Prediction of COVID-19 Infection Using Epidemiology Dataset

Ravi Satvik Gorthi, Omkar Chavan, Dev Patel, Smit Gandhi
Dhirubhai Ambani Institute of Information and Communication Technology,
Gujarat, India

***Abstract*- In this paper we are predicting and forecasting the COVID-19 outbreak in World on machine learning approach, the aim of the project is to provide data analysis of covid-19 (a pandemic started in December 2019) using different kind of ML models such as SVM, polynomial regression and Prophet. Using those three models we predicting conform, death and recover cases for the World and we take one country which is India. Through plotting of data, various cases have been studied like most affected countries due to this pandemic as well as most affected state in India. In this paper we are using Time series dataset provided by Johns Hopkins University. The model is predicting the number of confirmed, recover and death cases based on data available from 22nd January, 2020 till up to today. In this project, the predictions on various cases have been done and finally, the accuracy of the algorithm has been determined. Comparison graphs has also been plotted to analyses how much World and INDIA is getting affected/recover day by day.**

Keywords – World: India: COVID-19: Machine learning: Polynomial regression model: SVM prediction model: Prophet library: time series forecasting:

I. INTRODUCTION

The year 2020 has been a disastrous year for humankind. We humans, all around the globe have come across the Coronavirus. It was first identified in December 2019 in Wuhan, China. The World Health Organization declared the outbreak a Public Health Emergency of International Concern on 20 January 2020, and later a pandemic on 11 March 2020. As of 3 April 2021, more than 130 million cases have been confirmed, with more than 2.84 million deaths attributed to COVID-19, making it one of the deadliest pandemics in history. India witnessed an outbreak of COVID-19, during the last week of January 2020 when a few Indian students travelled to Kerala from Wuhan located in China. In 2020, from January to till today, we have not been able to get rid of the virus. As per the World Health Organization (WHO), numerous potential COVID-19 antibodies are being examined, and many voluminous clinical trials may report their results later at the near end of 2020 or the very beginning of 2021. WHO is working with partners around the world to help coordinate with the key steps in this process Companies such as Pfizer and Biotech have concluded a phase 3 study of the COVID-19 vaccine and claim to be 95% sufficient against the virus. How the epidemic in India will top or decrease is foremost concerning the issue. Therefore, it is pivotal to predict the trends of the pandemic, nationwide. With this view of helping the Government, we undertook this research to aid them in making informed decisions about the spread of coronavirus thereby taking precautionary measures. For this, we have analysed Johns Hopkins University Time series dataset using Polynomial regression model, Support vector machine and Prophet we predicted confirm cases , death cases , recovery cases of World as well as for India with proper visualization on world map and India map.

II. METHODOLOGY

This section discusses the different methods applied to world's dataset and India's dataset for COVID-19 to analysis, prediction, and forecasting of different cases. Fig. 1 shows the flowchart of our methodology which includes data collection, followed by data pre-processing, data visualization, implementation of different model models, time series forecasting approach, and their results.

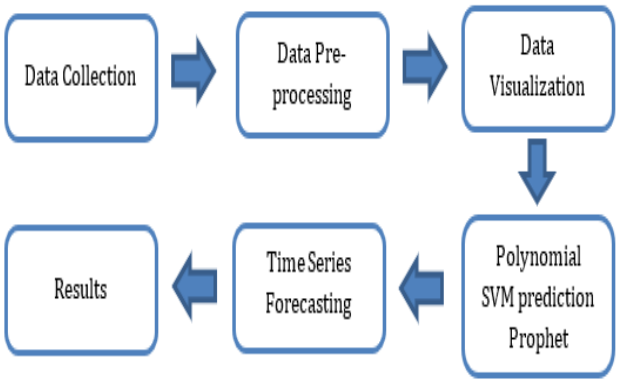

Fig1 Flowchart for the implemented methodology

A. Data Collection

The data for the ongoing Covid-19 outbreak in world as well as India is collected from [5]. The columns of this dataset include the Total number of Confirmed, Active, Cured, and Death cases of Covid-19 patients accumulating all the states with countries on a day-to-day basis from 22th January 2020 to last date of toady.

B. Data Pre-processing

In the section of data pre-processing, redundant or null values were removed by data cleaning. Further, we have set the column attributes as -"Confirmed, Active, Cured and Deaths cases" for which the dependent variable and "Dates" as the independent variable. To achieve this target, the data was then split into 75% for training purpose and 25% for testing purpose. Standardization of the variables pertinent to training and testing was done using the train_test_split() function and the fit_transform() function, the object was fit into data to transform these values into standard form.

C. Data Visualization

Figures 2 and 3 represent Heat Maps of data and information to supply an open way to see and get trends, exceptions, and patterns in information. A Heat Map visualization could be combination of coloured rectangles, each representing a quality component that permits clients to rapidly get a handle on the state and effect of an expansive number of factors at one time. For example, Maharashtra state has the highest number of cases which is shown by the high intensity of red coloration, whereas for Lakshadweep state, colour intensity is the least.

	Country Name	Number of Confirmed Cases	Number of Deaths	Number of Recoveries	Number of Active Cases	Mortality Rate
0	US	3060690	554103	0.000000	3055587.000000	0.018102
1	Brazil	12910082	328206	11277632.000000	1304244.000000	0.025422
2	India	12392260	164110	11569241.000000	658909.000000	0.013243
3	France	4902457	96438	304563.000000	4401456.000000	0.020081
4	Russia	4511973	97986	4138458.000000	275529.000000	0.021717
5	United Kingdom	4367969	127058	13190.000000	4227721.000000	0.029089
6	Italy	3629000	110328	2963377.000000	565295.000000	0.030402
7	Turkey	3400296	31892	3059462.000000	308942.000000	0.009379
8	Spain	3291394	75541	150376.000000	3065477.000000	0.022951
9	Germany	2682356	76940	2566530.000000	238886.000000	0.026693
10	Colombia	2428048	63777	2300887.000000	63384.000000	0.026267
11	Poland	2387511	54165	1911249.000000	422097.000000	0.022687
12	Argentina	2373153	56023	2121954.000000	195176.000000	0.023607
13	Mexico	2247357	203854	1765244.000000	278259.000000	0.090706
14	Iran	1908974	62876	1633949.000000	212149.000000	0.032937
15	Ukraine	1762713	35326	1381163.000000	346224.000000	0.020041
16	Peru	1568345	52331	1485582.000000	30432.000000	0.033367
17	South Africa	1550724	52946	1475398.000000	22380.000000	0.034143
18	Czechia	1545865	26765	1374823.000000	144277.000000	0.017314
19	Indonesia	1523179	41151	1361017.000000	121011.000000	0.027017
20	Netherlands	1308774	16733	16446.000000	1275595.000000	0.012785
21	Chile	1011485	23421	942413.000000	45651.000000	0.023155
22	Canada	998208	23011	924144.000000	51053.000000	0.023052

Fig 2 hit map of world covid-19

	Province_State	Last_Update	Confirmed	Deaths	Recovered	Active	Incident_Rate	Case_Fatality_Ratio
48	Andaman and Nicobar Islands	2021-04-03 04:20:44	6084	62	4981.000000	41.000000	1219.079408	1.215512
49	Andhra Pradesh	2021-04-03 04:20:44	904548	7225	888508.000000	8815.000000	1678.091025	0.798741
50	Arunachal Pradesh	2021-04-03 04:20:44	16846	56	16785.000000	5.000000	1072.680708	0.332423
51	Assam	2021-04-03 04:20:44	218533	1107	215517.000000	1909.000000	613.735391	0.506560
52	Bihar	2021-04-03 04:20:44	266677	1580	262733.000000	2364.000000	213.683620	0.592477
53	Chandigarh	2021-04-03 04:20:44	57543	381	24064.000000	3098.000000	2377.526278	1.383292
54	Chhattisgarh	2021-04-03 04:20:44	557976	4247	321873.000000	31858.000000	1216.113571	1.186395
55	Dadra and Nagar Haveli and Daman and Diu	2021-04-03 04:20:44	9703	2	3499.000000	202.000000	601.405825	0.054010
56	Delhi	2021-04-03 04:20:44	568814	11060	545770.000000	11994.000000	3674.457742	1.652178
57	Goa	2021-04-03 04:20:44	58564	632	55838.000000	1914.000000	3693.238771	1.420163
58	Gujarat	2021-04-03 04:20:44	312748	4539	294650.000000	13559.000000	489.644987	1.451328
59	Haryana	2021-04-03 04:20:44	294270	3174	280074.000000	11022.000000	1043.337045	1.078601
60	Himachal Pradesh	2021-04-03 04:20:44	64420	1056	60026.000000	3338.000000	864.471135	1.639242
61	Jammu and Kashmir	2021-04-03 04:20:44	131936	2003	126720.000000	3215.000000	969.681736	1.518137
62	Jharkhand	2021-04-03 04:20:44	125585	1115	120562.000000	3908.000000	325.400750	0.887845
63	Karnataka	2021-04-03 04:20:44	1006229	12591	959400.000000	34238.000000	1489.326520	1.251306
64	Kerala	2021-04-03 04:20:44	1129890	4646	1098526.000000	26718.000000	3165.007364	0.411190
65	Ladakh	2021-04-03 04:20:44	10189	130	9781.000000	278.000000	3714.695084	1.275886
66	Lakshadweep	2021-04-03 04:20:44	733	1	690.000000	42.000000	1137.686446	0.136426
67	Madhya Pradesh	2021-04-03 04:20:44	300634	4014	277484.000000	19936.000000	382.433983	1.334291
68	Maharashtra	2021-04-03 04:20:44	2904076	55379	2457494.000000	391203.000000	2398.272219	1.906940
69	Manipur	2021-04-03 04:20:44	29406	374	26975.000000	57.000000	951.174898	1.271849
70	Meghalaya	2021-04-03 04:20:44	14082	150	13868.000000	64.000000	418.271844	1.065190
71	Mizoram	2021-04-03 04:20:44	4487	11	4436.000000	40.000000	362.075586	0.245153
72	Nagaland	2021-04-03 04:20:44	12361	92	12138.000000	131.000000	549.452259	0.744276
73	Odisha	2021-04-03 04:20:44	341772	1921	337430.000000	2421.000000	737.271416	0.562071

Fig 2 hit map of India covid-19

(D) Polynomial regression model

We have employed the model of polynomial regression provides the relationship between the dependent variable Y and the independent variable X and is modelled as 2nd, 3rd, 4th, and 5th-degree polynomial in x. here we choose fix degree because of large covid-19 time series dataset world and India .The least-squares method is used while fitting these models. Using this method helps to minimize the fluctuation of the fair estimators of the coefficients. In general, we can demonstrate the anticipated value of y as an nth degree polynomial, generating the standard polynomial regression model.

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \dots\dots\dots\dots + \beta_n x^n + \varepsilon$$

where ε is an unexpected error with mean zero conditioned on a scalar variable x, β_0 is a constant, and β_1 to β_n are coefficients [6].

(E) SVM model

The second type of regression analysis that we have used is SVM (Support Vector Machines) time series prediction model. SVM are used for time series prediction and compared to radial basis function networks. we make use of two different cost functions for Support Vectors: training with (i) an ε insensitive loss and (ii) Huber's robust loss function and discuss how to choose the regularization parameters in these models. two applications are considered: data from (a) a noisy Mackey-Glass System (normal and uniform noise and (b) the Santa Fe Time Series Competition set D In both cases, Support Vector Machines show an excellent performance. In case (b), the Support Vector approach improves the best known result on the benchmark. Here we use direct library of SVM and fit the line and predict covid -19 confirm, death, recovery cases.[7]

(F) Prophet model

Time series forecasting is the use of a model to predict future values based on previously observed values. Models for time series data can have many forms and represent different stochastic processes. Prophet is an open-source tool from Facebook used for forecasting time series data which helps businesses understand and possibly predict the market.

It is based on the decomposition (trend + seasonality + holidays) models available in Python and R. It provides us with the ability to make time series predictions with good accuracy using simple intuitive parameters. Prophet follows the sklearn model API. We create an instance of the Prophet class and then call its fit and predict methods. [8]

(G) Error Analysis

We calculated Mean Absolute Error (MAE) and Mean Squared Error (MSE) is an important method in statistics that measures the prediction accuracy of forecasting. For instance, it is used as a loss function for regression problems in trend estimation. R-squared (R) is a mathematical measurement that speaks to the extent of the fluctuation for a subordinate variable that's clarified by a variable or variables in a relapse demonstrate. The goodness of fit of a model could be measured using the R squared score.[9]

III. EXPERIMENTAL RESULTS

The data for the models have been taken from 'Johns Hopkins University Coronavirus Data Stream' that combines World Health Organization (WHO) and Centre for Disease Control and Prevention (CDC) case data. For the exponential model, the data between March 11 and March 23 were used, when the number of reported infections were 62 and 499 respectively. For the SIR model, a longer range is required to obtain a reasonable estimate, and hence data for 21 days were considered starting from March 10, which is designated as the seed value. On this date, there were 56 individuals who had contracted the virus, out of which 39 were travel-related (stage-1) and 17 were person-to-person (stage-2) transmissions. There is no confirmed report of community transmission as of March 30. Hence, we consider that models with these seed values will give a good estimate for stage-1 and 2 transmissions. Further, like the study of Singh and Adhikari, we consider all cases to be symptomatic as it is not easy to estimate the number of asymptomatic cases. This could significantly underestimate the actual numbers of cases. Before showing the models for India, we compare the growth of infections with several different countries and states in the US. There have been several reports that have indicated that the initial slow growth of infections in India could be an artifact of its testing strategy, where testing is limited to specific individuals travelling from high-risk countries and their immediate contacts. We visualize the growth of infections in India from 1 to 1000, along with other countries and states with a reasonable number of daily international travellers.

The growth rate in India is much smaller than places like New York and New Jersey, where the spread is very fast and it took only 16-17 days to reach 1,000 cases. Italy and France took about 29 and 43 days respectively to reach the same figure. On the other hand, places like California and Washington took a relatively long time (about 55-58 days), which is similar to India that took 59 days. Further, the growth curve of India is very close to that of Washington. As India is 9 days behind Washington in outbreak history, this information could be very useful as one may look at the Washington data to make predictions for India.

We tried to predict the future of Covid-19 for world data using a polynomial regression model which performed with a Mean Average Error (MAE) of 9,822,335 and a regression score (r2) of 0.7168 for predicting the number of worldwide confirmed cases. The regression model performed with a MAE of 318,032.3326 and a r2 score of -0.9921284 for predicting the number of worldwide deaths which indicates that the model was arbitrarily worse. The polynomial regression model had a MAE of 9901995.3071 and a r2 score of 0.37817 in predicting the number of worldwide recoveries. Following the regression model, we tried to predict the future of Covid- 19 for world data using a SVM model which performed with a Mean Average Error (MAE) of 13,508,666.5274 and a regression

score (r2) of 0.56625 for predicting the number of worldwide confirmed cases. The SVM model performed with a MAE of 115,721,428.1219 and a r2 score of -16.3747 for predicting the number of worldwide deaths which indicates that the model was arbitrarily worse. The model had a MAE of 59,274,349.1673 and a r2 score of -3.7675 in predicting the number of worldwide recoveries. Hence we may conclude that the polynomial regression model was more accurate than the SVM model for predicting the worldwide data.

Further which, we predicted the future of Covid- 19 in India using two time series prediction models: Polynomial Regression and Support Vector Machine. The polynomial regression model performed with a MAE of 68,709,403.6491 for the number of confirmed cases, a MAE of 6,29,437.2952 for the number of deaths, and a MAE of 57,131,873.1565 for the number of recoveries. The Support Vector Machine (SVM) model performed with a MAE of 8,384,137.6358 for the number of confirmed cases, a MAE of 19,158,374.8358 for the number of deaths, and a MAE of 8,784,323.6995 for the number of recoveries.

World Time Series Prediction

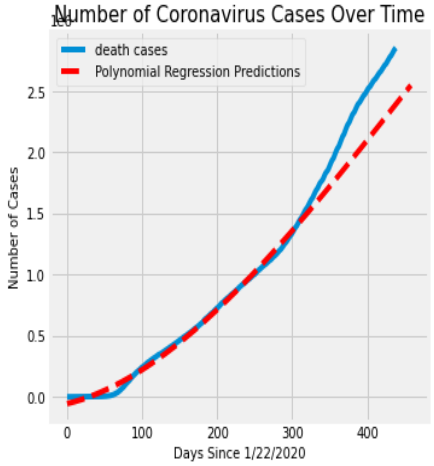

Fig 3.1 Death cases Prediction By Polynomial Reg.

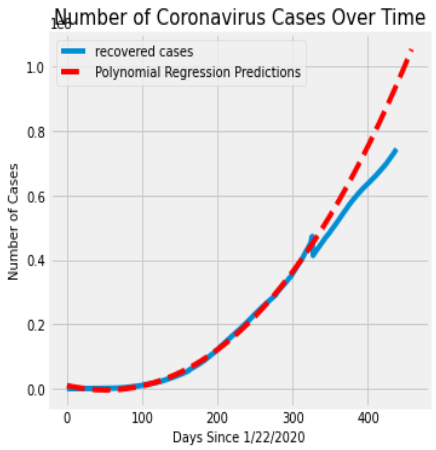

Fig 3.2 Recovery cases Prediction By Polynomial Reg.

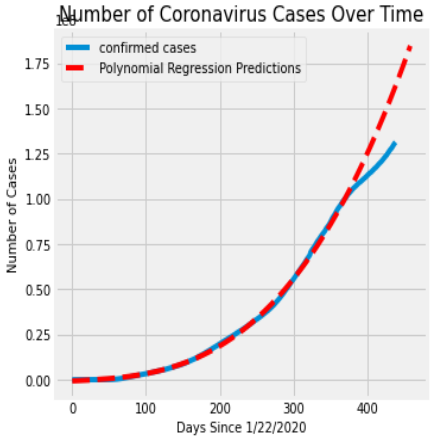

Fig 3.3 Confirmed cases Prediction By Polynomial Reg.

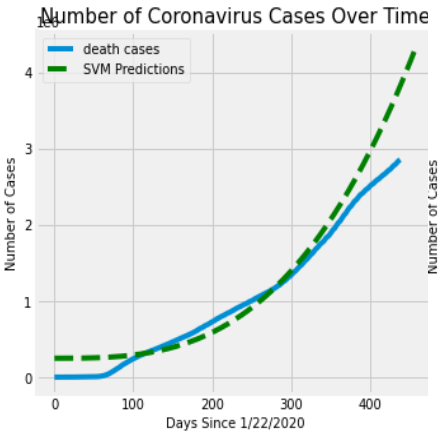

Fig 3.4 Death cases Prediction By SVM model.

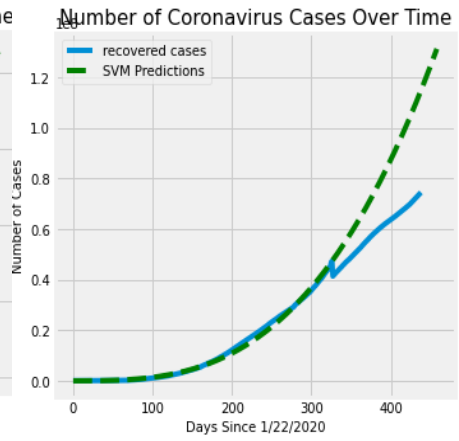

Fig 3.5 Recovery cases Prediction By SVM model.

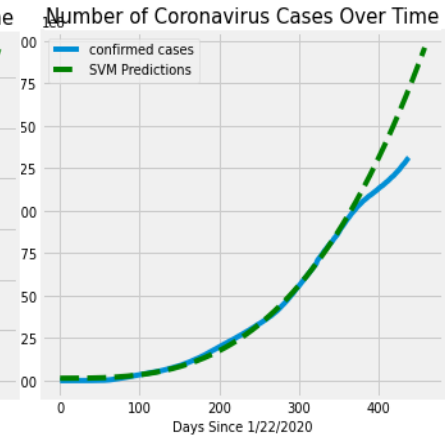

Fig 3.6 Confirmed cases Prediction By SVM model.

India Time Series Prediction

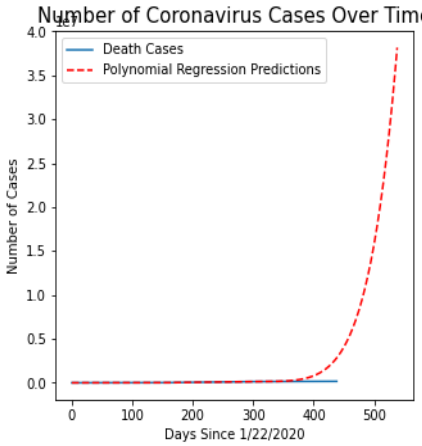

Fig 3.7 Death cases Prediction By Polynomial Reg.

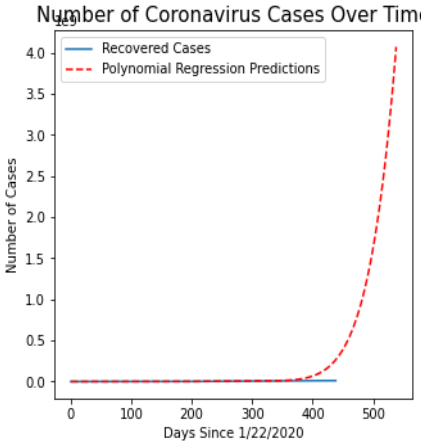

Fig 3.8 Recovery cases Prediction By Polynomial Reg.

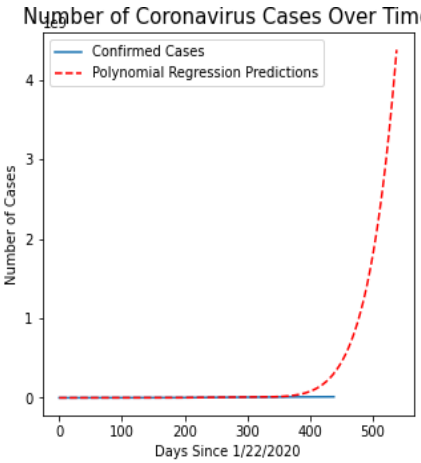

Fig 3.9 Confirmed cases Prediction By Polynomial Reg.

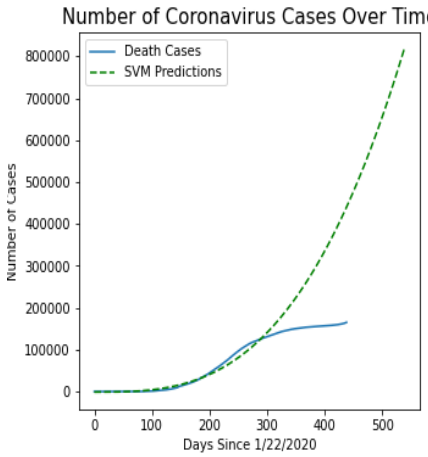

Fig 3.10 Death cases Prediction By SVM model.

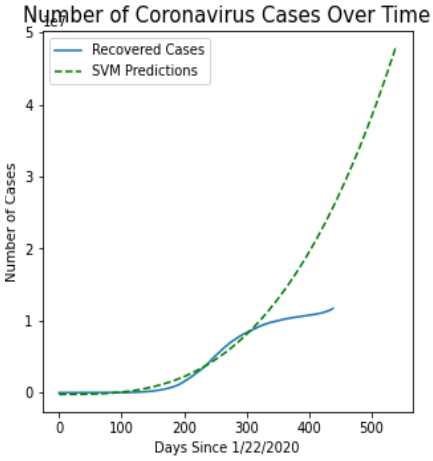

Fig 3.11 Recovery cases Prediction By SVM model.

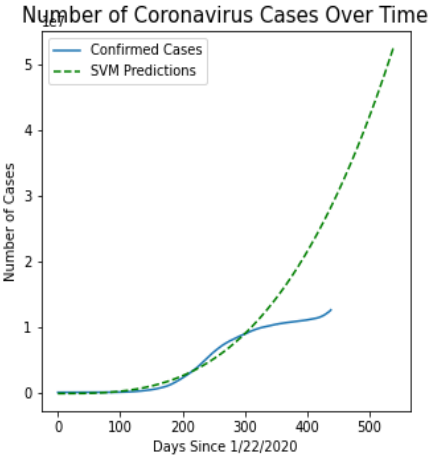

Fig 3.12 Confirmed cases Prediction By SVM model.

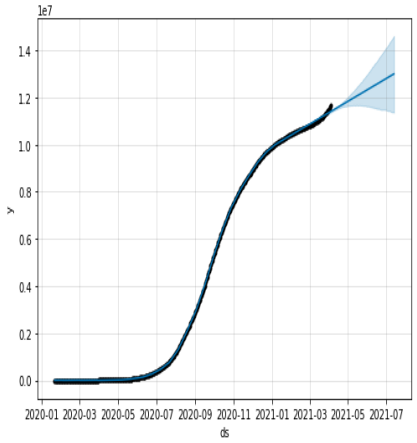

Fig 3.13 Death cases Prediction By Prophet model.

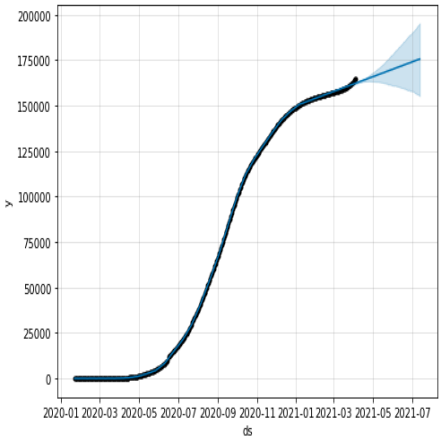

Fig 3.14 Recovery cases Prediction By Prophet model.

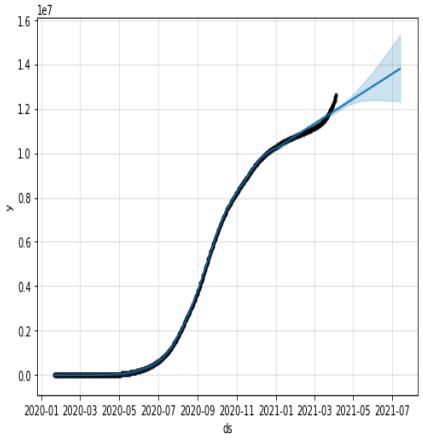

Fig 3.15 Confirmed cases Prediction By Prophet model.

The state-of-the-art model we used for comparing our models is Prophet which is a Facebook's open source project for time series prediction. The FBProphet model performed with a MAE of 1,680,490 for the number of confirmed cases, a MAE of 20817.8 for the number of deaths, and a MAE of 2,022,180 for the number of recoveries.

IV. CONCLUSION

In the present study, we conducted an experimental study in the forecasting of the COVID-2019 epidemic pattern and have also compared the differences of actual and predicted values in both principle and practical aspects. Moreover, based on weighted overlay, the district is classified into a very high, high, medium and low risk zone of COVID-2019. The Prophet model can acquire past values and consider current and preceding residual series historical knowledge. An efficient linear model to efficiently capture a linear pattern of the COVID-19 disease series was demonstrated in the Prophet model. In general, decomposition methods operate best when the sequence is compatible with the hypothesis for decomposition. The drawback of the model is that only the data from the time series can derive linear relationships. With events which may be influenced by multiple factors, including several meteorological and specific social influences, this does not work well. When used in other cases, the findings based on a particular disease may not be repeatable. Moreover, there are several other theories about the long-term trend in methods of decomposition, such as generalized models and Support Vector Machine (SVM), which assume a nonlinear function in the time series. Hence we may conclude that the SVM model performed better than the Polynomial regression model when predicting the future of Covid- 19 in India. Whereas the polynomial regression model performs better than the SVM model for prediction when considering the world data for Covid.

V. REFERENCES

[1] "An epidemiological modelling approach for COVID-19 via data assimilation" , Philip Nadler, Shuo Wang, Rossella Arcucci, Xian Yang & Yike Guo.

[2] "Prediction of epidemic trends in COVID-19 with logistic model and machine learning techniques" Peipei Wanga, Xinqi Zhengab, Jiayang Lia, BangrenZhua.

[3] "Rapid implementation of mobile technology for real-time epidemiology of COVID-19" David A. Drew1, Long H. Nguyen1, Claire J. Steves, Cristina Menn, Maxim Freydin, Thomas Varsavsky, Carole H. Sudre, M. Jorge Cardoso, Sebastien Ourselin, Jonathan Wolf, Tim D. Spector, Andrew T. Chan, COPE Consortium.

[4] "Real-time forecasts of the COVID-19 epidemic in China from February 5th to February 24th, 2020", K. Roosaa, Y .Leea, R. Luoa, A.Kirpicha, R. Rothenberg, J. M. Hyman, P.Yanc, G.Chowella.

[5] "time series dataset", Available: https://github.com/CSSEGISandData/COVID-19

[6] "polynomial regression", Available: https://en.wikipedia.org/wiki/Polynomial_regression

[7] "SVM(Support Vector Machines) time series prediction model", Available: https://www.sciencedirect.com/science/article/abs/pii/S0169743903001114

[8] "Prophet model Time series forecasting", Available: https://medium.com/analytics-vidhya/time-series-forecasting-of-covid19-data-with-fbprophet-3a73dba59106

[9] "Mean Absolute Error (MAE) and Mean Squared Error (MSE)", Available: https://www.bing.com/search?q=polynomial+regression+mae+mase++function&qs=n&form=QBRE&sp=-1&pq=polynomial+regression+mae+mase+function&sc=0-

39&sk=&cvid=BC3EC8D190B0456C8FE1FB2E826AB6C8

[10] Harapan Harapan , Naoya Itoh, Amanda Yufika, Wira Winardi, Synat Keam "Coronavirus disease 2019 (COVID-19): A literature review" by Journal of Infection and Public Health Volume 13, Issue 5, May 2020

[11] "Prediction of COVID-19 coronavirus pandemic based on time series data using a support vector machine". Vijander Singh,Ramesh Chandra Poonia ,Sandeep Kumar,Pranav Dass,Pankaj Agarwal,Vaibhav Bhatnagar &Linesh Raja