2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence) | 978-1-6654-1451-7/20/$31.00 ©2021 IEEE | DOI: 10.1109/Confluence51648.2021.9377137

Analysis and Prediction of COVID-19 using Regression Models and Time Series Forecasting

[1]Saud Shaikh, [2]Jaini Gala, [1]Aishita Jain, [1]Sunny Advani, [1]Sagar Jaidhara, [1]Dr. Mani Roja Edinburgh
Electronics and Telecommunication Department
[1]Thadomal Shahani Engineering College, Mumbai, India
[2]K.J. Somaiya College of Engineering, Mumbai, India

***Abstract*— In this paper, we are predicting and forecasting the COVID-19 outbreak in India based on the machine learning approach, where we aim to determine the optimal regression model for an in-depth analysis of the novel coronavirus in India. We are implementing the two regression models namely linear and polynomial and evaluating the two using the R squared score and error values. The COVID-19 dataset for India is being used to serve the research of this paper. The model is predicting the number of confirmed, recovered, and death cases based on the data available from March 12 to October 31, 2020. For forecasting the future trend of these cases, we are utilizing the time series forecasting approach of tableau. Furthermore, the time series forecasting method is being employed to forecast the total count of confirmed cases in the future.**

Keywords*— *India; COVID-19; machine learning; regression models; linear regression; polynomial regression; time series forecasting; tableau

I. INTRODUCTION

The year 2020 has been a disastrous year for humankind. We humans, all around the globe have come across the Coronavirus. India witnessed an outbreak of COVID-19, during the last week of January 2020 when a few Indian students traveled to Kerala from Wuhan located in China. In 2020, from January to November to date, 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 [1]. WHO is working with partners around the world to help coordinate with the key steps in this process. Companies such as Pfizer and BioNTech have concluded a phase 3 study of the COVID-19 vaccine and claim to be 95% efficient against the virus [2]. 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 analyzed India's COVID-19 dataset using regression models with supporting empirical evidence including error analysis and accuracy juxtapositions. Also, we have forecasted the trend of coronavirus cases using the Time Series Forecasting approach of Tableau. These methods are applied for four different types of cases: Confirmed, Active, Cured, and Death Cases as available in [3].

II. LITERATURE REVIEW

In [4], the outbreak of COVID-19 in India is anticipated by Sunita Tiwari, Sushil Kumar, and Kalpna Guleria based on the Chinese pattern of employing a Machine Learning approach. The predicted no. of confirmed cases, recovered cases, and death cases support the data available between January 22 and April 3, 2020, by employing statistical forecasting. According to them, the effects of the virus in India were expected to be high between the 3rd and 4th weeks of April 2020 and controlled towards the end of May 2020. The number of predicted confirmed cases was expected to reach 68978, and thus the total deaths to be 1557 by April 25, 2020.

In [5], Peipei Wang *et al.* have coordinated the COVID-19 information before June 16, 2020, into a calculated model to fit the cap of the epidemic trend, and after that bolster the cap value into the FbProphet model to determine the epidemic bend and anticipate its trend. The modeling results for entire Brazil, Russia, India, Peru, and Indonesia have summarized in 3 points namely the epidemic peak point, fastest-growing point, and turn point.

In this paper [6], Milind Yadav *et al.* have unraveled 5 distinctive errands such as anticipating the spread of coronavirus over districts, analyzing the development rates and the kinds of mitigation over nations, anticipating the epidemic's conclusion, analyzing the transmission rate of the infection, and relating it with climate conditions. They proposed an SVR method to analyze these tasks associated with the novel coronavirus. They utilized the supported vectors to encourage way better classification precision. The results demonstrate its prevalence in both proficiency and exactness.

In [7], Anuradha Tomar and Neeraj Gupta have utilized data-driven estimation strategies like LSTM and curve fitting for the expectation of the number of COVID-19 cases in India and the impact of preventive measures like social separation and lockdown on the spread of the same. The forecast of different parameters (no. of positive cases, no. of recovered cases, etc.) gotten by the proposed strategy is precise.

In [8], Deep Learning models are utilized for foreseeing the no. of coronavirus positive cases in India by Debanjan Parbat and Monisha Chakraborty. RNN based LSTM variations such as Deep LSTM, Convolutional LSTM, and Bidirectional LSTM are connected on Indian datasets to foresee the no. of positive cases. An online site is created where the state-wise expectations are upgraded utilizing the proposed demonstrate for specialists, analysts, and organizers.

III. METHODOLOGY

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

A. Data Collection

The data for the ongoing Covid-19 outbreak in India is collected from [3]. The columns of this dataset include the Total number of Confirmed, Active, Cured, and Death cases of Covid-19 patients accumulating all the states, on a day-to-day basis from 12th March 2020 to 31st October 2020.

B. Data Preprocessing

In the section of data preprocessing, 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 is Y and "Months" as the independent variable X. 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 StandardScaler() 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 a combination of colored 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, color intensity is the least.

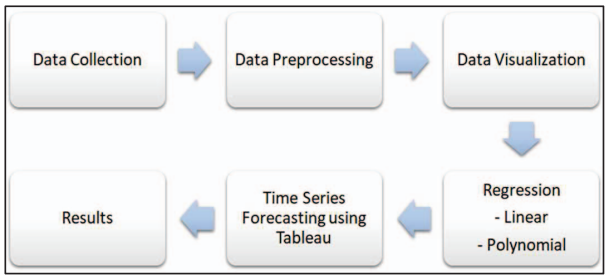

Fig. 1. Flowchart for the implemented methodology

	State/UT	Confirmed	Active	Cured	Death	Total cases
0	Andaman and Nicobar Islands	305070	41430	259586	4054	305070
1	Andhra Pradesh	51464573	7476387	43541522	446664	51464573
2	Arunachal Pradesh	695522	180666	513472	1384	695522
3	Assam	13482777	2513934	10920821	48022	13482777
4	Bihar	14995667	1855181	13065405	75081	14995667
5	Chandigarh	768341	149076	608572	10693	768341
6	Chhattisgarh	7520137	1962820	5490422	66895	7520137
7	Dadra and Nagar Haveli and Daman and Diu	255955	30965	224772	218	255955
8	Delhi	26792955	3331655	22845112	616188	26792955
9	Goa	2437497	416310	1991926	29261	2437497
10	Gujarat	13185893	2114405	10647463	424025	13185893
11	Haryana	9921229	1394331	8416223	110675	9921229
12	Himachal Pradesh	1067939	249179	806536	12224	1067939
13	Jammu and Kashmir	5791066	1282506	4413529	95031	5791066
14	Jharkhand	5749145	1041359	4656539	51247	5749145
15	Karnataka	45601763	9755561	35152851	693351	45601763
16	Kerala	15191531	4562411	10573992	55128	15191531
17	Ladakh	381422	91704	285804	3914	381422
18	Lakshadweep	0	0	0	0	0
19	Madhya Pradesh	10084199	1700773	8167382	216044	10084199
20	Maharashtra	112192904	24692767	84187205	3312932	112192904
21	Manipur	904989	244534	655191	5264	904989
22	Meghalaya	421216	145151	272714	3351	421216
23	Mizoram	148570	40204	108363	3	148570
24	Nagaland	516003	144274	370349	1380	516003
25	Odisha	15309909	2386290	12857369	66250	15309909
26	Puducherry	1914791	402236	1478644	33911	1914791
27	Punjab	7835588	1290120	6315464	230004	7835588
28	Rajasthan	11636988	1927764	9562266	146958	11636988
29	Sikkim	220023	48273	169415	2335	220023
30	Tamil Nadu	53039667	6676091	45522073	841503	53039667
31	Telangana	15913596	2658014	12947352	108230	15913596
32	Tripura	1832035	411622	1402348	18067	1832035
33	Uttar Pradesh	29993960	5187849	24341697	464414	29993960
34	Uttarakhand	3337851	675684	2616291	45876	3337851
35	West Bengal	21021876	3126404	17485345	430127	21021876

Fig. 2. Heat map of State-wise COVID-19 Cases in India

State/UT	Total Active
Maharashtra	24892767
Karnataka	9755561
Andhra Pradesh	7476367
Tamil Nadu	6676091
Uttar Pradesh	5187849
Kerala	4562411
Delhi	3331655
West Bengal	3126404
Telangana	2658014
Assam	2513934
Odisha	2386290
Gujarat	2114405
Chhattisgarh	1962820
Rajasthan	1927764
Bihar	1655181
Madhya Pradesh	1700773
Haryana	1394331
Punjab	1290120
Jammu and Kashmir	1282506
Jharkhand	1041359
Uttarakhand	675684
Goa	416310
Tripura	411622
Puducherry	402236
Himachal Pradesh	249179
Manipur	244534
Arunachal Pradesh	180666
Chandigarh	149076
Meghalaya	145151
Nagaland	144274
Ladakh	91704
Sikkim	48273
Andaman and Nicobar Islands	41430
Mizoram	40204
Dadra and Nagar Haveli and Daman and Diu	30965
Lakshadweep	0

Fig. 3. Heat map of State-wise Total Active Cases in India on 31st October 2020

D. Linear

We have employed the model of a linear regression which is a way to model the relationship between two variables. Model fitting is done using the least-squares approach. The condition has the form

$$y=a+bX+\in \quad (1)$$

Where Y and X are the dependent and independent variables respectively, b is the slope of the line, a is the y-intercept, and $\in$ is an unexpected error [9].

E. Polynomial

The second type of regression analysis that we have used is the polynomial regression provides the relationship between the dependent variable Y and the independent variable X and is modeled as 2nd, 3rd, 4th, and 5th-degree polynomial in x. 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.

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_1 x^n + \varepsilon \quad (2)$$

where ε is an unexpected error with mean zero conditioned on a scalar variable x, β_0 is a constant, and $\beta_1$to β_n are coefficients [10].

F. Time Series Forecasting using Tableau

The Time Series Forecasting is done utilizing Tableau, it happens when logical expectations are made utilizing authentic time-stamped information. It includes building models through authentic investigation and utilizing them to form perceptions and drive future vital decision-making. The importance of estimating is that long-run results can be evaluated through cautious examination and evidence-based priors utilized from the dataset.

To create a forecast in Tableau, at slightest two factors are required, for illustration, the number of dynamic cases that are to be forecasted is on the Row shelf and a ceaseless date field is on the Column shelf. Tableau visualizes evaluated future values of the measure concerning actual authentic values and after that, the assessed values are appeared by default in a lighter shade of the color utilized for the chronicled information [11].

Steps for forecasting using Tableau:

- Start
- Open Tableau Desktop
- Select the dataset having extension .csv
- Date in the Column shelf
- Number of cases in the Row shelf
- In the Analytics section, select the option: Forecast
- Change the number of months till you obtain the forecasted data
- Add label marks
- End

G. Error Analysis

Mean Absolute Percentage Error (MAPE) 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.

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{At}\right| \quad (3)$$

Where n is the no. of observations, A_t is the Actual Error and F_t is the Forecasted Error.

R-squared (R^2) 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^2 score [12].

IV. Results And Discussions

This section contains the experimental results, figures, tables, and explanations.

TABLE I. ACTUAL DATA

Months of 2020	Types of Cases			
	Confirmed	*Active*	*Cured*	*Death*
April	13788	10902	2451	435
May	94822	55396	36529	2896
June	351945	152324	188964	10656
July	1023435	353318	644520	25597
August	2604826	665008	1889706	50111
September	4970459	935915	3953097	81445
October	7315171	789945	6413690	111536

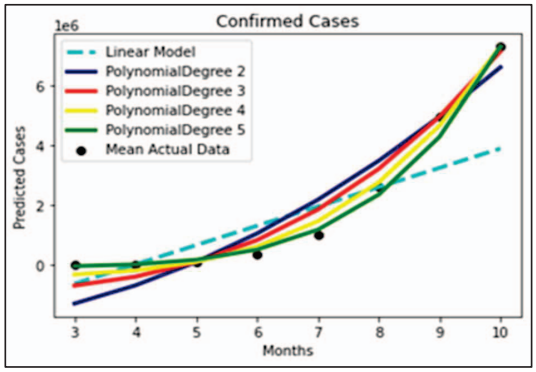

Fig. 4. Regression models applied for Confirmed Cases

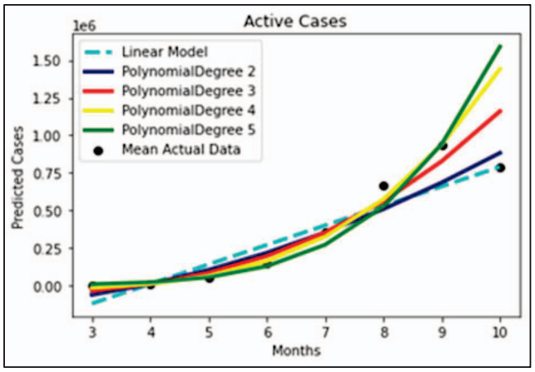

Fig. 5. Regression models applied for Active Cases

Figures 4 to 7 depict the prediction of the number of confirmed, active, cured, and death cases done by linear and polynomial regression models.

The polynomial models are categorized into four different degrees. These graphs present a comparison of the two models on [3], where we can observe that polynomial regression models of degree 4 and 5 outperform other degree models and the linear model. As the degree of polynomial increases, the curve adapts better towards the actual data of each month.

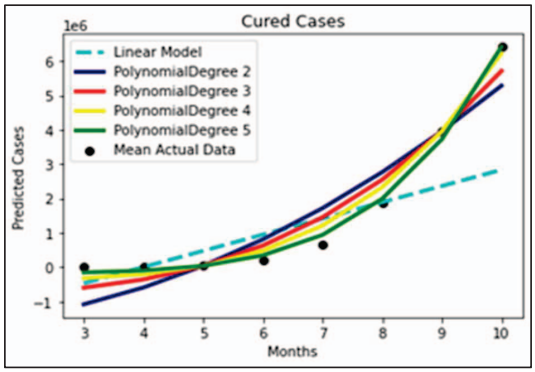

Fig. 6. Regression models applied for Cured Cases

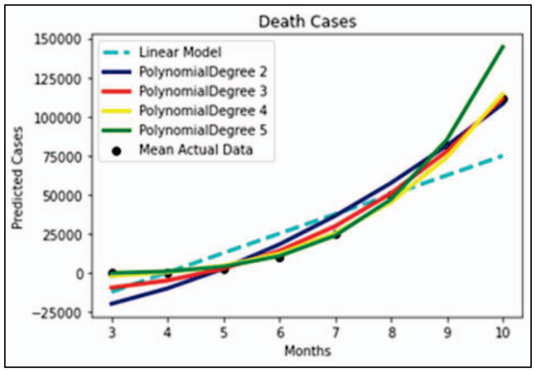

Fig. 7. Regression models applied for Death Cases

TABLE II. PREDICTED DATA BY POLYNOMIAL DEGREE 5

Months of 2020	Types of Cases			
	Confirmed	*Active*	*Cured*	*Death*
April	13846	22001.5	-101771	1149
May	168800	55275	36546.6	4192
June	511824	128933.9	342742.5	10928
July	1177884	271959	937293	24007
August	2355050.5	524735.6	1988076.5	47121
September	4293344	940951	3718271	85182
October	7313589.6	1589497.9	6414256.8	144488

TABLE III. MEAN SQUARE ERROR (MSE)

Models		Types of Cases			
		Confirmed	*Active*	*Cured*	*Death*
Polynomial	*Degree 2*	5.77	0	144	1.24
	Degree 3	1.92	0	52.65	0.32
	Degree 4	0.46	0	17.48	0
	Degree 5	0	0	4.34	0
Linear		5.69	0	153	1.39

Table III represents the formulation of errors generated after the prediction of confirmed, actual, cured, and death cases by the two regression models, and presents a comparison between four polynomial models and a linear model applied on [3], where we can observe that Polynomial Regression models of degree 5 outperform other degree models and the linear model. As the degree of polynomial increases, MSE decreases.

TABLE IV. ACCURACY OF COVID-19 ANALYSIS FOR INDIA BASED ON R^2 SCORE AND MEAN AVERAGE PERCENTAGE ERROR

Models		Confirmed		Active		Cured		Deaths	
		R^2 score	*MAPE*	*R^2 score*	*MAPE*	*R^2 score*	*MAPE*	*R^2 score*	*MAPE*
Polynomial	***Degree 2***	0.968357651	679.76	0.87213159	13.3	0.957003654	3395.68	0.976890065	315.66
	Degree 3	0.971244641	392	0.932280645	14.84	0.967077464	2052	0.976893557	161
	Degree 4	0.973863222	193.8	0.96323221	12.25	0.968015478	1182.8	0.978630318	9.45
	Degree 5	0.974727757	16.53	0.965823518	20.55	0.968675839	589.44	0.978745042	33.34
Linear		0.791163744	125.66	0.85897801	27.9	0.744672132	231.84	0.857393043	10.39

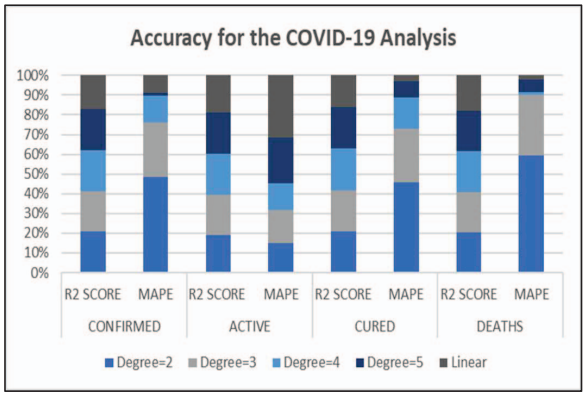

Fig. 8. Performance evaluation of the regression models

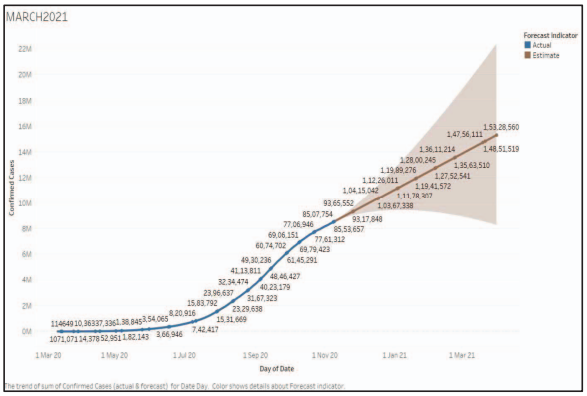

Fig. 10. Forecasting of Confirmed Cases till March 2021

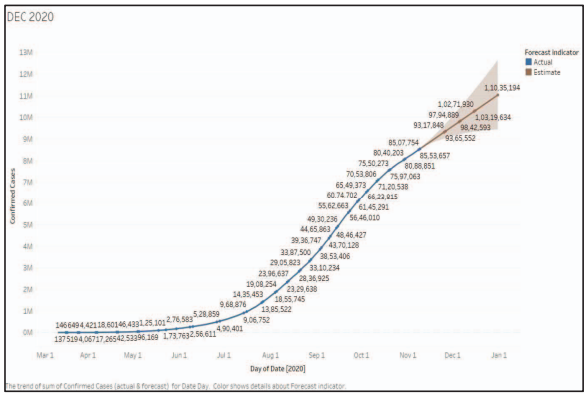

Fig. 9. Forecasting of Confirmed Cases till December 2020

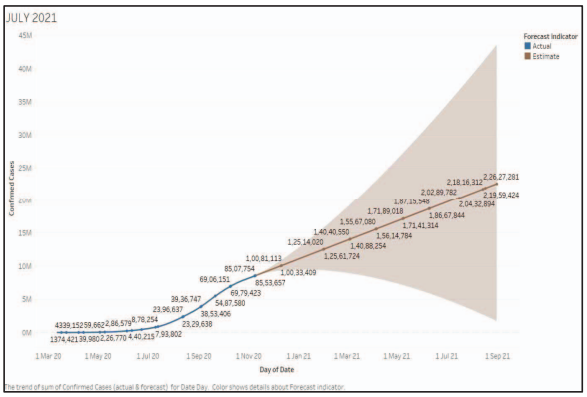

Fig. 11. Forecasting of Confirmed Cases till July 2021

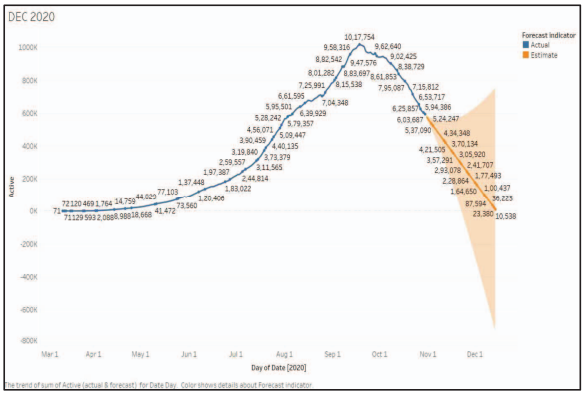

Fig. 12. Forecasting of Active Cases till December 2020

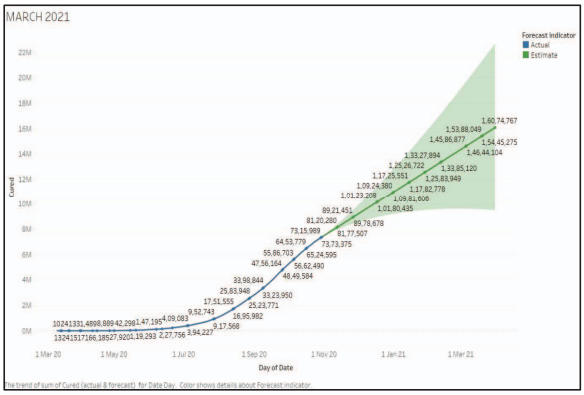

Fig. 15. Forecasting of Cured Cases till March 2021

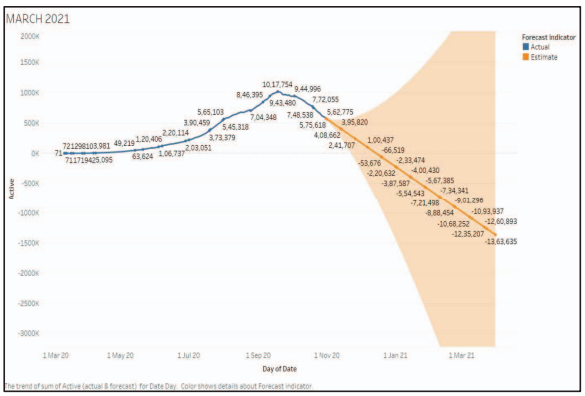

Fig. 13. Forecasting of Active Cases till March 2021

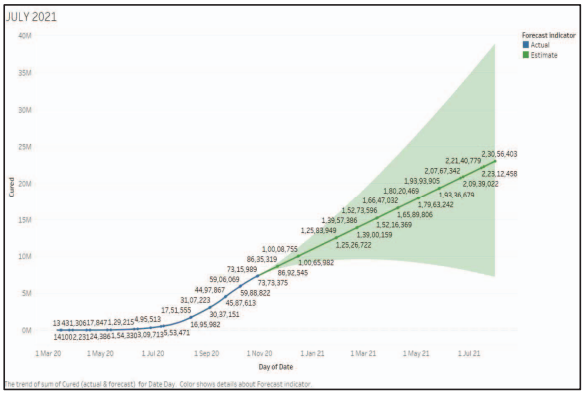

Fig. 16. Forecasting of Cured Cases till July 2021

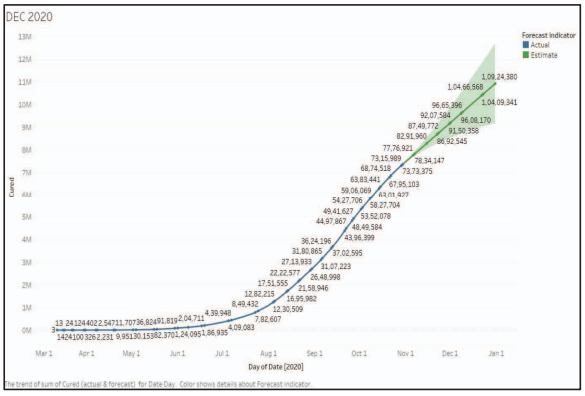

Fig. 14. Forecasting of Cured Cases till December 2020

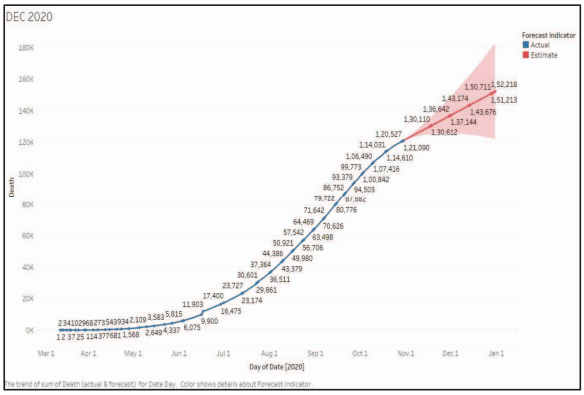

Fig. 17. Forecasting of Death Cases till December 2020

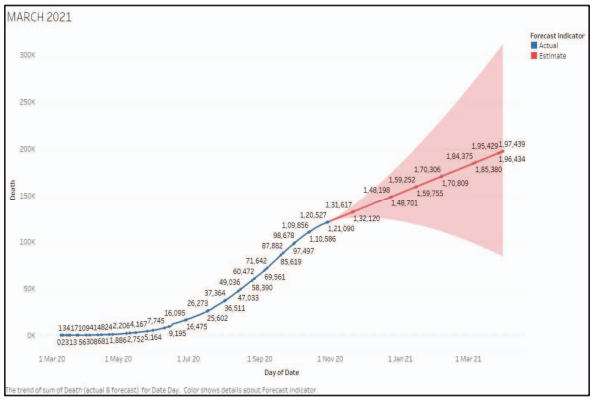

Fig. 18. Forecasting of Death Cases till March 2021

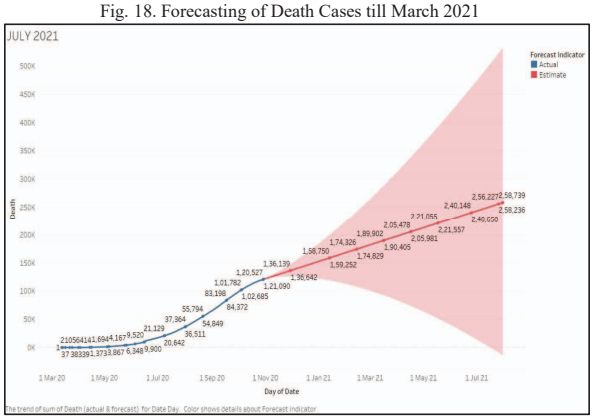

Fig. 19. Forecasting of Death Cases till July 2021

Figures 9 to 11 and 14 to 19 represent the forecasting of Confirmed, Cured, and Death Cases for the COVID-19 dataset in India up to December 2020, March 2021, and July 2021. This is done with the help of Time Series Forecasting [3]. The Active Cases forecasting is performed up to Dec 2020 and Mar 2021 represented by Figures 12 and 13. All of these cases contain a confidence interval of 95%. Fig. 8 depicts the future cases for Dec 2020 in which the trendline is observed to be an increasing one. This is because although the active cases might decrease, confirmed cases are rising because of cumulative density. However, Figures 10 and 11 represent the forecasted cases till Mar and July of 2021 respectively, and it is seen that when the lower bound of the confidence interval for both are considered, the forecasted cases may match the expected scenario as the no. of active cases in Figures 12 and 13 go on decreasing from Dec following a dip in Mar 2021. The actual number of Active cases in the 3rd week of September was reported to be 10,17,754 marking the peak, following which the number of cases began to decline. As per Fig. 12, in Dec, the cases are expected to become nearly zero towards the end of the month. Figures 14 to 16 depict the number of forecasted Cured cases in which the number of cured cases is increasing in all three scenarios because the active cases are going down. Fig. 17 depicts the forecasted no. of death cases for Dec 2020 in which the trend is observed to be an increasing one because of cumulative density. However, Figures 18 and 19 represent the forecasted cases till Mar 2021 and July 2021 respectively, and if the lower bound of the confidence interval for both are taken into consideration with the actual forecasted cases, the cases are expected to reduce. For all these results, some variation is likely in the no. of cases present in [3] due to factors like lack of testing facilities in certain areas, testing unavailability on the weekends. etc.

V. CONCLUSION

India's COVID-19 dataset [3] is analyzed to mine the trends of the cases that can help the government and citizens to ensure safety by taking precautionary measures in the future. Analysis of dataset is done using linear and polynomial regressions which involved metrics like accuracy, R^2 score, and MAPE. Based on Fig. 8 and Table IV, we can conclude that polynomial is better than linear regression. Forecasting is done using Tableau and the results are found to be satisfactory. However, the error rate in the future can be reduced by using a bigger dataset, better algorithms, and fine-tuning of the parameters.

REFERENCES

[1] "World Health Organization," 7 April 1948. [Online]. Available: https://www.who.int. [Accessed 17 September 2020].

[2] "Pfizer," Pfizer, 18 November 2020. [Online]. Available: https://www.pfizer.com/news/press-release/press-release-detail/pfizer-and-biontech-conclude-phase-3-study-covid-19-vaccine. [Accessed 19 November 2020].

[3] "PRS legislative research," 12 March 2020. [Online]. Available: https://prsindia.org/covid-19/cases. [Accessed 21 August 2020].

[4] S. Tiwari, S. Kumar and K. Guleria, "Outbreak Trends of Coronavirus Disease–2019 in India: A Prediction," *Disaster Medicine and Public Health Preparedness,* vol. 14, no. 5, pp. 1-6, 2020.

[5] P. Wang, X. Zheng, J. Li and B. Zhu, "Prediction of epidemic trends in COVID-19 with logistic model and machine learning technics," *Chaos, Solitons and Fractals,* vol. 139, no. 110058, pp. 1-7, 1 July 2020.

[6] M. Yadav, M. Perumal and M. Srinivas, "Analysis on novel coronavirus (COVID-19) using machine learning methods," *Chaos, Solitons and Fractals,* vol. 139, no. 110050, pp. 1-12, 2020.

[7] A. Tomar and N. Gupta, "Prediction for the spread of COVID-19 in India and effectiveness of preventive measures," *Science of the Total Environment,* vol. 728, no. 138762, pp. 1-6, 20 April 2020.

[8] D. Parbat and M. Chakraborty, "A python based support vector regression model for prediction of COVID19 cases in India," *Chaos, Solitons and Fractals,* vol. 138, no. 109942, pp. 1-5, 2020.

[9] S. Rath, A. Tripathy and A. R. Tripathy, "Prediction of new active cases of coronavirus disease (COVID-19) pandemic using multiple linear regression model," *Diabetes & Metabolic Syndrome: Clinical Research & Reviews,* vol. 14, no. 5, pp. 1467-1474, 2020.

[10] R. Tamhane and S. Mulge, "Prediction of COVID-19 Outbreak using Machine Learning," *International Research Journal of Engineering and Technology,* vol. 7, no. 5, pp. 5699-5702, 2020.

[11] A. Aggarwal, A. Rani, P. Sharma, M. Kumar, A. Shankar and M. Alazab, "Prediction of Landsliding using Univariate Forecasting Models," *Internet Technology Letters,* no. e209, pp. 1-6, 2020.

[12] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion and O. Grisel, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research,* vol. 12, pp. 2825-2830, 2011.

[13] 2020. [Online]. Available: https://www.tableau.com/. [Accessed 10 September 2020].

[14] P. Arora, H. Kumar and B. K. Panigrahi, "Prediction and analysis of COVID-19 positive cases using deep learning models: A descriptive case study of India," *Chaos, Solitons and Fractals,* vol. 139, no. 110017, pp. 1-9, 2020.