PREDICTIVE ANALYSIS OF COVID-19 AND ITS PREVALENCE

[1]D. V. Koushik, [2]D. Sai Vikas, [3]G. Anand Vardhan, [4]G. Venkata Sai Pavan Kumar, [5]Dr. M. Ramakrishna Murty

[1]Student, [2]Student, [3]Student, [4]Student, [5]Professor,
Department of Computer Science and Engineering,
Anil Neerukonda Institute of Technology and Sciences, Visakhapatnam, India.

Abstract: COVID-19 is a virus that is rapidly spreading across the World. Until now no full-fledged drug or vaccine is developed for this virus. Though people are recovering with the support of antibiotics, anti-viral drugs, C- Vitamin supplementation, and Remdesivir injections. It is now obvious that the world needs an immediate and faster solution to block and tackle the further prevalence of COVID-19 across the world with the aid of non-clinical methods such as data mining, AI, and other augmented intelligence techniques to ease the huge load on the healthcare system while providing the finest possible means for patients' diagnosis and prediction of the 2019-nCoV epidemic efficiently. There is an immediate requirement for a solution to contain the rise of COVID-19 cases across the World and this can be done with the help of Data Mining and Machine Learning by forecasting or predicting the trends in the COVID-19 cases. The models forecast the number of Confirmed, Recovered, and Death cases for the next week. By forecasting the trend of corona-virus we can support health and medical departments to make vital decisions regarding the pandemic.

Keywords **– COVID-19, Linear Regression, Support Vector Regression, Polynomial regression, ARIMA, Prophet, RMSE, MAE, MAPE.**

1. Introduction

In 2019, a deadly virus named Coronavirus or COVID-19 came into existence. It spreads quickly in individuals and its primary indications are fever, cold, cough, uneasiness in breathing. All these are similar to the flu. COVID-19 reached a pandemic level and resulted in deaths across the World and human-to-human transmission. There is a need for Data Mining techniques, ML techniques, and AI techniques to limit the further outbreak of COVID-19. It is a new pandemic with high prevalence rate. It requires a worldwide and united response of all national medical and healthcare organizations. Due to COVID-19, there is a rise in the demand for quick response and data exchange on swiftly spreading global pandemics. This virus has unique characteristics such as high transmissibility, extensive fatal deaths and is capable of disrupting the world socially and financially.

The pandemic has taken grasp over peoples' life. Since the start of the pandemic, some countries are facing the problem of cumulative cases. Through the analysis of COVID-19 data, one can know how countries all over the world are doing in terms of controlling the pandemic. Analyzing data leads to familiarize the prevention model of the countries that are doing great in terms of depressing the graph. Forecasts are made with the dataset available to the individual/country/organizations, thus assisting them to decide how long they can control the contagion or up to how much extent they should provide preventive measures. Several ML models were developed in this project to forecast the COVID-19 confirmed, recovered, and death cases. The health workers can gain information about the pandemic so that they can work accordingly utilizing these models. ML models developed are Linear Regression Model, SVM Regression Model, Polynomial Regression Model, ARIMA Model, and FB Prophet Model.

Through this project, a step towards helping people to understand the spread and predict the cases in their country is done. This project also gives an insight into how a country is doing in terms of limiting the spread. By forecasting the trend of the virus we can support health and medical departments to make vital decisions regarding the pandemic.

2. Literature Survey

Several kinds of research are done on forecasting the pandemics. "In 2007, a study on the spread of the H5N1 influenza virus in birds was done by Hall, Gani, Hughes, and Leach" [6]. They used regression analysis to forecast the influenza virus timing and prevalence. "A predictive model was built in 2020 by Remuzzi and Remuzzi that will help to understand the rise in confirmed cases which assists the medical department to take necessary decisions" [7]. It discussed the grave impact that was caused by the COVID-19 on China and Italy. "In 2020, Roosa et al. utilized phenomenological models and forecasted the reported cases in Hubei Province, China. Data from the National Health Commission of China is considered by them and their model also depicts the containment strategies implemented in China" [8]. "In 2020, Benvenuto, Giovanetti, Vassallo, Angeletti, and Ciccozzi published a research paper in which they proposed a simple ARIMA model to predict the COVID-19. They utilized the Johns Hopkins epidemiological data to predict the trend of prevalence of COVID-19" [9]. "Fatemeh Ahouz, Behbahan Khatam Alanbia University of Technology, Behbahan, Iran & Amin Golabpour proposed a machine learning model to forecast the number of COVID-19 patients' across the world" [1]. "Data Modelling & Analysing Coronavirus (COVID19) Spread using Data Science & Data Analytics in Python Code" [3] By Jatin Chaudhary utilizes the SIR model to predict the trends of COVID-19 across India. We have gone through other citations and websites also and the links related to them were given in the References section.

3. Data Acquisition and Description

“The data set used in this project was the real-time data from Johns Hopkins Centre for Systems Science and Engineering GitHub repository” [11]. The repository has separate datasets for confirmed cases, death cases and recovered cases. The datasets contain Province/State, Country/Region, Latitude, Longitude, and dates as columns. The data from these datasets were merged to obtain the parameterized datasets of the world from January 22, 2020, till March 11, 2021. These were used for the data analysis and visualization part. The dataset consisting of 345 days i.e. from January 22, 2020, till December 31, 2020 is considered for the prediction part. The main reason for opting for regression analysis is that it is the best procedure to predict continuous dependent variables from several independent variables and also the dataset is a continuous one.

4. Feature Selection

Feature selection is one of the steps in EDA that extracts and selects the finest outputs from our model. It shows the underlying structure of the data from this we can say that it has the best features. The performance of the model is affected significantly by using feature engineering. It includes some features that are split or combined to get new features or to collect the data from external sources. Conclusions of the dataset can be evaluated and drew with the help of dimensionality reduction. The dates are changed to date-time object and irrelevant attributes are removed to get better conclusions from the dataset. To get the parameterized datasets of the world from 22nd January, 2020 till 11th March, 2021 we have merged the data of these datasets.

6. System Architecture

The study aims to analyze COVID-19 data and future forecasting the trends of COVID-19 (Confirmed cases, Death cases, and Recovered cases). As mentioned above the dataset used in this study was taken from the repository of GitHub provided by the “Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE)” [11]. This repository contains the three datasets for confirmed cases, death cases, and the recovered cases respectively. The data are updated every day for all daily cases. To get better decisions from this dataset, irrelevant columns were removed. Data cleaning was performed by removing rows with zeros from the dataset. The dataset used in this project is considered for 345 days from January 22, 2020, to December 31, 2021. The architecture of this research work is shown in Figure 1.

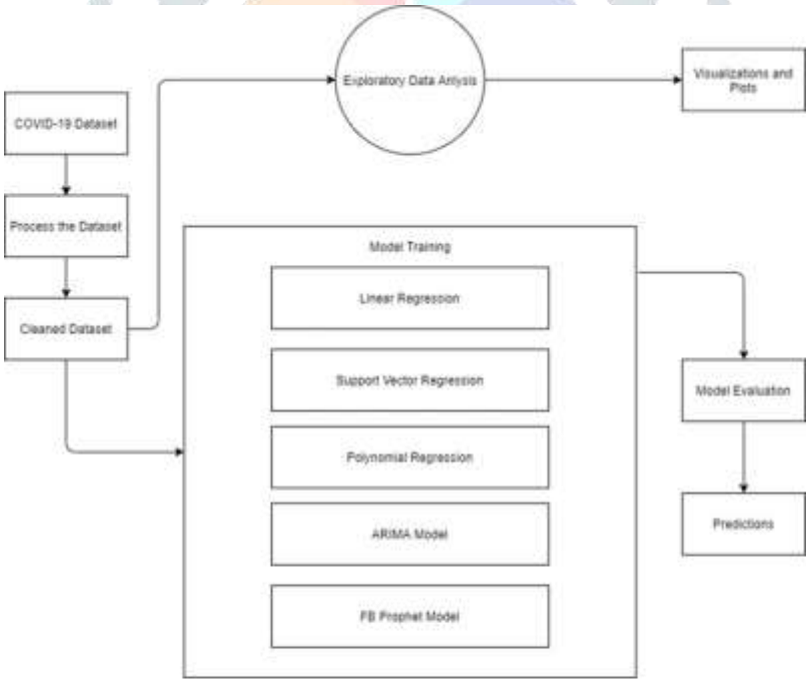


Fig 1. System Architecture

In the model training part, the training algorithm takes the training dataset with input COVID-19 cases to train the regression model with the regressor. After the completion of training, the regressor takes the test data as the input and predicts the output for it i.e. it generates a forecast for the given dataset. These outputs help us to evaluate the regression models using evaluation metrics. Then the whole dataset is trained with the regressor and the cases were forecasted for the next seven days. The Regression models applied in the paper are- Linear Regression Model, Support Vector Regression Model, Polynomial Regression Model, Auto Regression Integrated Moving Average (ARIMA) Model, and FB Prophet Model.

7. Data Analysis and Visualization

The process of analyzing datasets to make decisions on the information one has with a specialized software or a system is known as Data analytics. These technologies are now-a-days used to analyze pandemic situations around the World. These processes are automated into algorithms and mechanical processes which work on raw data. The graphical representation of information and data in a pictorial or graphical format like charts, graphs, and maps is known as Data Visualization. Trends, patterns in data and outliers are identified using Data Visualization tools.

Utilizing data in [11] simulations were carried out. Data is considered from January 22, 2020 to March 11, 2021. The libraries used to perform analysis are Panda library [10], Plotly library [10], and Folium library [10] in collaboration with Python. Figure 2. shows the cumulative number of confirmed, recovered, and death cases. According to it on March 12, 2021 there are 119.0605 million confirmed cases in the World. Figure 3. shows various summary statistics, by giving the mean, standard deviation, minimum and maximum values, and the quantiles of the data.

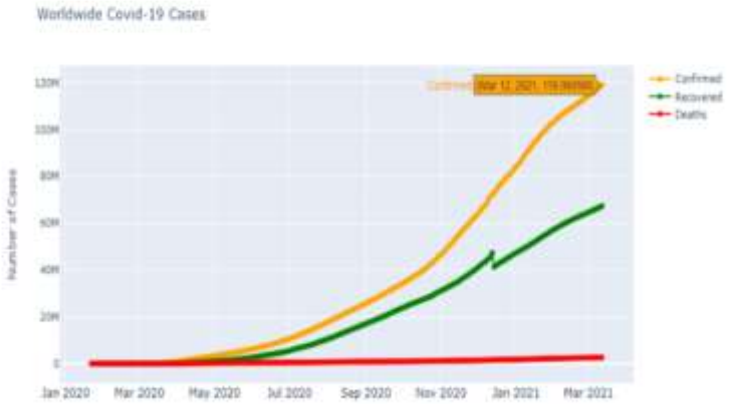
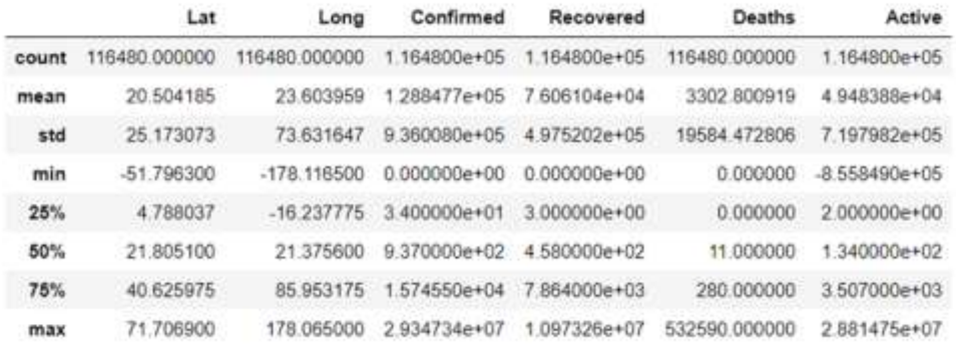


Fig 2. Cumulative number of confirmed, recovered, and death cases



	Lat	Long	Confirmed	Recovered	Deaths	Active
count	116480.000000	116480.000000	1.164800e+05	1.164800e+05	116480.000000	1.164800e+05
mean	20.504185	23.603959	1.288477e+05	7.606104e+04	3302.800919	4.948388e+04
std	25.173073	73.631647	9.360080e+05	4.975202e+05	19584.472806	7.197982e+05
min	-51.796300	-178.116500	0.000000e+00	0.000000e+00	0.000000	-8.558490e+05
25%	4.788037	-16.237775	3.400000e+01	3.000000e+00	0.000000	2.000000e+00
50%	21.805100	21.375600	9.370000e+02	4.580000e+02	11.000000	1.340000e+02
75%	40.625975	85.953175	1.574550e+04	7.864000e+03	280.000000	3.507000e+03
max	71.706900	178.065000	2.934734e+07	1.097326e+07	532590.000000	2.881475e+07

Fig 3. Summary statistics until March 11th, 2021 (Lat, Long, Confirmed, Recovered, Deaths, Active).

Figure 4. illustrates World-wide Covid-19 confirmed, recovered and death cases with time lapse over World Map. Figure 5. shows the cumulative Covid-19 cases over time (Active, Deaths, and Recovered) plotted using area plot. Figure 6. illustrates the countries and provinces affected by Covid-19 in the World plotted using Folium Maps. Figure 6. illustrates the number of new cases per day and number of countries affected daily in the World. Figure 7. shows the Newly emerged COVID-19 cases in the World and Figure 8. Depicts the Line plot for Number of days it took to get million cases after 100th case in the World. Figure 14. shows comparison of Confirmed, Deaths and Mortality of Covid-19 with other pandemics like SARS, EBOLA, MERS, H1N1. All these plots were plotted with the help of Plotly Library [10].

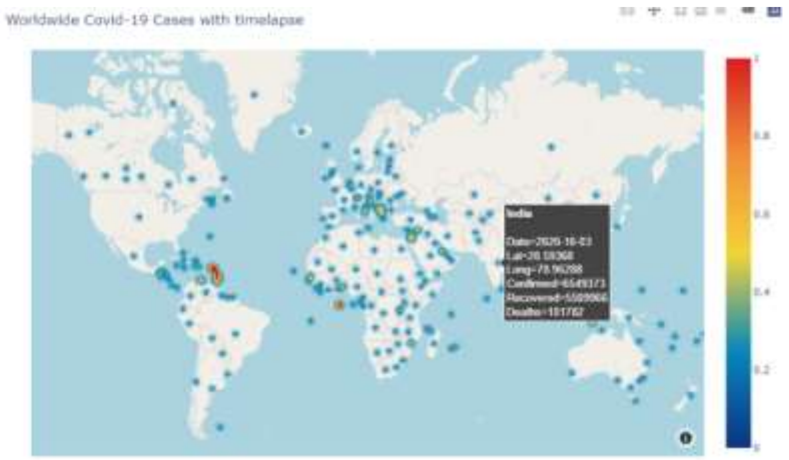


Fig 4. World-wide Covid-19 confirmed, recovered and death cases with time lapse over World Map.

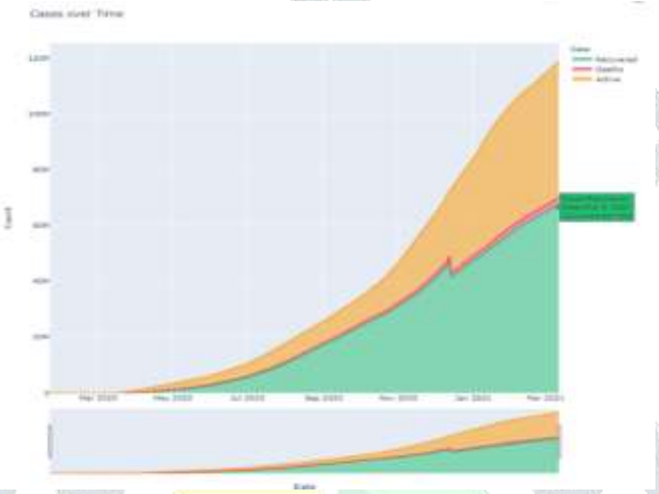


Fig 5. Cumulative Covid-19 cases over time.

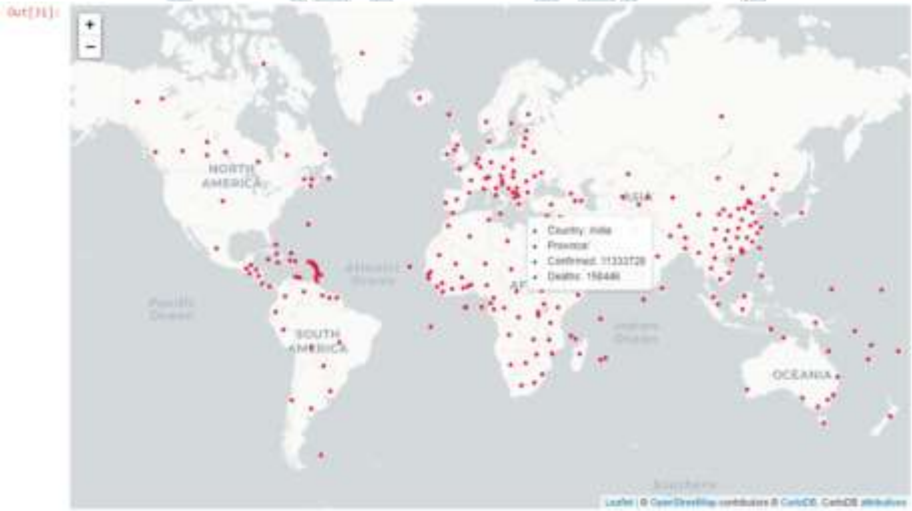


Fig 5. Countries and Provinces affected by Covid-19 in the World

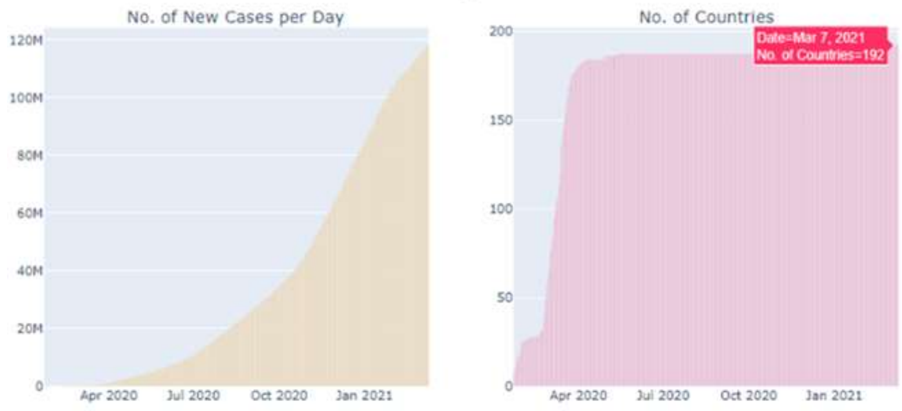


Fig 6. No. of new cases per day and No. of countries affected daily in the World.

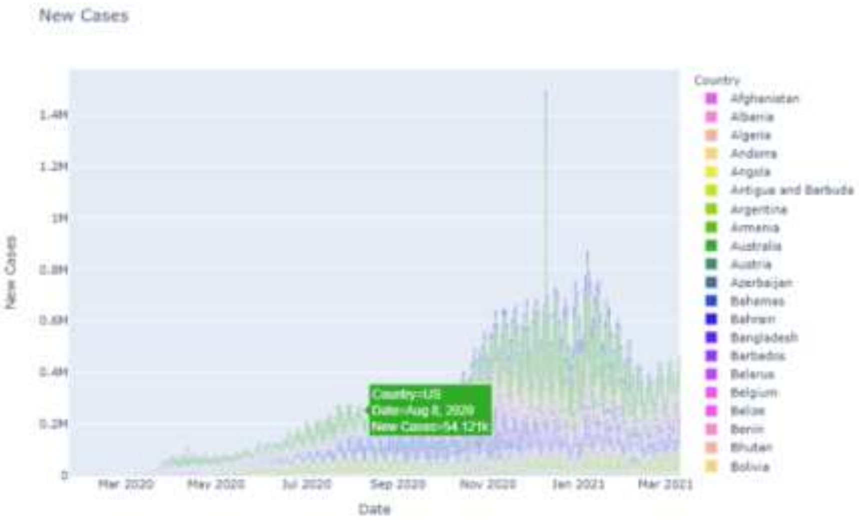


Fig 7. New Cases over the World.

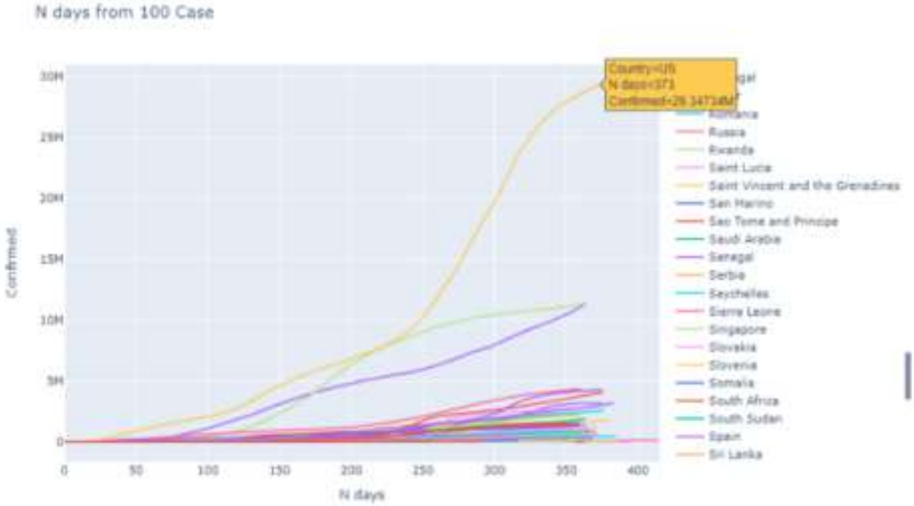


Fig 8. Number of Days from 100th case.

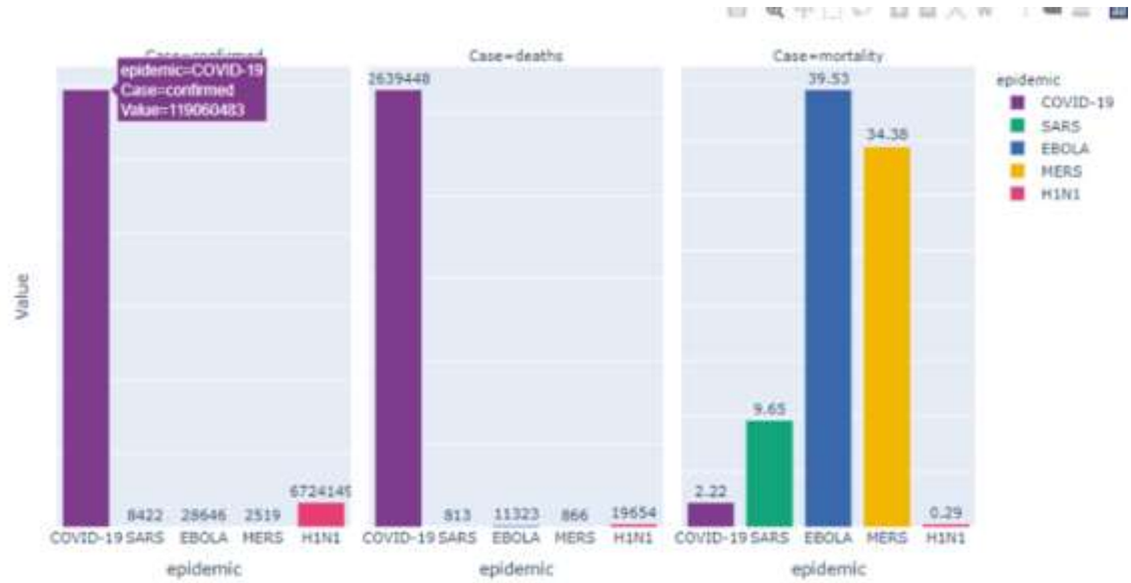


Fig 9. Comparison of Confirmed, Deaths and Mortality of Covid-19 with other pandemics like SARS, EBOLA, MERS, H1N1.

8. Prediction Models

8.1. Linear Regression Model

Linear regression is a model that assumes a linear relationship between the input variable and the output variable. Let, x be the input variable and y be the output variable then y can be calculated from a linear combination of variable x. Eq.1 is the general equation of Linear Regression.

$$y = w_0 + \sum_{i=1}^{n} w_i x_i + \varepsilon_i \qquad (1)$$

Here, w_0 is the parameter (linear) which is to be computed, x_i represents one or more predictors or independent variables and ε_i represents error values.

In our project, we have divided the dataset into training and testing parts and we have two variables namely x (independent variable) and y (dependent variable). We have divided the x and y into training and test sets respectively. Using x training set and y training set we trained the machine. By using x test set we predicted the output and compared it with the original y test set to know how the model performs. Figure 10. illustrates the flow diagram of Linear Regression process to find the upcoming Covid-19 cases in the World and also India. Figure 11. depicts the comparison between the Covid 19 Cases and the Linear Regression Predictions of the World and Figure 12. depicts comparison between the Covid 19 Cases and the Linear Regression Predictions of India. Table 1. shows the forecasts of the Linear Regression Model for the World and India.

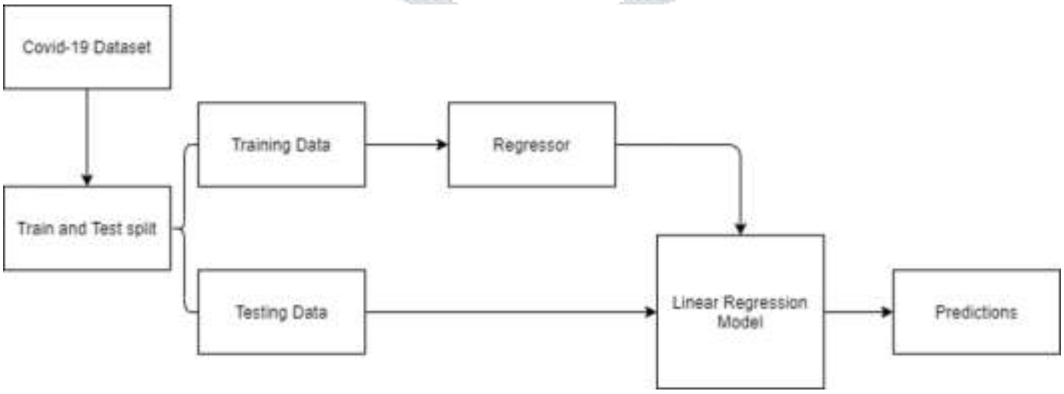


Fig 10. Flow diagram of Linear Regression Model.

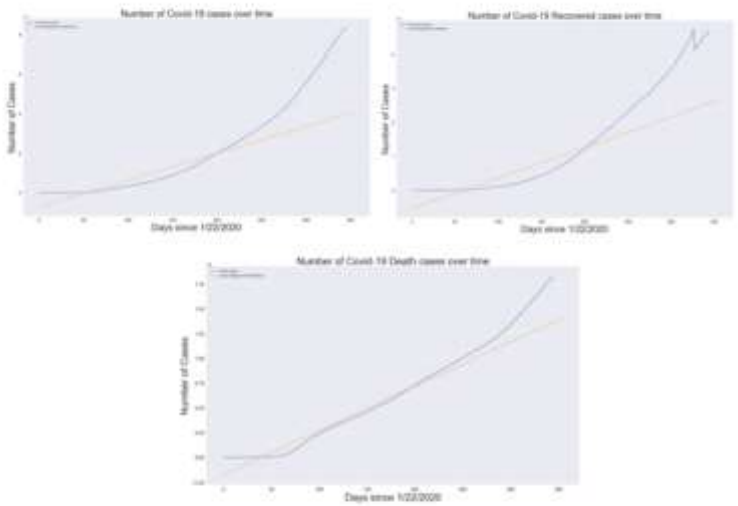


Fig 11. Comparison between the Covid 19 Cases and the Linear Regression Predictions of the World.

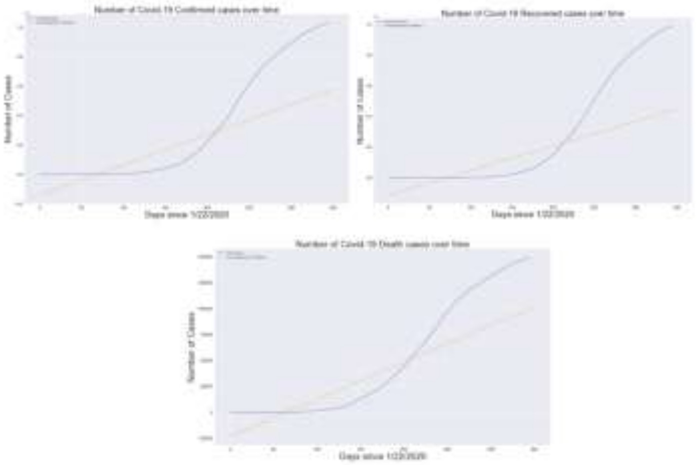


Fig 12. Comparison between the Covid 19 Cases and the Linear Regression Predictions of India.

Table 1. Forecasts of the Linear Regression Model for the World and India.

Linear	WORLD			INDIA		
Dates	**Confirmed**	**Recovered**	**Deaths**	**Confirmed**	**Recovered**	**Deaths**
1-1-2021	39592070	25619780	1368901	5659346	4380790	99512
1-2-2021	39727695	25709678	1373377	5679855	4396785	99867
1-3-2021	39863320	25799576	1377852	5700364	4412779	100222
1-4-2021	39998945	25889475	1382328	5720874	4428774	100578
1-5-2021	40134570	25979373	1386804	5741383	4444769	100933
1-6-2021	40270194	26069271	1391279	5761892	4460764	101289
1-7-2021	40405819	26159169	1395755	5782402	4476759	101644

8.2. Support Vector Regression Model

A regression algorithm that works on the principle of a Support Vector Machine (SVM) is known as the SVM Regression algorithm. It supports both linear and non-linear regressions. SVM regressor is used to predict continuous ordered variables. The main idea of SVR is to contain the error within a threshold i.e., it approximates the best value within a given margin. For Classification or outlier detection in an n-D space, a hyperplane or hyperplanes are constructed The best parameters are considered for the SVM-Regressor and the model is fitted using the training set and the test set is used to forecast the values.

Figure 13. illustrates the flow diagram of SVM Regression process to find the upcoming Covid-19 cases in the World and also India. Figure 14. depicts the comparison between the Covid 19 Cases and the Support Vector Regression Predictions of the World and Figure 25. depicts comparison between the Covid 19 Cases and the Support Vector Regression Predictions of India. Table 2. shows the forecasts of the SVM Regression Model for the World and India.

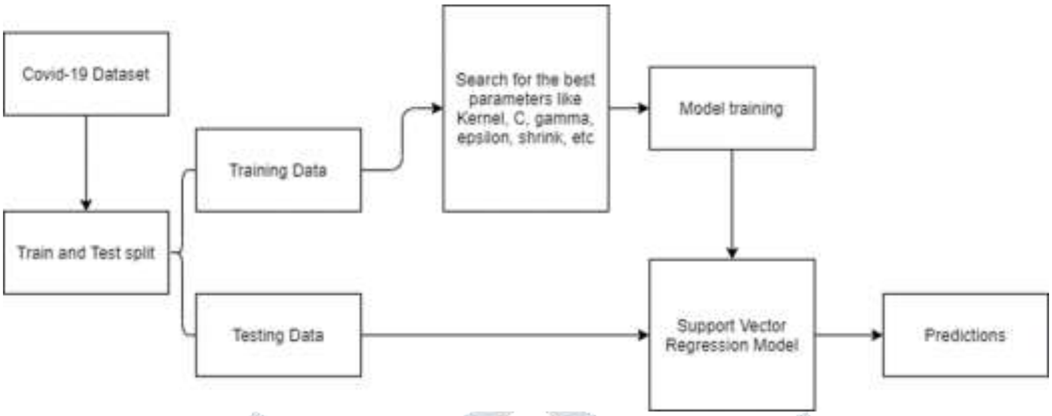

Covid-19 Dataset
Train and Test split
Training Data
Testing Data
Search for the best parameters like Kernel, C, gamma, epsilon, shrink, etc
Model training
Support Vector Regression Model
Predictions

Fig 13. Flow diagram of SVM Regression Model.

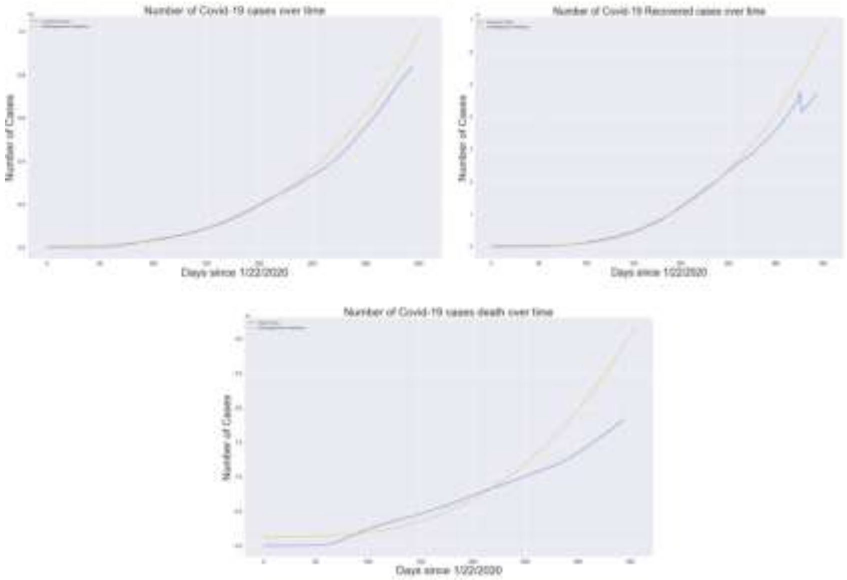


Fig 14. Comparison between the Covid 19 Cases and the Support Vector Regression Predictions of the World.

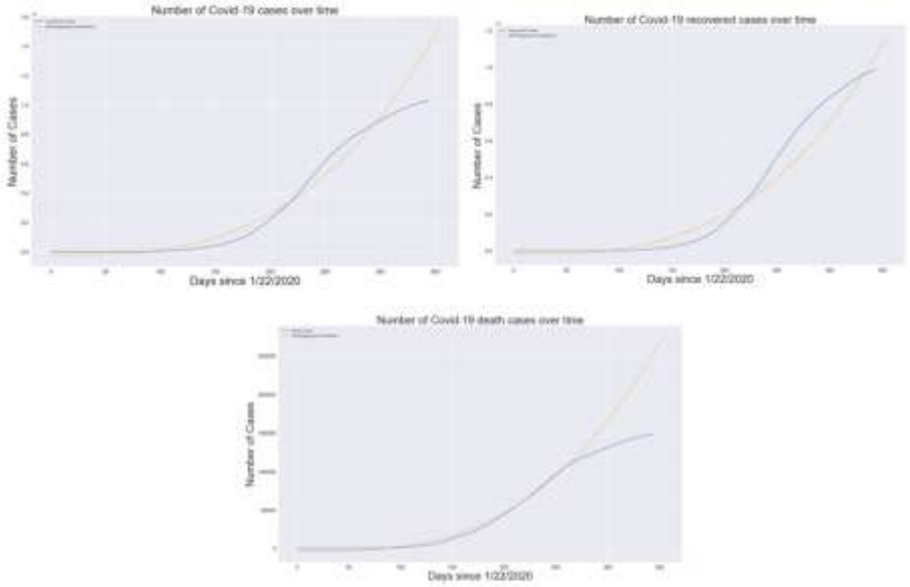


Fig 15. Comparison between the Covid 19 Cases and the Support Vector Regression Predictions of India

Table 2. Forecasts of the Support Vector Regression Model for the World and India.

SVR	WORLD			INDIA		
Dates	**Confirmed**	**Recovered**	**Deaths**	**Confirmed**	**Recovered**	**Deaths**
1-1-2021	94382128	62321394	2915896	14102058	10791785	254253
1-2-2021	95197468	62868096	2940206	14227024	10887513	256504
1-3-2021	96017535	63417968	2964657	14352715	10983797	258767
1-4-2021	96842342	63971017	2989249	14479132	11080637	261044
1-5-2021	97671903	64527255	3013984	14606277	11178036	263334
1-6-2021	98506231	65086689	3038860	14734154	11275994	265637
1-7-2021	99345341	65649329	3063879	14862763	11374513	267954

8.3. Polynomial Regression Model

The power of the y (independent variable) is greater than one in the Polynomial Regression Model. The best line fitting is in the form of a curve in this regression process. Polynomial regression is used to fit a wide range of curve. A better dependent-independent variable inter-relationship can be provided by the Polynomial regression. Based on the target variable and the predictor variable relationship the degree of polynomial is chosen. If we choose a degree equal to 1 then it behaves like simple linear regression. So, the degree should be more than 1 to be considered as Polynomial regression. For this project, we have chosen degree as 2.

$$y = w_0 + \sum_{i=1}^{n} w_i x^i + \varepsilon_i (1)$$

In Eq.1, x is the independent variable score, y is the estimated score of dependent variable, w_0 is the constant variable, w1, w2, ..., wn = are the weights and n states the polynomial degree.

Figure 16. illustrates the flow diagram of Polynomial Regression process to find the upcoming Covid-19 cases in the World and also India. Figure 17. depicts the comparison between the Covid 19 Cases and the Polynomial Regression Predictions of the World and Figure 18. depicts comparison between the Covid 19 Cases and the Polynomial Regression Predictions of India. Table 3. shows the forecasts of the Polynomial Regression Model for the World and India.

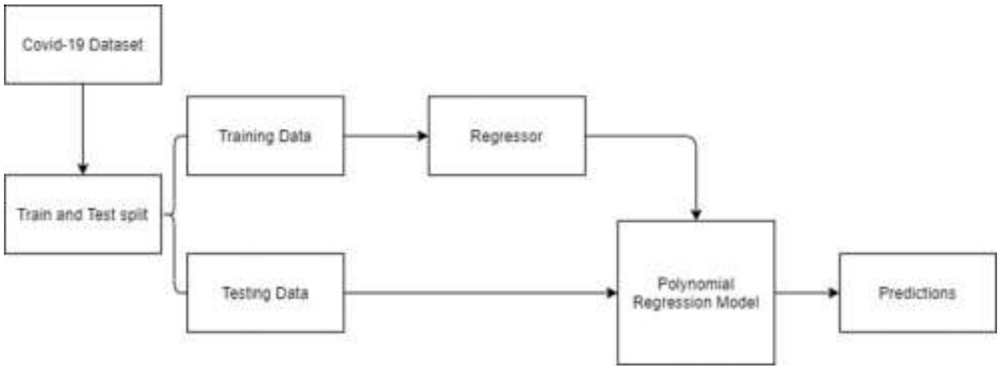


Fig 16. Flow diagram of Polynomial Regression Model.

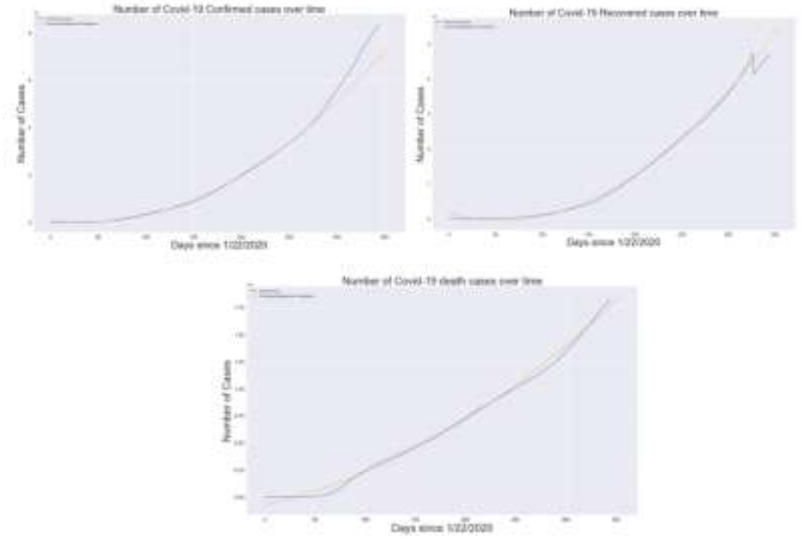


Fig 17. Comparison between the Covid 19 Cases and the Polynomial Regression Predictions of the World.

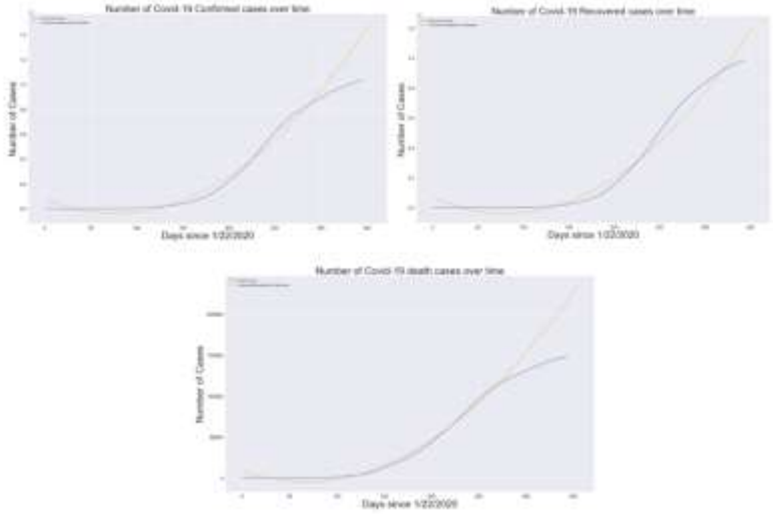


Fig 18. Comparison between the Covid 19 Cases and the Polynomial Regression Predictions of India.

Table 3. Forecasts of the Polynomial Regression Model for the World and India.

Polynomial	WORLD			INDIA		
Dates	**Confirmed**	**Recovered**	**Deaths**	**Confirmed**	**Recovered**	**Deaths**
1-1-2021	69887335	51392568	1768426	13924623	11351236	219932
1-2-2021	70341122	51753132	1777097	14031934	11440434	221552
1-3-2021	70796375	52114945	1785788	14139646	11529970	223178
1-4-2021	71253094	52478004	1794498	14247757	11619844	224809
1-5-2021	71711279	52842311	1803227	14356269	11710054	226447
1-6-2021	72170931	53207865	1811976	14465180	11800602	228090
1-7-2021	72632048	53574667	1820744	14574492	11891488	229739

8.4. ARIMA Model

ARIMA stands for Auto-Regressive Integrated Moving Average. Time series data is analyzed and forecasted using this model. It is a generalization model of a simpler version of ARMA model and is included with idea of integration. Its acronym is expressive where 'AR' stands for 'Auto Regression' which states that this model utilizes observation-lagged observations inter-relational dependency. 'I' represents 'Integrated' which states that the time series is made stationary by the model utilizing the differences between the fresh observations. 'MA' stands for 'Moving Average' states that the model utilizes the inter-observation dependencies and the residual error obtained by applying the model to the lagged observations.

The parameters utilized in the Auto Regressive integrated Moving Average Model are:

p: It denotes lag order which states the number of lags in the model.

d: It denotes the degree of difference and states the count of differences between the raw observations is defined by it.

q: It is known as the order of the (MA) Moving Average. It states the area coverage of the MA window

Eq.1 is the general equation for a 'p' order AR model is as follows:

$$y_t = C + \phi_1 y_{t-1} + \phi_2 y_{x-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t \qquad (1)$$

Here y_t is the data on which the model is to be applied. The parameters $\phi_1, \phi_2, \ldots$ are coefficients of AR.

Eq.2 is the general equation for a 'q' order MA model is as follows:

$$y_t = C + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q} (2)$$

Here y_t is the data on which the model is applied and $\theta_1, \theta_2, \ldots$ and so on are coefficients of MA.

From Eq.1 and Eq.2, Eq.3 is obtained.

Eq.3 is the general equation for an ARIMA (p, q) model is as follows:

$$y_t = C + \phi_1 y_{t-1} + \phi_2 y_{x-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q} \qquad (3)$$

Here y_t is the data on which the model is applied and ϕ_1, ϕ_2, ..., θ_1, θ_2, ... are the coefficients of AR and MA respectively.

Figure 19. illustrates the flow diagram of ARIMA process to find the upcoming Covid-19 cases in the World and also India. Figure 20. depicts the comparison between the Covid 19 Cases and the ARIMA Predictions of the World and Figure 21. depicts comparison between the Covid 19 Cases and the ARIMA Predictions of India. Table 4. Shows the forecasts of the ARIMA Model for the World and India.

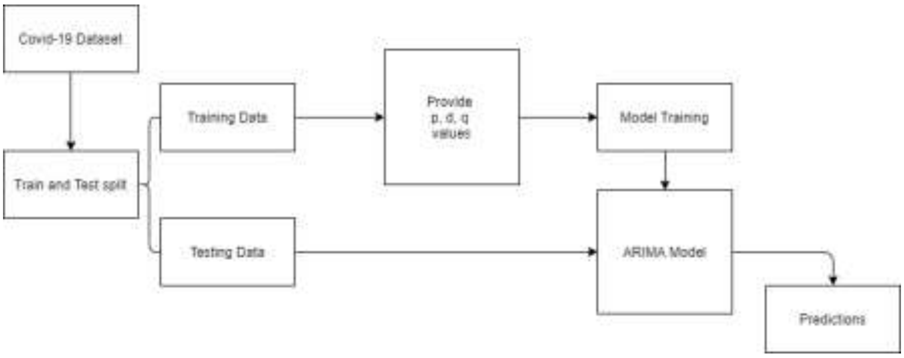


Fig 19. Flow diagram of ARIMA Model.

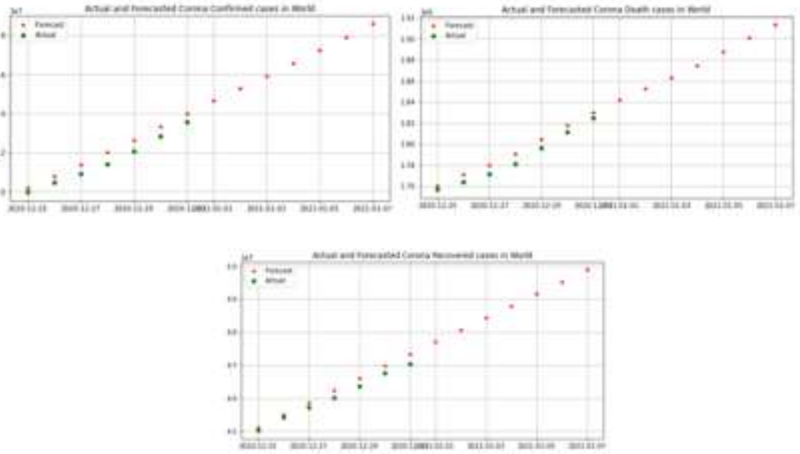


Fig 20. Comparison between the Covid 19 Cases and the ARIMA Predictions of the World.

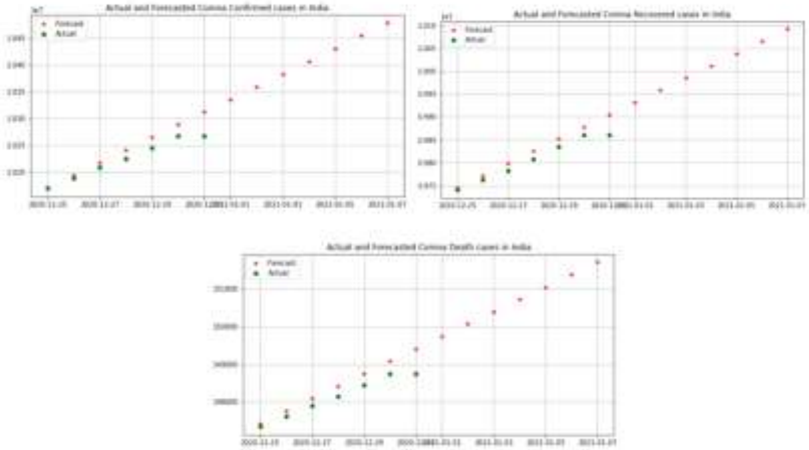


Fig 21. Comparison between the Covid 19 Cases and the ARIMA Predictions of India.

Table 4. Forecasts of the ARIMA Model for the World and India.

ARIMA	WORLD			INDIA		
Dates	**Confirmed**	**Recovered**	**Deaths**	**Confirmed**	**Recovered**	**Deaths**
1-1-2021	84140135	47295638	1834261	10273010	9872753	148951
1-2-2021	84652724	47579332	1842325	10279464	9885268	149165
1-3-2021	85148358	47864544	1850369	10286036	9897826	149379
1-4-2021	85715037	48154828	1861285	10292725	9910424	149593
1-5-2021	86380667	48443525	1875652	10299530	9923065	149807
1-6-2021	87093727	48733915	1890640	10306451	9935746	150022
1-7-2021	87775449	49023024	1903521	10313487	9948469	150236

8.5 FB Prophet Model

For an improper data FB Prophet model accurately analyzes and forecasts the time series data. The values in the dataset may contain yearly, monthly, daily, and weekly and may also contain holiday effects. The prophet model accurately works for the model which has historical data of several seasons. It is the best model to deal with data that is missing, handles outliers and trending shifts. Future values are equal to the average of historical data, for an average method.

$$y(t) = \mathrm{g}(t) + s(t) + h(t) + \varepsilon_t$$

$\mathrm{g}(t)$: linear or logistic curve that shows the changes in the time series

$s(t)$: periodic seasonality changes

$h(t)$: effects of holidays

ε_t: errors

Figure 22. illustrates the flow diagram of Prophet process to find the upcoming Covid-19 cases in the World and also India. Figure 23. depicts the comparison between the Covid 19 Cases and the Prophet Predictions of the World and Figure 24. depicts comparison between the Covid 19 Cases and the Prophet Predictions of India. Table 5. shows the forecasts of the Prophet Model for the World and India.

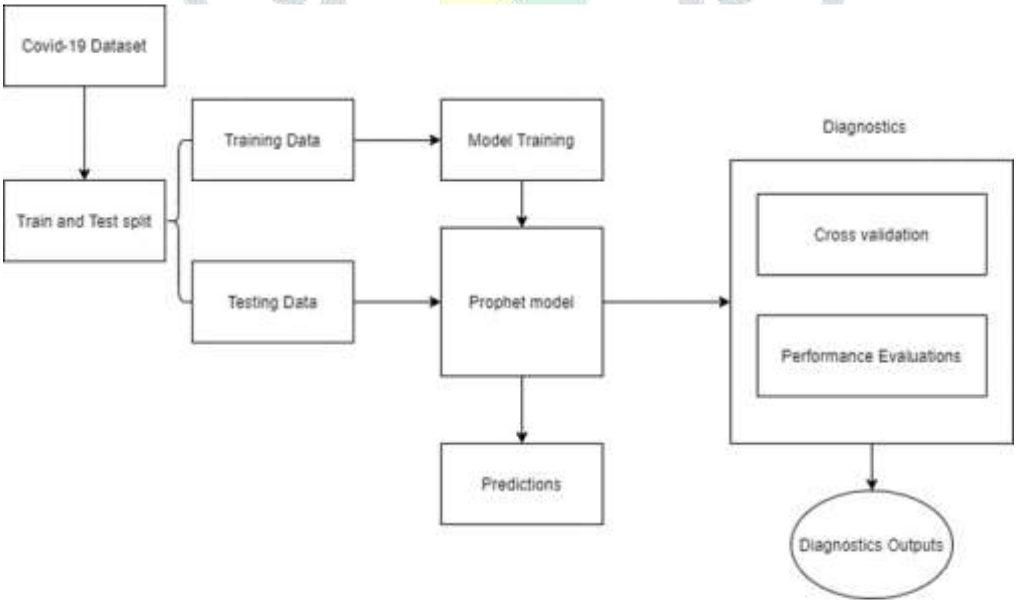


Fig 22. Flow diagram of Prophet Model.

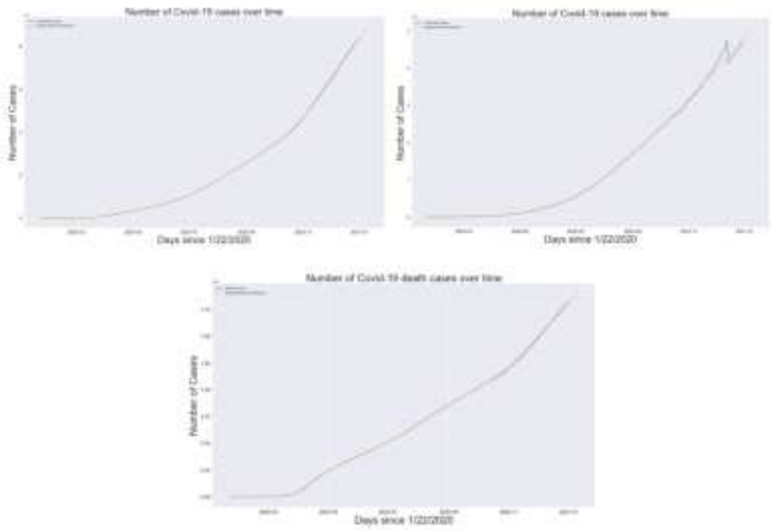


Fig 23. Comparison between the Covid 19 Cases and the Prophet Predictions of the World.

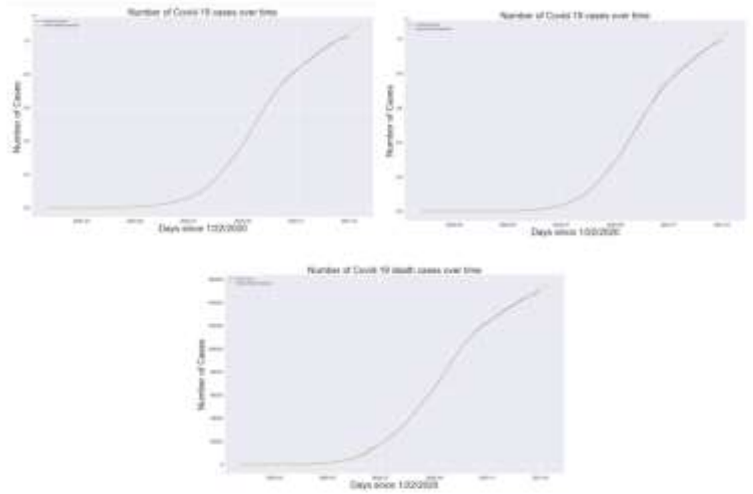


Fig 24. Comparison between the Covid 19 Cases and the Prophet Predictions of India.

Table 5. Forecasts of the Prophet Model for the World and India.

Prophet	WORLD			INDIA		
Dates	**Confirmed**	**Recovered**	**Deaths**	**Confirmed**	**Recovered**	**Deaths**
1-1-2021	83668757	48259357	1808546	10542594	10160076	151703
1-2-2021	84271639	48566448	1818127	10580837	10203043	152163
1-3-2021	84839702	48830975	1826517	10617123	10243881	152597
1-4-2021	85417339	48983943	1835599	10649703	10285849	153012
1-5-2021	86022783	49282256	1846404	10685289	10327488	153507
1-6-2021	86648687	49581612	1857297	10723240	10369590	153983
1-7-2021	87296610	49868073	1867947	10760075	10410616	154450

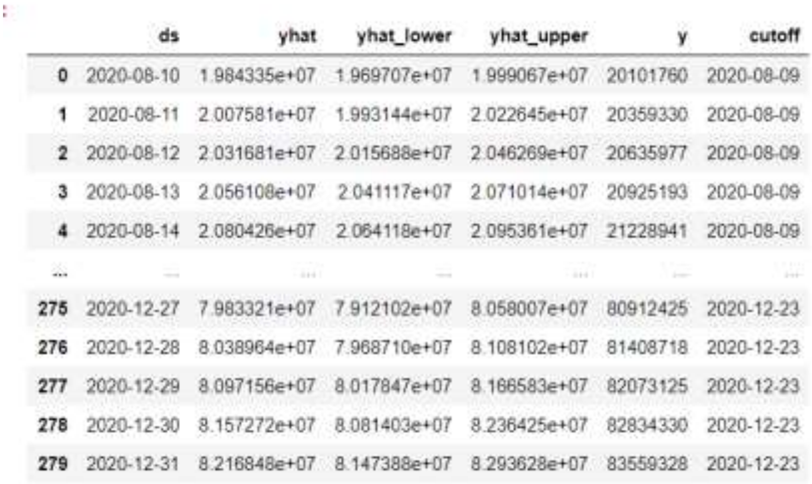
8.5.1. Diagnostics

Utilizing historical data Prophet measures prediction error using a functionality known as cross_validation. Cut-off points in the history are selected and for each of them and the model is fitted using data up to that cut-off point. Then the actual and predicted values are compared. The cross_validation function automatically validates for a series of historical cut-offs. By default, the primary training period is set to three times the horizon, and cut-offs are made every half a horizon. The forecast horizon, and the size of the primary training period (initial), and the cut-off dates spacing also known as period can be specified by the user.

Data frame with y and yhat which are true values and the forecasted values respectively along with simulated forecast date for each cut-off date is the output for cross_validation. For every observed point between cut-off and cut-off + horizon, a forecast is made. The data frame is used to compute the performance measures of yhat vs y.

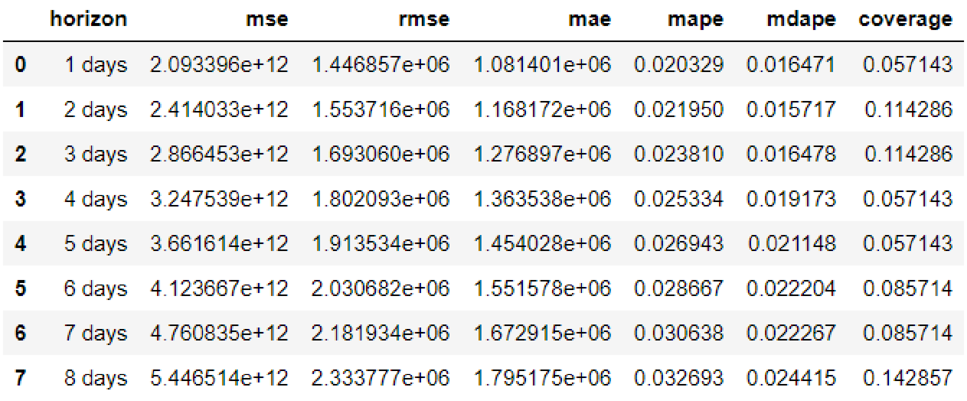
For instance, consider COVID-19 Cases of the World. Using the cross_validation of the diagnostics the FB Prophet makes 35 forecasts with cutoffs between 2020-08-09 00:00:00 and 2020-12-23 00:00:00. [cv = cross_validation(model = m, initial = '200 days', horizon = '8 days')].

Figures 25,26. Illustrates the Cross Validation and Performance Metrics for World-wide Covid-19 Cases. Figure 27. Depicts the Performance Metrics for World-wide Covid-19 Cases plotted individually.



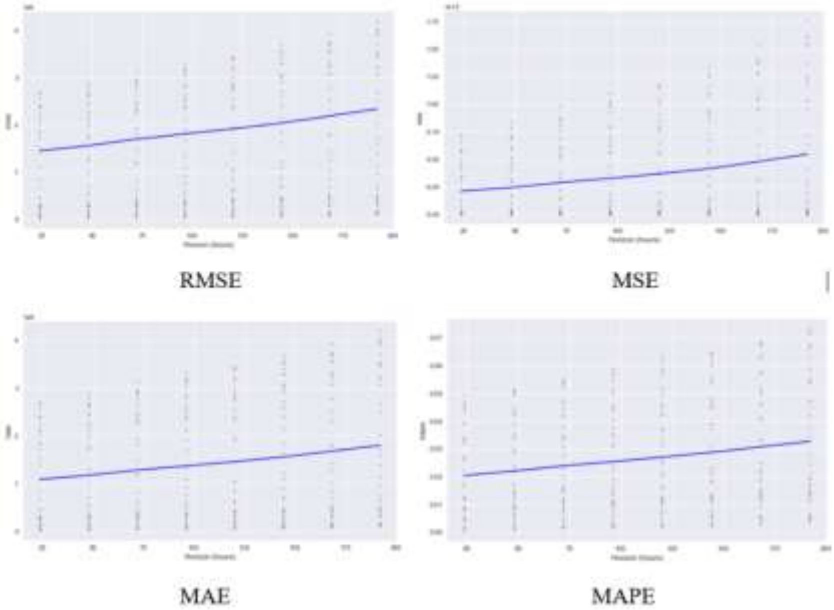
	ds	yhat	yhat_lower	yhat_upper	y	cutoff
0	2020-08-10	1.984335e+07	1.969707e+07	1.999067e+07	20101760	2020-08-09
1	2020-08-11	2.007581e+07	1.993144e+07	2.022645e+07	20359330	2020-08-09
2	2020-08-12	2.031681e+07	2.015688e+07	2.046269e+07	20635977	2020-08-09
3	2020-08-13	2.056108e+07	2.041117e+07	2.071014e+07	20925193	2020-08-09
4	2020-08-14	2.080426e+07	2.064118e+07	2.095361e+07	21228941	2020-08-09
...	...	...	...	...	...	...
275	2020-12-27	7.983321e+07	7.912102e+07	8.058007e+07	80912425	2020-12-23
276	2020-12-28	8.038964e+07	7.968710e+07	8.108102e+07	81408718	2020-12-23
277	2020-12-29	8.097156e+07	8.017847e+07	8.166583e+07	82073125	2020-12-23
278	2020-12-30	8.157272e+07	8.081403e+07	8.236425e+07	82834330	2020-12-23
279	2020-12-31	8.216848e+07	8.147388e+07	8.293628e+07	83559328	2020-12-23

Fig 25. Cross Validation for World-wide Covid-19 Cases.



	horizon	mse	rmse	mae	mape	mdape	coverage
0	1 days	2.093396e+12	1.446857e+06	1.081401e+06	0.020329	0.016471	0.057143
1	2 days	2.414033e+12	1.553716e+06	1.168172e+06	0.021950	0.015717	0.114286
2	3 days	2.866453e+12	1.693060e+06	1.276897e+06	0.023810	0.016478	0.114286
3	4 days	3.247539e+12	1.802093e+06	1.363538e+06	0.025334	0.019173	0.057143
4	5 days	3.661614e+12	1.913534e+06	1.454028e+06	0.026943	0.021148	0.057143
5	6 days	4.123667e+12	2.030682e+06	1.551578e+06	0.028667	0.022204	0.085714
6	7 days	4.760835e+12	2.181934e+06	1.672915e+06	0.030638	0.022267	0.085714
7	8 days	5.446514e+12	2.333777e+06	1.795175e+06	0.032693	0.024415	0.142857

Fig 26. Performance Metrics for World-wide Covid-19 Cases.



RMSE

MSE

MAE

MAPE

Fig 27. Performance Metrics for World-wide Covid-19 Cases plotted individually.

8.5.2. Trend Changepoints

Trend Changepoints are detected by the Prophet Model in two steps:

- Firstly, a large number of possible changepoints are defined at which there is a change in rate.
- Secondly, a sparse prior is kept on the rate change magnitudes.

Prophet includes a large number of changepoints which makes it a better model. By default there are twenty-five possible changepoints in Prophet and these are placed in the initial eighty percent of the series. n_changepoints is the argument that is set by using the number of potential changepoints and it can be better tuned by adjusting the regularization. To forecast the trend-forward and also to lessen the chances of overfitting at the end trend changepoints are suggested for the first eighty percent of the time series. By utilizing changepoint_range argument we can set the changepoint range as the default case will not work as expected in all situations.

For instance, m = Prophet(changepoint_range=0.9).

Figure 31. shows the trend changepoints for the Covid-19 cases in the World and Figure 32. Depicts the trends and weekly increase of Covid-19 cases in the World.

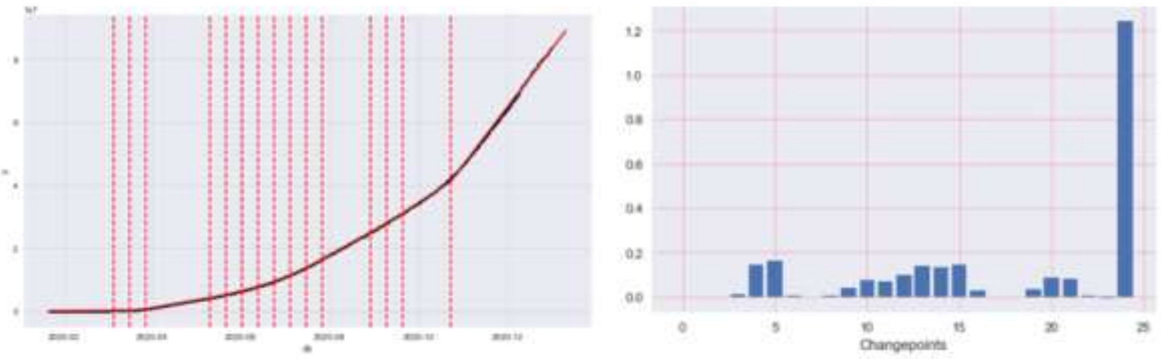


Fig 28. Trend Changepoints for Covid-19 cases in the World.

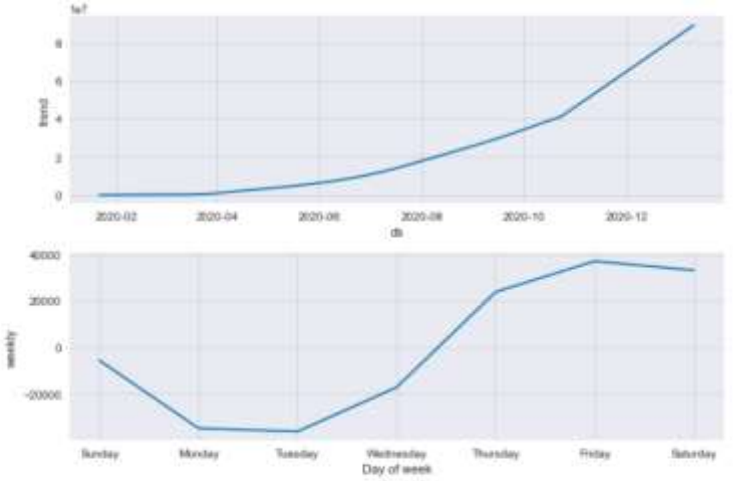


Fig 29. Trends and weekly increase of Covid-19 cases in the World.

9. Performance Measures

Using some metrics or combination of metrics we "measure" the objective performance of a regression model. This is known as Performance metrics or Performance evaluation. For this the forecasted data and the actual data are considered. Using them following performance evaluation metrics were performed, they are: RMSE, MAE, and MAPE. “Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are; RMSE is a measure of how spread out these residuals are. In other words, it tells you how concentrated the data is around the line of best fit” [12]. “The Mean Absolute Error (MAE) measures the average magnitude of the errors in a set of forecasts, without considering their direction. It measures accuracy for continuous variables. The equation is given in the library references. Expressed in words, the MAE is the average over the verification sample of the absolute values of the differences between forecast and the corresponding observation. The MAE is a linear score which means that all the individual differences are weighted equally in the average” [13]. “The mean absolute percentage error (MAPE) is a measure of how accurate a forecast system is. It measures this accuracy as a percentage, and can be calculated as the average absolute percent error for each time period minus actual values divided by actual values” [14]. The RMSE, MAE, and MAPE formulae respectively are as follows:

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2} \ , \frac{1}{n}\sum_{i=1}^{n}|y_i-\hat{y}_i| \ , \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i-\hat{y}_i}{y_i}\right|$$

Here y_i is the actual values, $\hat{y}$ is the forecasted values and n is the number of observations taken into consideration by the model.

Performance evaluation is done for the regression models that were constructed namely- Linear Regression Model, Support Vector Regression Model, Polynomial Regression Model, ARIMA Model, and Facebook Prophet Model. Table 6. illustrates the comparison of Performance metrics for all models.

10. Conclusion and Perspectives

We have used the graphs for analyzing the COVID-19 pandemic. We have predicted the values of confirmed, recovered, and deaths for the next 7 days in both World and India. Prediction is done using machine learning techniques and we also found the performance metrics of these techniques among them ARIMA and Prophet are the best. The prediction values are useful for the government to control the covid 19 pandemic and also for allocating medical resources as required. There is a huge spike in the mid-2020 and the cases were rising nonstop. Though there are many getting recovered from the virus simultaneously the death rate is also increasing. In order to stop this and reduce the cases rate Government should take necessary precautions and measures like imposing strict lockdowns where the Confirmed rate is high and use containment strategies wherever necessary. Even from the peoples’ side everyone must co-operate and work hand-in-hand with the Government by following WHO guidelines like Washing Hands, Using Masks and Sanitizers, etc.

Table 5. Performance metrics comparison chart of all models derived for both World and India.

Models	Cases	MAE	MAPE	RMSE
ARIMA Model	**Confirmed World**	436119.28	0.005332	453278.93
	Recovered World	179354.14	0.003873	195304.37
	Deaths World	6765.85	0.003787	7036.98
	Confirmed India	17006.42	0.001659	21606.77
	Recovered India	18099.42	0.001841	21600.96
	Deaths India	280	0.001886	332.34
Prophet Model	**Confirmed World**	1018737.39	0.012479	1024816.88
	Recovered World	1535355.31	0.033392	1541748.44
	Deaths World	38643.93	0.021597	39004.29
	Confirmed India	281824	0.027551	285190.07
	Recovered India	324862.3	0.033106	328445.8
	Deaths India	2806.74	0.018937	2849.03
Polynomial Model	**Confirmed World**	5661368.81	0.082993	7358477.54
	Recovered World	1244671.9	0.023058	1854483.79
	Deaths World	46062.39	0.025835	50503.74
	Confirmed India	1161487.81	0.084541	1530532.78
	Recovered India	729905.87	0.054346	809044.26
	Deaths India	26297.04	0.183498	33697.34
SVR Model	**Confirmed World**	6805598.8	0.12094	6920622.76
	Recovered World	5893172.14	0.150599	7114284.12
	Deaths World	635198.29	0.443585	677141.03
	Confirmed India	1207425.05	0.055557	1547005.62
	Recovered India	1075185.98	0.12956	1195094.31
	Deaths India	40692.93	0.292112	50979.94
Linear Model	**Confirmed World**	23845556.05	0.391486	26293846.37
	Recovered World	14577022.01	0.391337	15329638.55
	Deaths World	211076.43	0.14246	243639.46
	Confirmed India	4090609.29	0.461662	4123672.78
	Recovered India	4533621.41	0.549306	4603528.26
	Deaths India	46593.57	0.357575	46768.23

12. References

[1] "Predicting the COVID-19 Prevalence Rate Using Data Mining" By Fatemeh Ahouz, Behbahan Khatam Alanbia University of Technology, Behbahan, Iran & Amin Golabpour, Shahrood University of Medical Sciences, Iran. https://www.researchsquare.com/article/rs-21247/v1

[2] "Predictive Data Mining Models for Novel Coronavirus (COVID-19) Infected Patients' Recovery" By L. J. Muhammad & Sani Sharif Usman, Department of Mathematics and Computer Science, Faculty of Science, Federal University of Kashere, P.M.B. 0182, Gombe, Nigeria. Md. Milon Islam & Safial Islam Ayon, Department of Computer Science and Engineering, Khulna University of Engineering & Technology, Khulna, 9203, Bangladesh. 21th June, 2020. https://doi.org/10.1007/s42979-020-00216-w

[3] "Data Modelling & Analysing Coronavirus (COVID19) Spread using Data Science & Data Analytics in Python Code" By Jatin Chaudhary https://in.springboard.com/blog/data-modelling-covid/

[4] "COVID-19 DATA ANALYSIS" By Ashutosh Kumar, M. tech (1st Year), Computer Science and Engineering, National Institute of Technology, Jamshedpur. https://www.researchgate.net/publication/342277441_COVID19_DATA_ANALYSIS_httpsgithubcomashukumar7Covid19

[5] Arun, Shreyas & Iyer, Ganesh. (2020). On the Analysis of COVID19 - Novel Corona Viral Disease Pandemic Spread Data Using Machine Learning Techniques. 1222-1227.10.1109/ICICCS48265.2020.9121027. https://www.researchgate.net/publication/342325703_On_the_Analysis_of_COVID19__Novel_Corona_Viral_Disease_Pandemic_Spread_Data_Using_Machine_Learning_Techniques

[6] “Real-time epidemic forecasting for pandemic influenza” By Hall, Gani, Hughes, and Leach. https://pubmed.ncbi.nlm.nih.gov/16928287/

[7] “COVID-19 and Italy: what next?” By Andrea Remuzzi and Giuseppe Remuzzi, published on March 13, 2020. https://www.thelancet.com/article/S0140-6736(20)30627-9/fulltext

[8] “Real-time forecasts of the COVID-19 epidemic in China from February 5th to February 24th, 2020” By Roosa K., Tariq A., Yan P., Hyman J.M., Chowell G. https://www.sciencedirect.com/science/article/pii/S2468042720300051#!

[9] “Application of the ARIMA model on the COVID-2019 epidemic dataset” By Benvenuto, Giovanetti, Vassallo, Angeletti, and Ciccozzi. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7063124/

[10] For Libraries:

- Pandas - https://pandas.pydata.org/pandas-docs/stable/development/internals.html
- Plotly - https://plotly.github.io/plotly.py-docs/generated/plotly.html
- Folium - https://pypi.org/project/folium/0.1.5/

[11] For Data Sets:

- https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series
- https://github.com/laxmimerit/Covid-19-Preprocessed-Dataset

[12] RMSE (Root Mean Squared Error)- https://www.statisticshowto.com/probability-and-statistics/regression-analysis/rmse-root-meansquareerror/#:~:text=Root%20Mean%20Square%20Error%20(RMSE)%20is%20the%20standard%20deviation%20of,the%20line%20of%20

[13] MAE (Mean Absolute Error)- https://medium.com/human-in-a-machine-world/mae-and-rmse-which-metric-is-better-e60ac3bde13d

[14] MAPE (Mean Absolute Percentage Error)- https://www.statisticshowto.com/mean-absolute-percentage-error mape/#:~:text=The%20mean%20absolute%20percentage%20error,values%20divided%20by%20actual%20values