A data analytics approach for COVID-19 spread and end prediction (with a case study in India)

## **Abstract**

World is now experiencing the new pandemic caused by COVID-19 virus and all countries are affected by this disease. Due to socio-economic problems of this disease, it is required to predict the trend of the outbreak and propose a beneficial method to find out the correct trend. In this paper, we compiled a dataset including the number of confirmed cases, the daily number of death cases and the number of recovered cases. Furthermore, by combining case number variables like behaviour and policies that are changing over time and machine-learning (ML) algorithms such as logistic function using inflection point, we created new rates such as weekly death rate, life rate and new approaches to mortality rate and recovery rate.

### **Data collection and preparation**

Variables in our research include number of newly found cases, new death cases, and newly found recovered cases. All information was collected and categorized from reputable sources such as WHO (World Health Organization). These variables are chosen to utilize in our prediction methods due to their numerical nature.

The high prevalence rate of COVID-19 and the need for estimated calculations necessitate collecting the required data sets from reliable sources including WHO and world meter. Research data including observation data are obtained from a collection of samples’ reports in three parts

Find the factors influencing the spread of the virus

A study of world conditions

Collecting the dataset

|  |
| --- |
|  |

Investigate the influential variables in forecasting

Walk-Forward Validation

Selecting India

Predicting the trend of virus infecting in India

Navie Method

Time -Series

Predicting Tools

Predicting the rates in India

Investigating the rates in India

Determination of deaths, recovered and confirmed rates

Comparing India with the World

A study of India conditions

Python software is used to implement operations such as data cleaning, data manipulation, and machine-learning algorithms. Some operations such as cleaning data, joining, summarizing, defining functions, building, and changing data frames are applied. “Numpy”, “Pandas”, “datetime”, “sklearn”, “math”, “seaborn”, “matplotlib”, “statsmodels” and “scipy” packages are utilized in this analysis.

#### **Descriptive analytics**

Data analytics tools are categorized in three classes: descriptive analytics, predictive analytics, and prescriptive analytics Descriptive analytics, extracts information from raw data appeared in business and management reports regarding sales, customers and operations. It helps organizations in grasping the reasons of the events happened in the past. Combined with simple graphical analysis, they shape the quantitative data analysis basis

In this section, descriptive analysis is done to review the position. Choosing the death rates and analysing them more accurately is another way to survey them as the main factor and make new rates. Therefore, the mortality rate which is calculated as Death/Recovered cases, is smoothing, as well as Death/Confirmed cases. This is mainly because of a drastic increase in the number of deaths at first and appropriate reactions to them in countries like Iran.

#### **Prescriptive analytics**

Prescriptive analytics phase provides studies with adaptive, automated, time-dependent, and optimal decisions. In general, prescriptive analytics is predictive analytics that prescribes one or more courses of action and shows the likelihood of outcome/influence of each action. Prescriptive analytics is purely built on the “what-if” scenarios. The main tools of prescriptive analytics are optimization, simulation, and evaluation methods. Simply, it provides advice based on predictions

Based on the literature, analysis in this section provides suggestions for future behaviours and policies with attention to rates instead of cases. Some known rates such as recovery and death are limited in usage. Thus, two new rates are considered and defined here.

Rates are the best indexes to predict the future and end of disease, some useful rates are as follows:

* The “Weekly recovered rate” is the ratio of the average weekly number of people have cured over the average weekly number of people infected with COVID-19.
* The “Weekly death rate” is the ratio of the average weekly number of dead people over the average weekly number of people have infected with COVID-19.
* The “Life rate” increase or decrease is the difference between the weekly recovered rate and the weekly death rate. It shows an increase when it is positive, while it indicates a decrease when it is negative.

#### **Predictive analytics**

Predictive analytics phase incorporates the descriptive analytics output as well as some ML algorithms and simulation techniques to build accurate models that predict the future trends. Predictive analytics assists studies in identifying future opportunities and likely risks by distinguishing specific patterns over the historical data. Some outstanding techniques which are utilized in this phase are data mining (DM), text/web/media mining, and forecasting approaches such as regression, SVM, K-Nearest neighbours and Random forest

In the following, a question arises regarding where and when COVID-19 will be eradicated.

##### **Cumulative cases prediction**

As predicting the future of outbreak is something nearly impossible, a great approach toward several each case including confirmed, death, and recovered, is enforced through using density distribution for fitting cumulative amounts of each category. Due to the cumulative nature of these numbers, the cumulative distributions under consideration should not follow the normal distribution.

**Discussion**

### This study focused on predicting the trend of COVID-19 prevalence and find the best approximate fitting pattern to predict the peak and end in different periods of time including short-term, mid-term and, long-term, and also defining and analyzing mortality rate, recovery rate, death rate, weekly rates, life rate, and infection rate. Descriptive and prescriptive analytics is done on these rates and then predictive analytics was performed with applying machine-learning algorithms. In our results, Gaussian functions are the best method to predict the outbreak since it has enough power to estimate its curve, peak and end. These analytics are needed to find out how effective were policies for prohibiting COVID-19 spread by the government including isolation or closing markets have been made to diminish commuting and traffic have been chosen.

**Challenges faced**

There are a lot of drawbacks for predicting the trend of COVID-19 disease and analyzing countries’ conditions to explain the situation exactly such as lack of information about social and political measurements and reactions to illness spread, state categorized data is not available for the public completely, many algorithms do not work in this restricted dataset due to its time-series nature.