International Journal of Science and Research (IJSR)
ISSN: 2319-7064
ResearchGate Impact Factor (2018): 0.28 | SJIF (2019): 7.583

A Data Driven Analysis of a Global Crises (COVID-19) using Machine Learning Algorithms

Achyut Tiwari

Abstract: ***At the time of writing this paper, the COVID-19 (Corona Virus) outburst has been declared a pandemic by the WHO, with confirmed cases in practically every country in the worldwide. It is the latest epidemic. It is one of the hot areas of research in the present scenario. In this paper we are going to understand the complete scenario and flow of this crisis by collecting the data and information from the secondary sources and also analyze the statistics to find out the solutions of the various questions generated in the study. We try to concentrate on COVID-19 predictions using Machine Learning Techniques. Data such as total number of person infected by this virus in India, total number of person infected by this virus in the world, number of person infected per day by the virus in the world per day, total number of deaths due to this virus, total number of deaths per day due to this virus is used as an input and then these features are modeled for analysis and prediction.***

Keywords: COVID-19, CORONAVIRUS, Machine Learning, Arima, Predictive Analysis, Epidemic, Pandemic, Fbprophet, Forecasting Model

1. Introduction

In our existence people face various health issues but recently we heard a replacement word Corona Virus (COVID-19). Until recently, the general public round the world will never have heard about the COVID-19 (corona viruses). The primary case of COVID-19 was reported in Wuhan, China and has spread to over 50 countries within the world. World Health Organization has been declared COVID-19 as a Public Health Emergency of International Concern (PHEIC) on 30 January 2020. Naturally, a rising communicable disease involves fast spreading; endangering the health of enormous numbers of individuals, and thus requires immediate actions to forestall the disease at the community level.

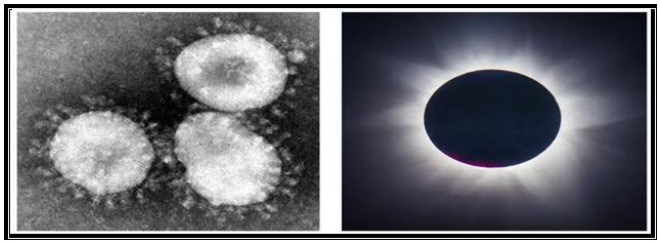
This paper aims to predict and forecast total number of cases, active cases, deaths, and recoveries through predictive modeling. The model helps to deduce patterns of outlook on disseminating related health information, and assess political and economic influence of the spread of the virus.

1.1 Historical Background of COVID-19

Who first discovered Corona Viruses?
"The opinion on Corona Viruses were first identified by a group of virologists (J D Almeida, D M Berry, C H Cunningham, D. Hamre, M. S Hofstad, L Mallucci, K McIntosh, and D A J Tyrrell), who relayed their findings in 1968 to the journal Nature."

Why are they called it as Corona Viruses?
The journal Nature reported in 1968, "these viruses are members of a previously unrecognized group which [the virologists] suggest should be called the corona viruses, to recall the characteristic appearance by which these viruses are identified within the microscope." The word "corona" has many various meanings. But it absolutely was the sun that the virologists had in mind after they chose the name corona viruses. As they wrote, they compared "the characteristic 'fringe' of projections" on the surface of the virus with the solar corona (not, as some have suggested, the points on a crown).

Figure 2 illustrates this.



[https://www.cebm.net/wp-content/uploads/2020/03/JA2.png]
Figure 2: Left: The versions of Corona Viruses;

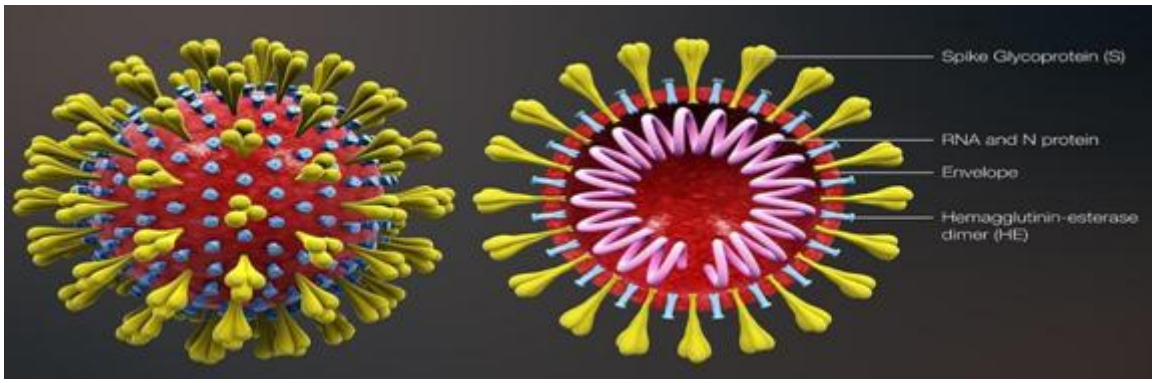
Right: The corona of the sun seen during an eclipse.

The corona-virus family is large, and causes respiratory infections humans, mammals and birds. Most members of the family cause only mild symptoms in healthy patients and, as they're explanation for about 15% of cases of the respiratory disorder, it's likely that you just have successfully fought off a corona virus infection without even noticing. However, the family also includes the highly pathogenic viruses chargeable for SARS, Near East respiratory syndrome (MERS), and in fact COVID-19. These diseases aren't only highly infectious but particularly virulent, with estimated death rates between 2 and 5%. The risks of highly aggressive corona- virus diseases like COVID-19 mustn't be understated.

The corona virus is primarily spread between people during close contact, most often via small droplets produced by sneezing, and coughing .Less commonly, people may become infected by touching a contaminated surface so touching their face. It's most contagious during the first three days after the onset of symptoms, although spread is possible before symptoms appear, and from those that don't show symptoms.

Common symptoms of Corona Virus include fever, cough, fatigue, shortness of breath, and loss of sense of smell. Complications may include pneumonia and acute respiratory distress syndrome. The time from exposure to onset of symptoms is typically around five days but may range from 2 to 14 days. There is not any known vaccine or specific

treatment for this.



2. Literature Review

Sn.	Paper Title	Author	DOI / ISSN /ISBN	Objective	Methodology	Conclusion
1	Corona Tracker: World-wide COVID-19 Outbreak Data Analysis and Prediction	Fairoza Amira Binti Hamzah, et al.	http://dx.doi.org/10.2471/BLT.20.255695	analyze the queried news, and classify the news into negative and positive sentiments, to understand the influence of the news to people's behavior both politically and economically	Susceptible-Exposed-Infectious-Recovered (SEIR) predictive modeling	At the time of writing this paper, the number of confirmed cases is expected to exceed 76000 cases, and reach the peak of this outbreak before 20 February 2020. The average Infected-Suspected ratio was found to be 2.399 which we used to initialize the number of Exposed individuals as a product of the number of Infected individuals on 20 Jan 2020. This outbreak is assumed to reach its peak in late May 2020 and will start to drop around early July 2020.
2	Presumed Asymptomatic Carrier Transmission of COVID-19	Drs Yao and Wang	https://jamanetwork.com/journals/jama/fullarticle/10.1001/jama.2020.2565	test the impact of corona treatment	Diagnostic treatment	effective
3	Predictive Analytics of COVID-19 using Information, Communication and Technologies	Parikshit N. Mahalle, et al.	https://doi.org/10.20944/preprints202004.0257.v1	Analysis of various predictive analytics methods	Application of Prophet predictive analytics algorithm on Kaggle dataset its predictions are also presented in this paper.	Simulation result of this model shows that the confirmed COVID-19 infected cases would be 1.6 million and 2.3 million by end of May and June respectively.
4	COVID-19 Epidemic Analysis using Machine Learning and Deep Learning Algorithms	Narinder Singh Punn · Sanjay Kumar Sonbhadra · Sonali Agarwal	https://doi.org/10.1101/2020.04.08.20057679	to analyze the transmission growth at the earliest and forecast the forthcoming possibilities of the transmission	predicted trend of the COVID-19 using SVR, PR, DNN, and LSTM	The results show that polynomial regression (PR) yielded a minimum root mean square error (RMSE) score over other approaches in forecasting the COVID-19 transmission. However, if the spread follows the predicted trend of the PR model then it would lead to huge loss of lives as it presents the exponential growth of the transmission worldwide
5	Mapping the landscape of Artificial Intelligence applications against COVID-19	Joseph Bullock, et al.	arXiv:2003.11336v2 [cs.CY] 23 Apr 2020	In this review, they present an overview of recent studies using Machine Learning and, more broadly, Artificial Intelligence, to tackle many aspects of the COVID-19 crisis at	AI TOOL	This research mapping exercise suggests that ML and AI can support the response against COVID-19 in a broad set of domains. In particular, we have highlighted emerging applications in drug discovery and development, diagnosis and

				different scales including molecular, clinical, and societal applications.		clinical outcome prediction, epidemiology, and infodemiology

Research Objectives

This paper contains following objectives:

- Analyze the rate of growth of cases in the world and India.
- Compare the growth rate of corona cases in different countries.
- Forecasting of the situation for next upcoming days.
- Prepare a forecasting model.

3. Research Methodology

It is a procedural framework which contains various steps like: data collection, data manipulation, data analysis, data interpretation etc. In this research paper for the Data Collection, I have referred various web crawlers like: journals, research papers, case studies and websites. I also referred to a dataset of global cases from Github for the work. I have analyzed the COVID-19 cases from 1/22/2020 to 6/22/2020.

For Data Manipulation and visualization I have applied various ML codes, for Data Analysis I have applied Time Series prediction model to forecast the possible deaths and active cases

As a forecasting tool, I have used ARIMA (Auto Regressive Integrated Moving Average) and for development of a model I have applied Prophet Model.

ARIMA: It is very popular and widely used statistical procedure for statistic forecasting. **ARIMA is an acronym that stands for Auto Regressive Integrated Moving Average**. It's a category of a model that captures a set of various standard temporal structures in statistic data. It explicitly caters to a set of normal structures in statistic data, and in and of itself provides a straightforward yet powerful method for creating skillful statistic forecasts.

Briefly,

- **AR: Auto Regression:** A model that uses the dependent relationship between an observation and a few number of lagged observations.
- **I: Integrated:** The employment of differencing of raw observations (e.g. subtracting an observation from an observation at the previous time step) so as to create the statisticstationary.
- **MA: Moving Average**: A model that uses the dependency between an observation and a residual error from a moving average model applied to lag observations.

Each of those components is explicitly laid out in the model as a parameter. A customary notation is employed of ARIMA (p, d, q) where the parameters are substituted with integer values to quickly indicate the precise ARIMA model being employed.

The parameters of the ARIMA model are defined as follows:

- **p**: The quantity of lag observations included within the model, also called the lag order.
- **d:** The amount of times that the raw observations are differenced, also called the degree of differencing.
- **q:** The dimensions of the moving average window, also called the order of moving average.

A rectilinear regression model is made including the required number and kind of terms, and therefore the data is ready by a degree of differencing so as to create it stationary, i.e. to get rid of trend and seasonal structures that negatively affect the regression model. A value of 0 is often used for a parameter, which indicates to not use that element of the model.

This way, the ARIMA model will be configured to perform the function of an ARIMA model, and even an easy AR, I, or MA model. Adopting an ARIMA model for a statistic assumes that the underlying process that generated the observations is an ARIMA process. This might seem obvious, but helps to motivate the necessity to verify the assumptions of the model within the raw observations and within the residual errors of forecasts from the model.

Prophet Model

Prophet is a procedure for forecasting time series data based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects. It works best with time series that have strong seasonal effects and several seasons of historical data. Prophet is robust to missing data and shifts in the trend, and typically handles outliers well.

Prophet is **open source software** released by Facebook's Core Data Science team.

How Prophet works. At its core, the **Prophet** procedure is an additive regression model with four main components: A piecewise linear or logistic growth curve trend. **Prophet** automatically detects changes in trends by selecting change points from the data

4. Data Analysis and Interpretation

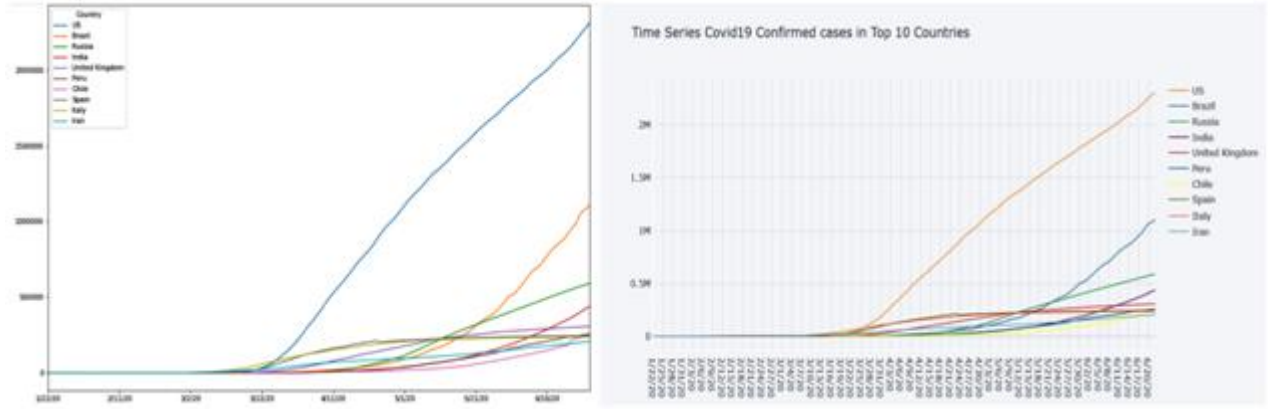


Figure 1: A Time Series Graph shows the status of exponential curve of outburst in top ten countries in the world. How growth in no of cases is increased?

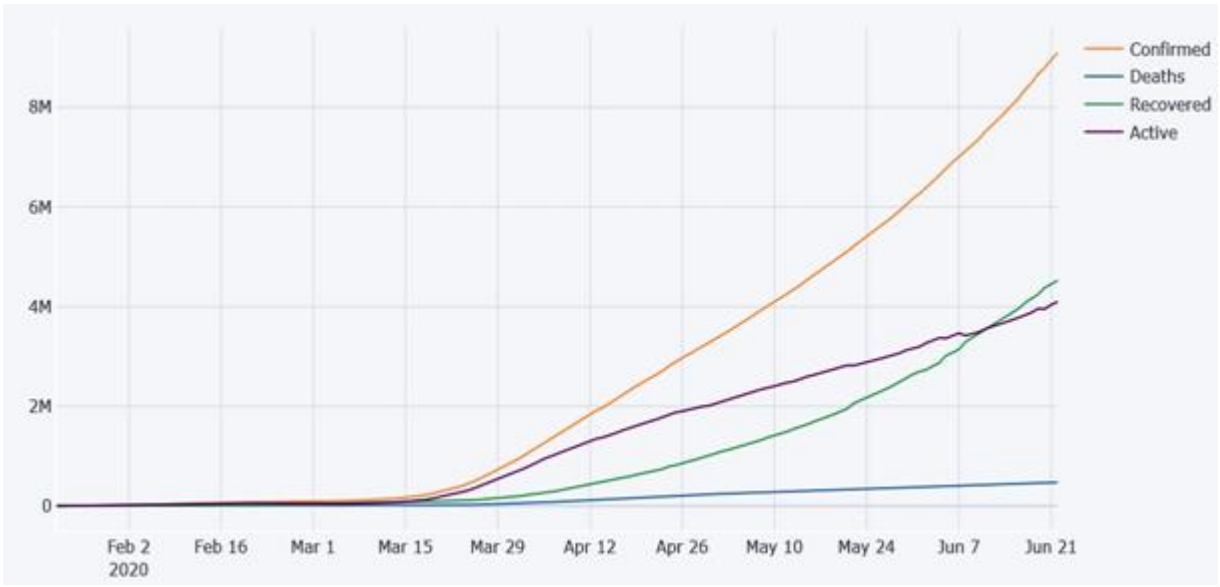


Figure 2: Represents the flow of confirmed cases, active cases, deaths and recoveries. As per the results graph inferred the status of different states of infected people in the world.

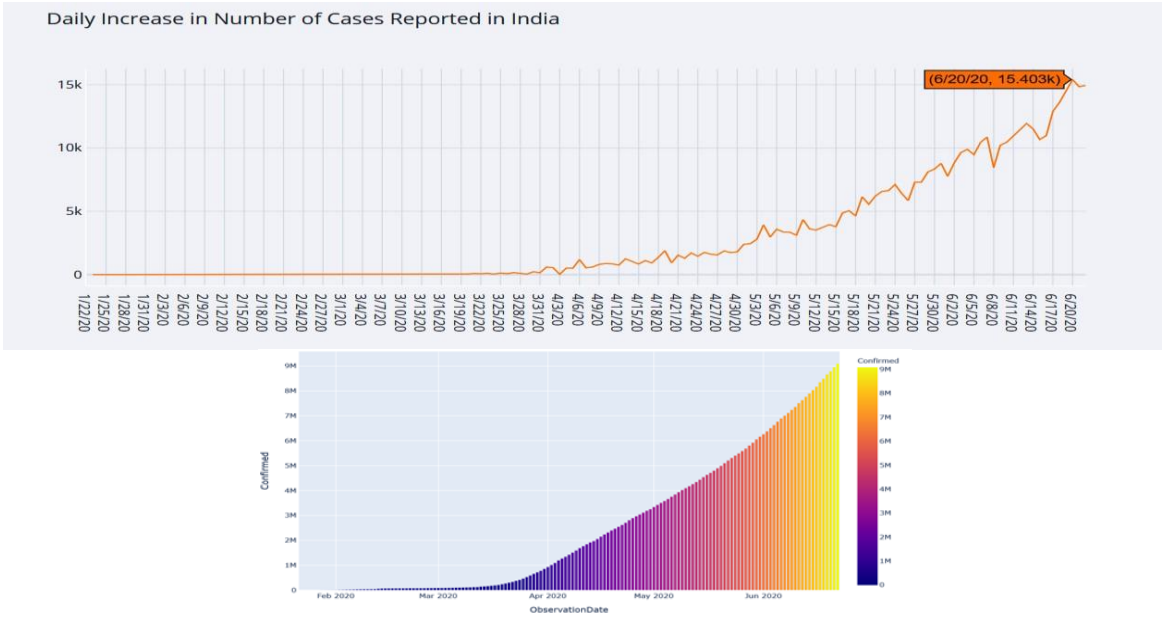


Figure 3: In this we can see the daily increase in the number of cases in India. As per the results the highest increase in the number of cases was on 20/06/2020, that was 15.403k.

Observation Date	Confirmed	Deaths	Recovered	Active	Mortality Rate
2020-01-22	555.0	17.0	28.0	510.0	3.063063
2020-01-23	653.0	18.0	30.0	605.0	2.756508
2020-01-24	941.0	26.0	36.0	879.0	2.763018
2020-01-25	1438.0	42.0	39.0	1357.0	2.920723
2020-01-26	2118.0	56.0	52.0	2010.0	2.644004
...	...	...	...	...	...
2020-06-02	6378238.0	380250.0	2729527.0	3268461.0	5.961678
2020-06-03	6508635.0	385947.0	2804982.0	3317706.0	5.929769
2020-06-04	6632985.0	391136.0	2869963.0	3371886.0	5.896832
2020-06-05	6770170.0	396121.0	3013132.0	3360917.0	5.850976
2020-06-06	6896917.0	400000.0	3085833.0	3411084.0	5.799693

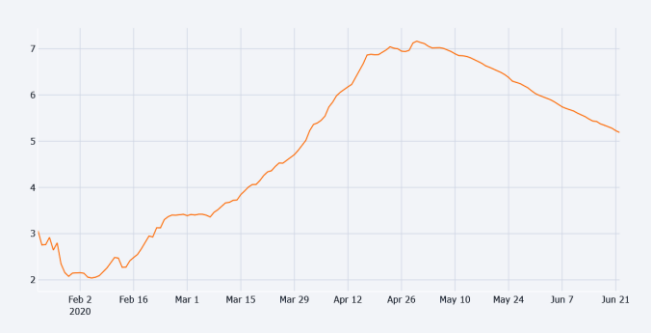


Figure 4: Here we can see date wise the mortality rate, which is calculated by (deaths/ confirmed cases)*100.As per results the highest mortality rate was 5.96 which inferred that on 02/06/2020 the rate is highest.

Figure 5: World map of affected region, where the darker the redness in the map, the more no. of affected people were there

5. Observations

A big rise can be seen in the number of confirmed cases from the month of April and it comes it its peak in mid June.

The mortality rate was at its peak in between 15 April to 10 may, but after 10 may death rate started decreasing that means no. of recoveries started increase.

As per data set the total no of confirmed cases: 5495061; total no of deaths: 346232; total no of recovered cases: 2231738; total no of active cases: 2917091

Top ten countries, most affected by corona:

1) UNITED STATES

2) BRAZIL
3) RUSSIA
4) INDIA
5) UNITED KINGDOM
6) PERU
7) CHILES
8) SPAIN
9) ITLAY
10) IRAN

Prediction Model for Next 10 Days: For prediction of cases I have applied Autoregressive Integrated Moving Average (ARIMA) Model and FBPROPHET Model. These are widely used methods for time series forecasting.

ObservationDate	Confirmed	Deaths	Recovered	Active	Mortality Rate	ARIMA Model Prediction
2020-06-07	7010582.0	402744.0	3140920.0	3466918.0	5.744801	7.094984e+06
2020-06-08	7119002.0	406543.0	3293408.0	3419051.0	5.710674	7.299826e+06
2020-06-09	7242313.0	411436.0	3375673.0	3455204.0	5.681003	7.533318e+06
2020-06-10	7360239.0	416201.0	3454807.0	3489231.0	5.654721	7.798989e+06
2020-06-11	7514481.0	421458.0	3540696.0	3552327.0	5.608611	8.088048e+06
2020-06-12	7632802.0	425394.0	3613277.0	3594131.0	5.573235	8.423959e+06
2020-06-13	7766952.0	429736.0	3698304.0	3638912.0	5.532878	8.798150e+06
2020-06-14	7900924.0	433066.0	3769712.0	3698146.0	5.481207	9.182323e+06
2020-06-15	8034504.0	436899.0	3857339.0	3740266.0	5.437784	9.629394e+06
2020-06-16	8173940.0	443685.0	3955169.0	3775086.0	5.428043	1.011515e+07
2020-06-17	8349950.0	448959.0	4073955.0	3827036.0	5.376787	1.063976e+07
2020-06-18	8488977.0	453985.0	4155099.0	3879893.0	5.347935	1.122749e+07
2020-06-19	8664986.0	460018.0	4245777.0	3959191.0	5.308930	1.184275e+07
2020-06-20	8791804.0	464465.0	4378255.0	3949084.0	5.282932	1.252896e+07
2020-06-21	8955536.0	468365.0	4447086.0	4040085.0	5.229894	1.329347e+07
2020-06-22	9098641.0	472171.0	4526333.0	4100137.0	5.189467	1.409136e+07

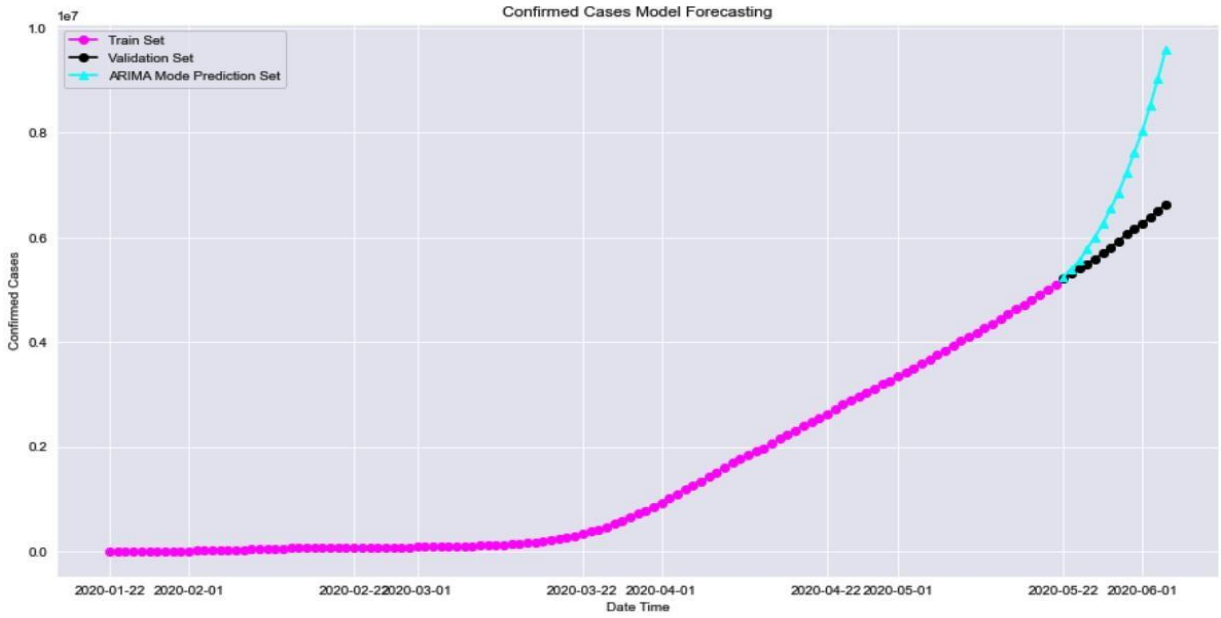


Figure 6: The above chart and graph clearly represents that the rate of increase of cases are getting worse day by day.

Prophet Model: A model of forecasting:

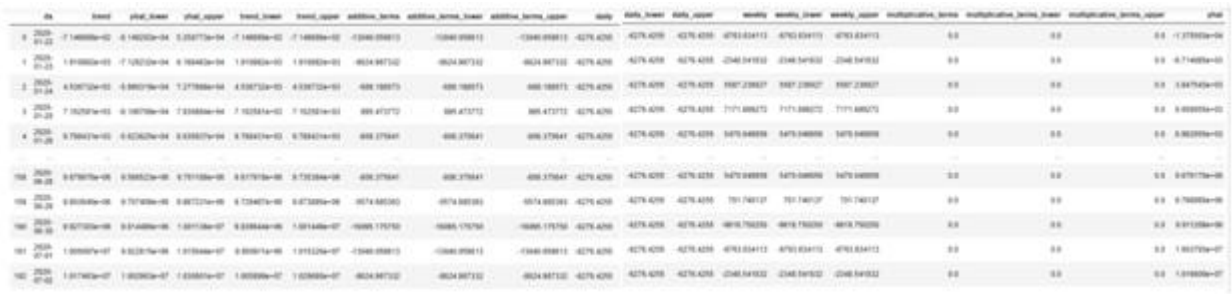


Figure 7: The above chart shows the number of confirmed cases with upper limits and lower limits.

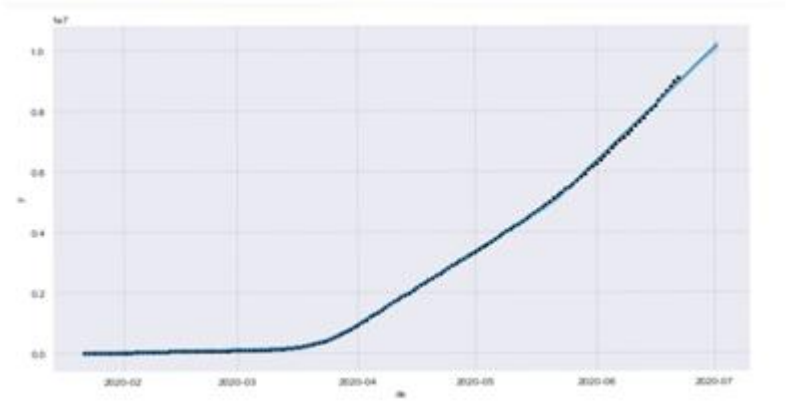


Figure 8: As per the prediction model, the above graph shows the limit of increase

By using this prediction, it can be concluded that the number of confirmed cases will increases on the daily basis in upcoming days. Hence all the nationals of the world have to follow the preventions to cope up with this epidemic. The preventive strategies are focused on the isolation of patients and careful infection control, including appropriate measures to be adopted during the diagnosis and the provision of clinical care to an infected patient. For instance, droplet, contact, and airborne precautions should be adopted during specimen collection, and sputum induction should be avoided.

6. Conclusion

At the same time as the COVID-19 situation is serious, and far from under control, it is important to remember that whatever the outcome, future outbreaks of new COVID-19 (Corona Virus) strains or further emergent diseases are inevitable.

References

[1] https://www.mewburn.com/news-insights/coronavirus-a-case-study-in-emerging-disease
[2] https://en.wikipedia.org/wiki/Coronavirus
[3] https://www.ncbi.nlm.nih.gov/books/NBK554776/
[4] https://www.esmo.org/oncology-news/findings-from-a-retrospective-case-study-of-covid- 19-infection-in-cancer-patients-in-wuhan-an-emphasis-on-severe-events
[5] https://www.healthmap.org/COVID-19/
[6] https://www.nejm.org/coronavirus
[7] https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/#:~:text=ARIMA%2C%20short%20for%20'Auto%20Regressive, used%20to%20forecast%20future%20values.
[8] https://www.safeture.com/coronavirus/
[9] Managing Research Risks during an Emerging Infectious Diseases Event: A Coronavirus Case Study by Advarra.
[10] Features, Evaluation and Treatment Coronavirus (COVID-19) by Marco Cascella; Michael Rajnik; Arturo Cuomo; Scott C. Dulebohn; Raffaela Di Napoli
[11] Understanding epidemic data and statistics: A case study of COVID-19, Amirhoshang Hoseinpour Dehkordi, Majid Alizadeh , Pegah Derakhshan , Peyman Babazadeh , Arash Jahandideh First published:24 April 2020, https://doi.org/10.1002/jmv.25885
[12] Understanding Epidemic Data and Statistics: A case study of COVID-19 **Preprint (PDF Available)** · March 2020 *with* 964 Reads DOI: 10.1101/2020.03.15.20036418
[13] The Novel Corona virus Pneumonia Emergency Response Epidemiology Team. The Epidemiological Characteristics of an Outbreak of 2019 Novel Corona virus Diseases (COVID-19)—China, 2020. China CDC Weekly. 2020;2:113-22
[14] WHO. Novel Corona virus(2019-nCoV) Situation Report - 39. 2020 [cited 2020 March 1]; Available from: https://www.who.int/docs/defaultsource/coronaviruse/situationreports/2 0200228-sitrep-39-covid19.pdf?sfvrsn=5bbf3e7d_2
[15] WHO. Novel Corona virus (2019-nCoV) Situation Report - 52. 2020 [cited 2020 March 12]; Available from:

https://www.who.int/docs/defaultsource/coronaviruse/ situation- reports/20200312-sitrep-52-covid19.pdf?sfvrsn=e2bfc9c0_4
[16] The center of Evidence-Based Medicine Develops, Promotes and disseminates better evidence for healthcare CEBM-University of OXFORD report [cited 2020 April 6] https://www.cebm.net/COVID-19/COVID-19-what-proportion-are-asymptomatic/
[17] "Coronavirus Disease 2019 (COVID-19), " Centers for Disease Control and Prevention, 2020.[Online].Available:https://www.cdc.gov/coronavirus/2019-ncov/about/symptoms.html. [Accessed 27 February 2020].
[18] Forecasting: Principles and Practice, Rob J Hyndman and George Athanasopoulos Monash University, Australia

Author Profile

Achyut Tiwari is a B.tech 2nd year student in department of Computer Science and Engineering College, Ghaziabad, UP, India. He is also awarded with Young Scientist Award by UP Government. His area of interest is Web Development, Machine learning and Deep learning.