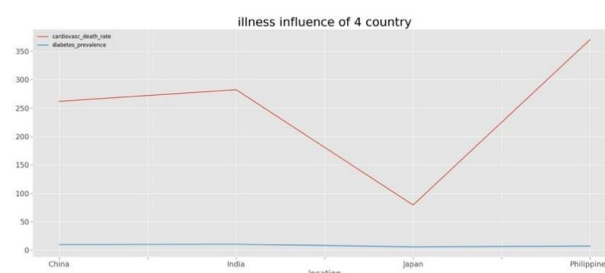


Figure 3-18 Disease rates in the 4 countries

The proportion of male/female smokers, based on the data of the most recent year.

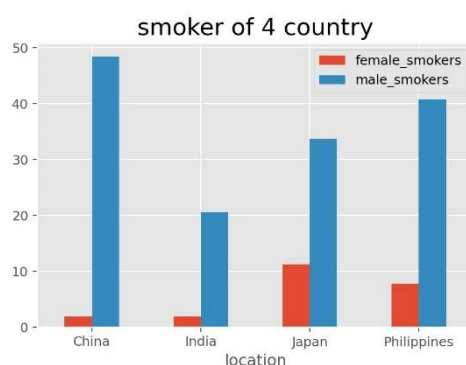


Figure 3-19 Proportion of men and women who smoke in the 4 countries

Life expectancy at birth in 2019.

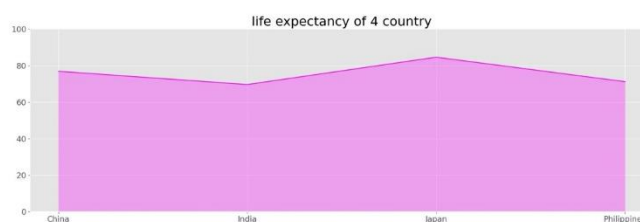


Figure 3-20 Life expectancy at birth in the 4 countries in 2019

A composite index of average achievement in three basic areas of human development - longevity and healthy living, knowledge and a decent standard of living. Based on the data of 2019.

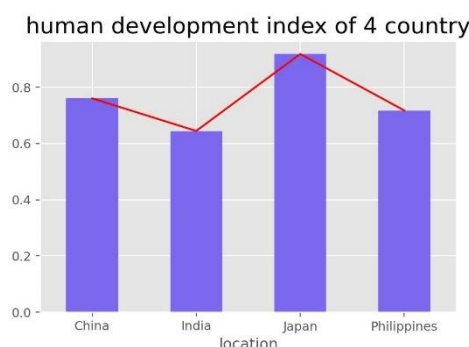


Figure 3-21 Human Development Index of the 4 countries

The percentage difference between the number of weekly or monthly deaths reported in 2020-2021 and the projected number of deaths based on the same period in previous years.

Table 3-1 Percentage differences in deaths in the 4 countries during the same period

Location	Excess Mortality
China	0.000000
India	0.000000
Japan	-0.012506
Philippines	0.496880

4. Build a COVID-19 Case Prediction Model

4.1. Split dataset

To train any machine learning model, regardless of the type of dataset used, the dataset must be divided into training data and test data. In this paper, it is split using the "train_test_split" library of sklearn. Here in this paper, "train_test_split" is used to split the data at an 80:20 ratio, i.e. 80% of the data will be used to train the model, while 20% of the data will be used to test the model built by it. Setting the random_state to a fixed value guarantees that the same sequence of random numbers will be generated each time the code is run.

4.2. One-hot encoding

One-hot coding is the process of converting categorical variables into pre-processing that can be provided to ML algorithms for better predictions.

The numerical value of the element in the dataset is represented by the category value. This essay completes the categorical value assignment using sklearn's LabelEncoder. The issue with Label Encoder is that it makes the assumption that a category is better the higher the categorical value. The foundation of this organizational structure is guided learning based on categorical values. The model's predictions will contain a large number of mistakes if it internally calculates the average. The model's accuracy may be significantly increased by creating a category of "thresholding" and using it as a feature to train the model.

4.3. Build the Random Forest model

During the training phase, we used the grid search strategy to find the best hyper-parameters in order to optimize the model. In this study, the total number of instances is predicted using the Random Forest model. The major focus of this study is the optimization of the Random Forest model in Python using Scikit-Learn. Despite being data-based, the model optimization in this work is completely independent of the data. Since parameter tuning is more dependent on experimental outcomes than theory, it is preferable to explore a variety of combinations and gauge how each model performs before settling on the ideal settings. But basing model evaluation only on training data can result in overfitting, one of the core issues with machine learning. Our model will perform well on the training

set but be unable to generalize to the new data if we optimize the model for the training data. We average each fold's performance at the conclusion of training to get the model's final validation measure.

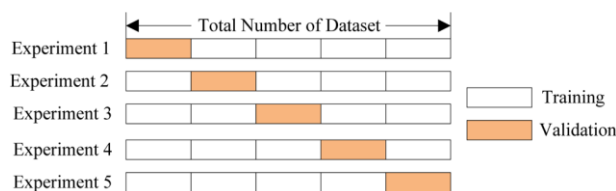


Figure 4-1 5-fold cross validation

This research executes numerous rounds of the full K-Fold CV process, utilizing a new model configuration each time, for hyper-parameter optimization. The best model is then chosen after comparisons of all the models, and it is trained on a complete training set and assessed on the test set.

4.4. Model based on Random Forest

4.4.1 Traditional Random Forest regression

Numerous classification techniques, including Logistic Regression, Support Vector Machine, Naive Bayes classifier, and Decision Tree, are available thanks to data science. But in terms of proximity to the top of the classifier hierarchy, the Random Forest classifier and regressor are the best.

Other significant options for the RandomForestRegressor include min samples split, min samples leaf, n jobs, and others that may be read in the documentation provided by sklearn. In order to obtain a more optimal collection of parameters, we first utilize grid search and cross validation to identify the fundamental values that go into the Random Forest regression model. After testing, our dataset performs best under the parameters in the following figure:

```
RandomForestRegressor(max_depth=15, n_estimators=600)
```

Figure 4-2 Random Forest best hyper-parameters

4.4.2 GBDT model

Gradient Boosting Decision Tree (GBDT) is a type of decision tree boosting technique that is optimized from three perspectives: the method itself, the algorithm's operational efficiency, and the algorithm's resilience.

After testing, it performs best under the parameters of the following figure:

```
GradientBoostingRegressor(max_depth=5, n_estimators=600)
```

Figure 4-3 GBDT best hyper-parameters

4.4.3 Extra Trees

Extra Trees and Random Forest are two very similar methods. Both approaches have many characteristics. Both include a sizable number of decision trees, with each tree's predictions factored into the final choice. The choice of a cut point to separate the node is another distinction. Extra Trees makes a random selection, whereas Random Forest selects the optimal split. However, after choosing the split point, the two algorithms determine the best point out of all the feature subsets. As a consequence, Extra Trees keeps the optimizations while also adding randomness. The Extra Trees approach is quicker when it comes to calculating expenses and execution durations because the entire process is the same and the split point is chosen at random rather than by computing the ideal point.

5. Model Evaluation

5.1. Feature importance analysis

Feature importance is useful, the reasons are as follows: Data understanding, Model improvements and Model interpret-ability.

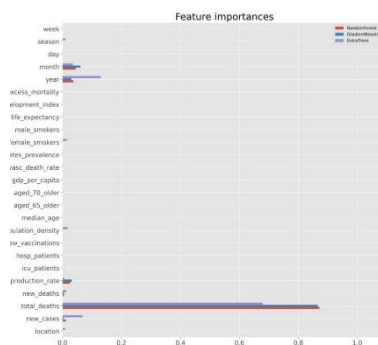


Figure 5-1 Overall importance of each feature except for the 3 countries except India

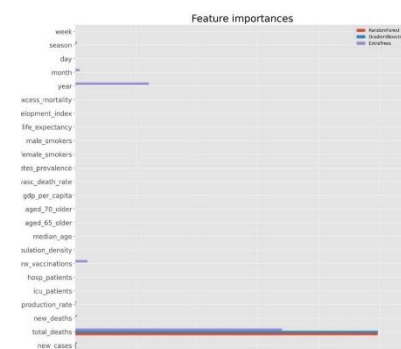


Figure 5-2 The importance of the features of India to the model

5.2. Evaluation index calculation

In this paper, the model is evaluated by calculating the accuracy. Accuracy represents the percentage of the total number of cases that are correctly predicted, and the formula for its calculation is as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (1)$$

Due to the differences in data between India and the other 3 countries, taking into account that India might also have differences in accuracy performance, this paper has carried out analysis on India separately. The following figures show the accuracy performance of 3 countries except India and the 3 models of India.

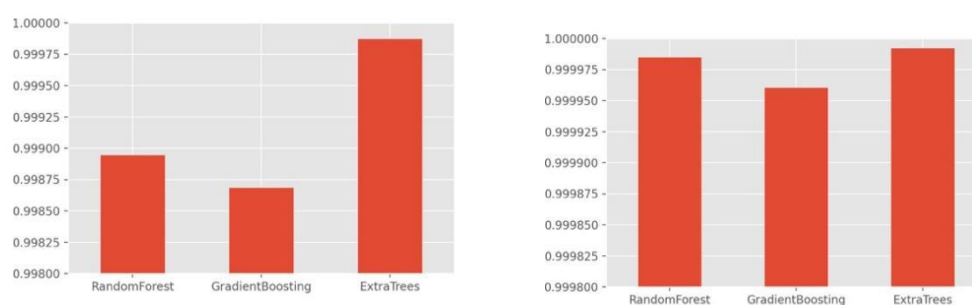


Figure 5-3 Accuracy performance of the model

6. Conclusion

A powerful prediction tool is Random Forest. The Law of Large Numbers prevents them from over-fitting. Adding the right amount of randomness to make them reliable classifiers and regression models. The framework also offers insight into the prediction powers of Random Forest based on the strength of each individual predictor variable and their association. The theoretical values of strength and correlation can be materialized using the out-of-bag estimation.

Random Forest does not progressively alter the training set; instead, it uses boosting and adaptation bags to provide results that are competitive. Their effectiveness implies that their function is to lessen prejudice. It's unclear how it works. Bayesian process is another name for Random Forest.

Random features and inputs yield strong classification results but poor regression outcomes. Bagging and random feature are the two types of randomization that were employed in this study. It's conceivable that different kinds of injection randomization might have more beneficial outcomes.

In this study, the dataset is first pre-processed and descriptively examined. The grid search method is then utilized to assess the performance of the Random Forest, GBDT, and Extra Trees algorithms.

The Extra Trees approach outperformed the other two in the comparisons, although it is important to note that the dataset's prediction accuracy differs dramatically between India and the other 3 nations. In conclusion, this research offers a fresh approach for forecasting the number of epidemic cases that will occur in a certain region in the future.

References

- [1] Samuel, J., Ali, G.G., Rahman, M., Esawi, E. and Samuel, Y. (2020) Covid-19 Public Sentiment Insights and Machine Learning for Tweets Classification. *Information*, 11, 314.<https://doi.org/10.3390/info11060314>
- [2] Ivanov, D. (2020) Predicting the Impacts of Epidemic Outbreaks on Global Supply Chains: A Simulation-Based Analysis on the Coronavirus Outbreak (COVID-19/ SARS-CoV-2) Case. *Transportation Research Part E: Logistics and Transportation Review*, 136, Article ID: 101922.<https://doi.org/10.1016/j.tre.2020.101922>
- [3] Zhan, C., Tse, C.K., et al. (2020) Prediction of COVID-19 Spreading Profiles in South Korea, Italy and Iran by Data-Driven Coding.<https://www.medrxiv.org/content/10.1101/2020.03.08.20032847v1>
- [4] Zheng, Z., Wu, K., Yao, Z., et al. (2020) The Prediction for Development of COVID-19 in Global Major Epidemic Areas through Epidemical Trends in China by Utilizing State Transition Matrix Model.<https://www.medrxiv.org/content/10.1101/2020.03.10.20033670v1>
- [5] Williamson, E.J., et al. (2020) Factors Associated with COVID-19-Related Death Using OpenSAFELY. *Nature*, 584, 430-436.
- [6] Huang, C., et al. (2020) Clinical Features of Patients Infected with 2019 Novel Coronavirus in Wuhan, China. *The Lancet*, 395, 497-506.[https://doi.org/10.1016/S0140-6736\(20\)30183-5](https://doi.org/10.1016/S0140-6736(20)30183-5)
- [7] Guan, W.J., et al. (2020) Comorbidity and Its Impact on 1590 Patients with COVID-19 in China: A Nationwide Analysis. *European Respiratory Journal*, 55, Article ID: 2000547.<https://doi.org/10.1183/13993003.01227-2020>
- [8] Li, X., et al. (2020) Impact of Cardiovascular Disease and Cardiac Injury on In-Hospital Mortality in Patients with COVID-19: A Systematic Review and Meta-Analysis. *Heart*, 106, 1142-1147.<https://doi.org/10.1136/heartjnl-2020-317062>
- [9] Centers for Disease Control and Prevention (2020) National Diabetes Statistics Report. US Department of Health and Human Services.<https://www.cdc.gov/diabetes/data/statistics-report/index.html>
- [10] Casqueiro, J., Casqueiro, J. and Alves, C. (2012) Infections in Patients with Diabetes Mellitus: A Review of Pathogenesis. *Indian Journal of Endocrinology and Metabolism*, 16, S27.<https://doi.org/10.4103/2230-8210.94253>