Prediction of COVID'19 Outbreak by Using ML-Based Time-Series Forecasting Approach

Devesh Kumar Shrivastava, Akhilesh Kumar Sharma, and Sachit Bhardwaj

Abstract

The COVID-19 now became a pandemic and rising rapidly and spreading in all parts of the world like fire. India reported its first COVID-19 case on January 30, when a student arrived in Kerala from Wuhan. Thousands of people are acquiring this deadly virus daily and with many people dying from it. The major concern of all the countries is to protect its citizens and try to eradicate this disease as fast as possible. This paper aims to perform exploratory analysis using the concepts of data science on the confirmed cases, total deaths, and total recovered cases of this virus. The research work predicts the spread of the outbreak for the next five days by using time-series forecasting algorithms. This paper deals with learning about how the corona virus is spreading and using that trend to predict for the upcoming days. It would be able to predict to a suitable accuracy which can help the government learn about the statistics of this disease and prepare further for protection against this. The results are discussed at last with prediction and error estimates.

Keywords

COVID-19 • Machine learning • Time-series forecasting • SVM • HL model • AR model • MA model • HW model • FP model • ARIMA

1 Introduction

In machine learning, patterns are observed using statistics and the learning of model is performed by continuous iterations until an estimation of data prediction is observed. This study most commonly synthesizing useful concepts from the historical data. Machine learning is a method of training machines, i.e., computers to make a prediction based on some training data and experience. Application of machine learning is limitless, i.e., from health care industry to statistical-based conditions of a country. Machine learning is not just limited to a particular field, but it can be used to improve the existing knowledge of a field by learning from the previous data and predictions. In briefly, machine learning is an application of artificial intelligence that automates analytical model building by using an algorithm that iteratively learns from data without being explicitly programmed (Sharma et al., 2018). We can predict the further outcome or predicted outcome by analyzing through time-series order. Patterns like seasonality, trends, irregularity and cyclicity are used as features to predict the upcoming variable of interest. There are various application of time-series forecasting like earthquake prediction, stock market prediction, etc. The performance of the time-series forecasting models can be compared by evaluation error rate. The most in use error rate is root mean squared error.

2 Literature Review

Stephanie et al. (2020) in her research work analyzed the impact of COVID-19 with respect to geographical differences over features like population density, distribution of age, diagnostic capacity, etc. Leeb et al. (2020) in his research work analyzed the impact of COVID-19 targeting the specific age group, i.e., school-going children.

Lim et al. (2020) in his research work analysis the impact of COVID-19 targeting the specific section, i.e., interns

D. K. Shrivastava · A. K. Sharma (✉) · S. Bhardwaj
Department of Information Technology, Manipal University Jaipur, Jaipur, Rajasthan, India
e-mail: akhileshshm@gmail.com

D. K. Shrivastava
e-mail: devesh988@yahoo.com

S. Bhardwaj
e-mail: sachitbhardwaj@yahoo.in

P. K. Singh et al. (eds.), *Innovations in Information and Communication Technologies (IICT-2020)*,
Advances in Science, Technology & Innovation,
https://doi.org/10.1007/978-3-030-66218-9_33

working in University Hospital. Hayashi et al. (2020) in her research work analyzed the changed seasonal effect of influenza virus and SARS-Cov-2 due COVID-19 rules and regulations.

Wilson et al. (2020) in their research work performed clustering approach over COVID-19 impact on University Campus. Hawas (2020) in his research work performed time-series prediction for daily infection (COVID-19) rates in Brazil using RNN.

Alonso et al. (2020) in their research work discussed the various challenges for post-COVID-19 era and proposed some strategies for them. Lee and Lin (2020) in their research work analyzed the COVID-19 precaution relationships with other common infections and studied the impact of their outburst due to COVID-19. Cardil and de-Miguel (2020) proposed a scenario in his research work where COVID-19 rules and regulations could directly intervene and cause more damage by natural disasters. Filimonau et al. (2020) observed the section in their research work and manifested their commitment toward the job.

3 Methodology

Metric is a measurement of errors between the grouped observations that express the same phenomenon (Sharma, 2015). The most common metric used is root mean squared error (RMSE) which is also used in this paper.

Root mean squared error (RMSE) is calculated as the square root of average of squared differences between actual and predicted observations as shown in Eq. 1. RMSE can be used to penalize large errors as the error is squared before taking average.

$$\text{RMSE} = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}} \tag{1}$$

where $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n$ are predicted values, $y_1, y_2, \ldots, y_n$ are observed values and n is the number of observations.

Three time-series csv file datasets were fetched from Kaggle dating from January 22 to March 15. The three datasets are—confirmed, deaths, recovered. There are a of total 451 rows and 58 columns for each of these three datasets with features as province/state, country, latitude, longitude, dates from January 22, 2020 to March 15, 2020. The analysis has been done on all the features excluding the latitude and the longitude of the given locations.

The three csv file datasets were loaded to the Jupiter notebook using Pandas library and were converted into data frames that helped in representing the data through its in-memory 2d table. All the columns were extracted from the datasets using the.key () function. From this, further all the date columns were extracted using the loc () function. The analysis performed on the datasets and then appended into lists was total confirmed cases, total deaths, total recovered. Using these values from the lists, we calculated the mortality rate (total deaths/total confirmed), recovery rate (total recovered/total confirmed). All the dates and cases were converted into Nd array using NumpyPy. The dates stored into datasets were type casted into date-time format from integers and for better visualization. Using the loc () function on the Nd arrays, latest confirmed cases, latest death cases, latest recovered cases were displayed, i.e., from March 5, 2020 to March 15, 2020. The total number of confirmed cases per country was calculated. The unique values of provinces/states were stored, and it was observed that there were a lot of not a number (NAN) values assigned to these provinces/states which were removed using the pop () function. Top ten countries which had the greatest number of cases were calculated. As China was the first country to get affected by this deadly disease, it had the highest number of cases. To analyze this situation, a comparison was made between China and rest of the world on the basis of total number of confirmed cases. The dataset is pre-processed. As the model deals with dependent and independent variables, the dataset is split into training set and test set using the train test split 70% of the dataset is trained first and 30% of the dataset is kept for testing.

Polynomial Regression A polynomial function is used with the concept of curve fitting to forecast the variable of interest as shown in Eq. 2

$$f(x) = c_0 + c_1 x + c_2 x^2 \cdots c_n x^n \tag{2}$$

where n is the degree of the polynomial and c is a set of coefficients.

Support Vector Machine Regression Poly, sigmoid and Rbf (Gaussian) functions have been set inside the kernel which would further perform parallel processing of the data and produce the optimal function for the most appropriate prediction (Sharma & Shrivastav, 2020).The equation of hyperplane is shown in Eqs. 3 and 4, The Lagrangian form is minimized for w and b, where w is width of the margin and b is the constant.

$$g(x) = w^T x + b \tag{3}$$

$$J(w, b, \alpha) = \frac{1}{2} w^T w - \sum_{i=1}^{N} \alpha_i d_i \left(w^T x_i + b\right) + \sum_{i=1}^{N} \alpha_i \tag{4}$$

Holt's Linear Model Method proposed by Holt involves two smoothing relations, i.e., trend (b_t) and level (ℓ_t) with a forecast equation ($\hat{y}_{t+ht}$) as shown in Eqs. 5–7.

$$\hat{y}_{t+ht} = \ell_t + hb_t \tag{5}$$

$$\ell_t = \alpha y_t + (1-\alpha)(\ell_{t-1} + b_{t-1}) \tag{6}$$

$$b_t = \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)b_{t-1} \tag{7}$$

where $0 \le \alpha \le 1$ (level smoothing parameter) and $0 \le \beta^* \le 1$ (trend smoothing parameter).

Holt's Winter Model Method proposed by Holt involves three smoothing relations, i.e., trend (b_t), level (ℓ_t) and season (s_t) with a forecast equation ($\hat{y}_{t+ht}$) as shown in Eqs. 8–11.

$$\hat{y}_{t+ht} = \ell_t + hb_t + s_{t+h-m(k+1)} \tag{8}$$

$$\ell_t = \alpha(y_t - s_{t-m}) + (1-\alpha)(\ell_{t-1} + b_{t-1}) \tag{9}$$

$$b_t = \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)b_{t-1} \tag{10}$$

$$s_t = \gamma(y_t - e_{t-1} - v_{t-1}) + (1-\gamma)s_{t-m} \tag{11}$$

where k is the integer part of $(h-1)/m$, $0 \le \alpha \le 1$ (level smoothing parameter) $0 \le \beta^* \le 1$ (trend smoothing parameter) and $0 \le \gamma^* \le 1$ (seasonal smoothing parameter).

AutoRegressive Model(AR Model) In (autoregressive model (AR model) with the successor of past values of variables, variable of interest can be forecasted using linear combinations. An order of p AR model is shown in Eq. 12.

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t \tag{12}$$

where ε_t is white noise.

Moving Average Model (MA Model) In (moving average model (MA model) with the help of past forecast errors, variable of interest can be fore- casting in a regression alike model by using Eq. 13.

$$y_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q} \tag{13}$$

where ε_t is white noise.

Autoregressive Integrated Moving Average Model (ARIMA Model) (Autoregressive integrated moving average model (ARIMA model) is the combination of moving average and autoregression model as shown in Eq. 14. It follows the same stationary and invertibility environment as autoregressive and moving average models.

$$y_t' = c + \phi_1 y_{t-1}' + \cdots + \phi_p y_{t-p}' + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t \tag{14}$$

where y_t' is the differenced series.

Facebook's Prophet Model Facebook's Prophet Model predicts over nonlinear variables like trend, seasonality, holidays, idiosyncratic changes which is shown in the Eq. 15.

$$y(t) = g(t) + s(t) + h(t) + e(t) \tag{15}$$

where $g(t)$ is trend models non-periodic changes, $s(t)$ is seasonality which presents periodic changes, $h(t)$ is ties in effects of holidays and $e(t)$ covers idiosyncratic changes not harbored by the model.

The methodology section includes the stepwise algorithmic description where the research article presents the systematic research conducted and analyzed the trend on the COVID-19 dataset. The various algorithms applied and studied, and the conclusion has been presented at last. The mentioned methodology depicts the technical analysis related to the disease trends and the directions of the pandemic with respect to the duration/timing. The stipulated time is increasing, and the resultant spread also increases with the much-affected persons seems to be deadly among individuals Fig. 1 shown methodology.

4 Result

4.1 Polynomial Regression

As shown in Fig. 2, we have predicted the trend of confirmed cases using polynomial regression for the next five days, i.e., September 24, 2020 to September 29, 2020.

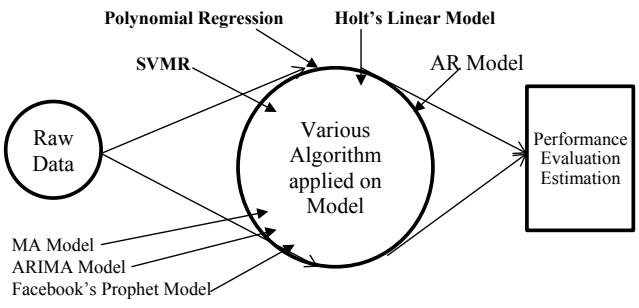

Fig. 1 Methodology

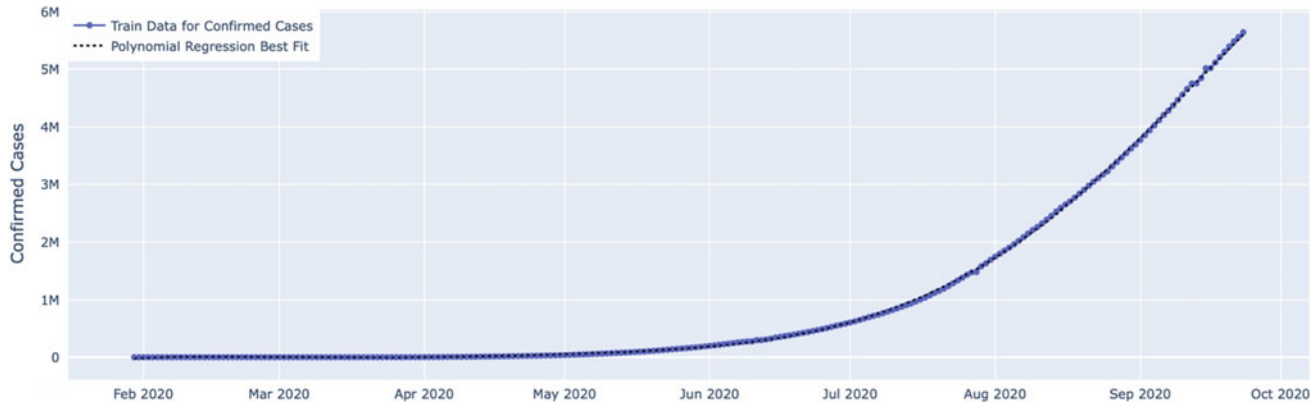

Fig. 2 Polynomial regression

4.2 Support Vector Machine Regressor

As shown in Fig. 3, we have predicted the trend of confirmed cases using support vector machine regressor for the next five days, i.e., September 24, 2020 to September 29, 2020. The root mean squared error (RMSE) was observed to be 1,005,399.938. Predicted confirmed cases are 7,387,905, 7,606,222, 7,830,090, 8,059,624, 8,294,945. This model performed worst.

4.3 Holt's Linear Model

As shown in Fig. 4, we have predicted the trend of confirmed cases using Holt's linear model for the next five days, i.e., September 24, 2020 to September 29, 2020. The root mean squared error (RMSE) was observed to be 113,382.878. Predicted confirmed cases are 5,909,289, 6,006,013, 6,102,737, 6,199,461, 6,296,186.

4.4 Holt's Winter Model

As shown in Fig. 5, we have predicted the trend of confirmed cases using Holt's winter model for the next five days, i.e., September 24, 2020 to September 29, 2020. The root mean squared error (RMSE) was observed to be 224,526.107. Predicted confirmed cases are 6,142,376, 6,284,535, 6,412,858, 6,583,385, 6,740,125.

4.5 AR Model

As shown in Fig. 6, we have predicted the trend of confirmed cases using autoregressive model (AR model) for the next five days, i.e., September 24, 2020 to September 29, 2020. The root mean squared error (RMSE) was observed to be 134,715.533. Predicted confirmed cases are 5,954,376, 6,057,034, 6,160,121, 6,263,644, 6,367,599.

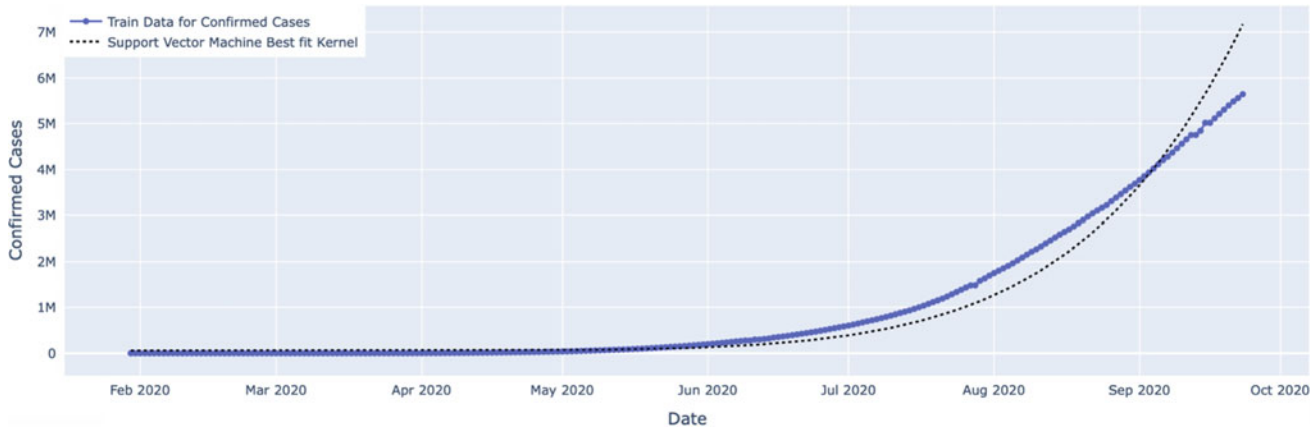

Fig. 3 Support vector machine regressor

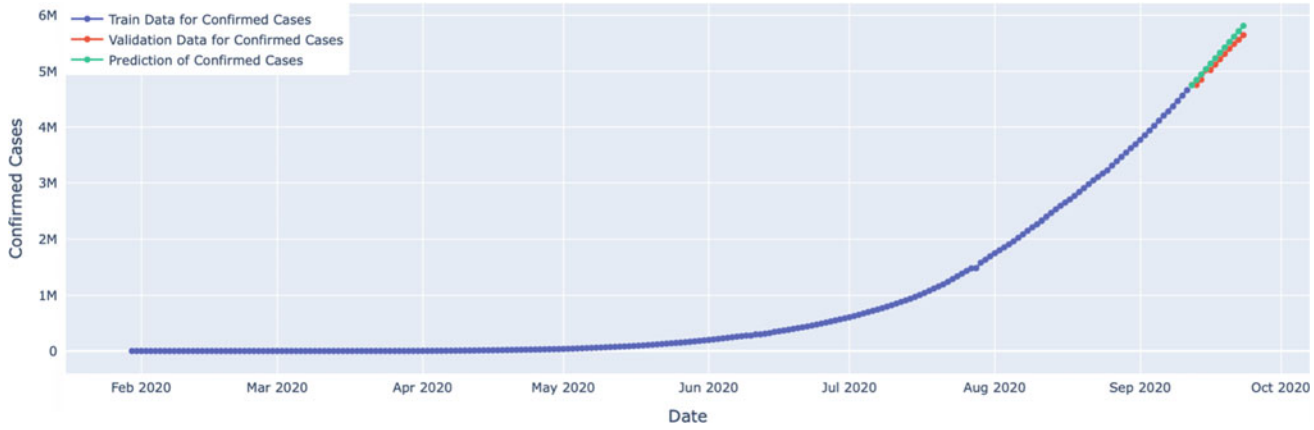

Fig. 4 Holt's linear model

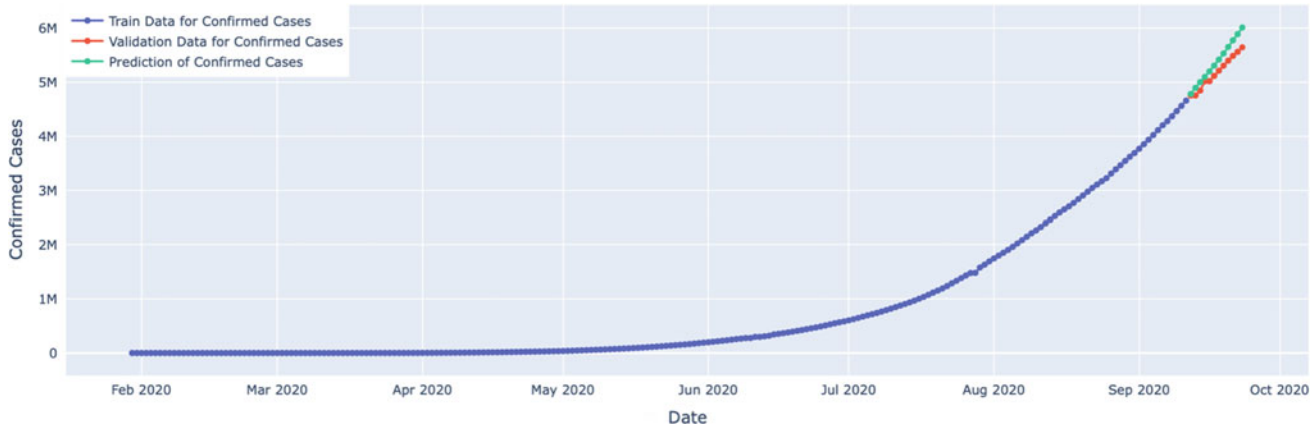

Fig. 5 Holt's winter model

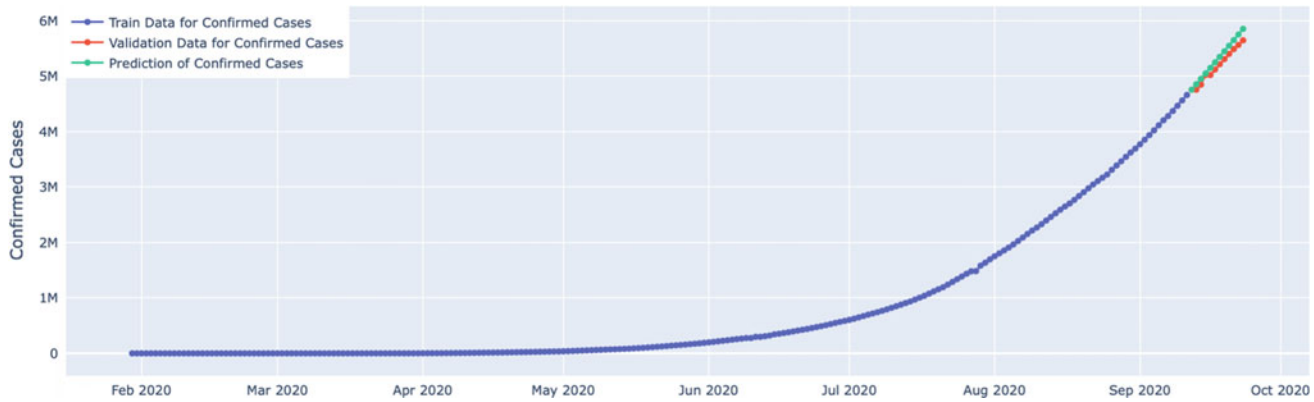

Fig. 6 AR model

4.6 MA Model

As shown in Fig. 7, we have predicted the trend of confirmed cases us polynomial regression for the next five days, i.e., September 24, 2020 to September 29, 2020. The root mean squared error (RMSE) was observed to be 37,850.063. Predicted confirmed cases are 5,682,521, 5,760,792, 5,839,125, 5,914,426, 5,989,599. This model performed as second best for MA model.

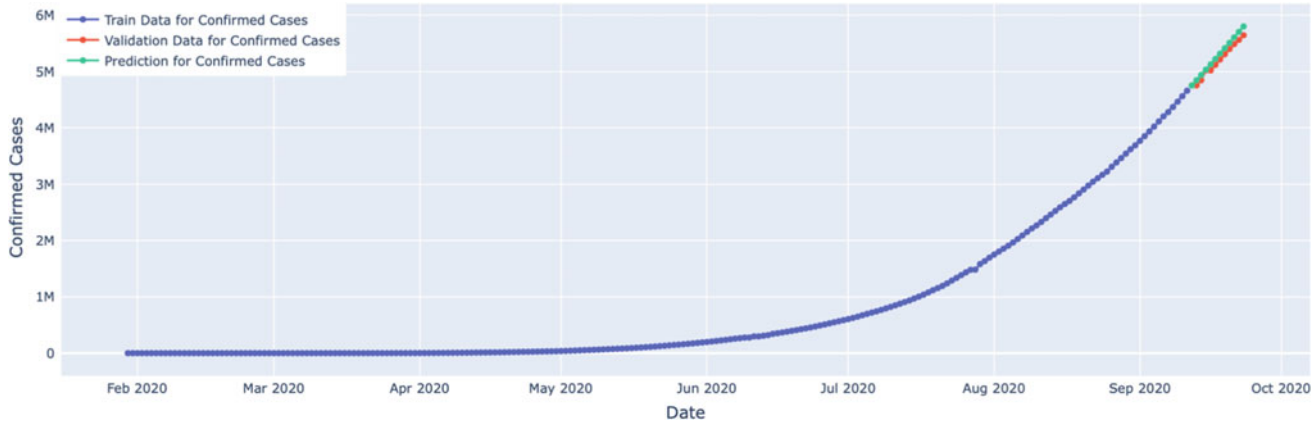

Fig. 7 MA model

4.7 ARIMA Model

As shown in Fig. 8, we have predicted the trend of confirmed cases using (autoregressive integrated moving average model (ARIMA model) for the next five days, i.e., September 24, 2020 to September 29, 2020. The root mean squared error (RMSE) was observed to be 112,111.823. Predicted confirmed cases are 5,912,914, 6,012,312, 6,112,132, 6,212,373, 6,313,037.

4.8 Facebook's Prophet Model

As shown in Fig. 9, we have predicted the trend of confirmed cases using Facebook's Prophet model for the next five days, i.e., September 24, 2020 to September 29, 2020. The root mean squared error (RMSE) was observed to be 36,248.027. Predicted confirmed cases are 5,598,438, 5,674,748, 5,751,535, 5,825,041, 5,899,729, whereas the upper bounds for the respective days are 5,666,743, 5,745,622, 5,824,853, 5,897,610, 5,968,758. This model performed bet.

4.9 Average of All Models

The average of all the prediction model's prediction for confirmed cases in the period between September 24–28 are observed as 6,017,199, 6,127,316, 6,236,641, 6,350,228, 6,462,896 (Figs. 10 and 11).

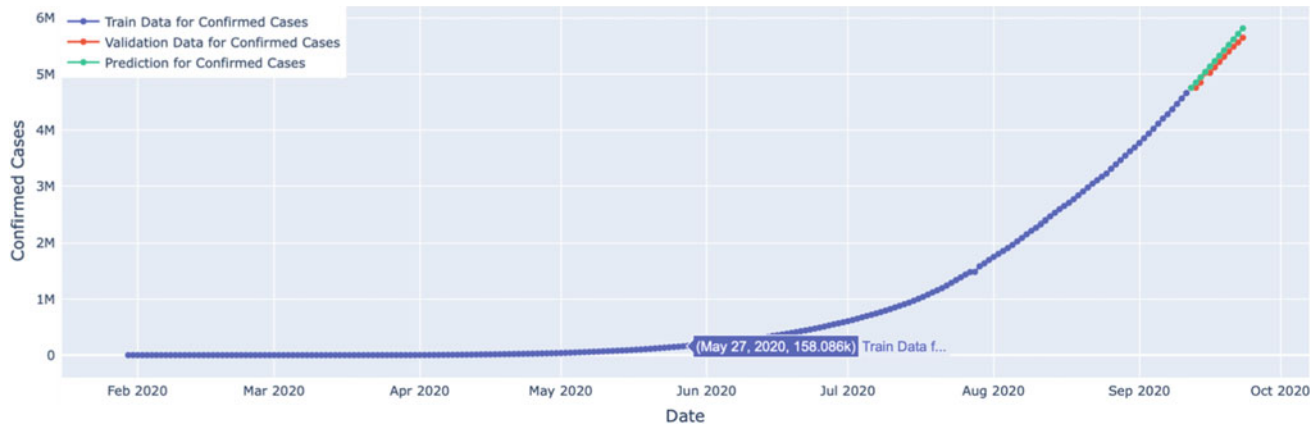

Fig. 8 ARIMA model

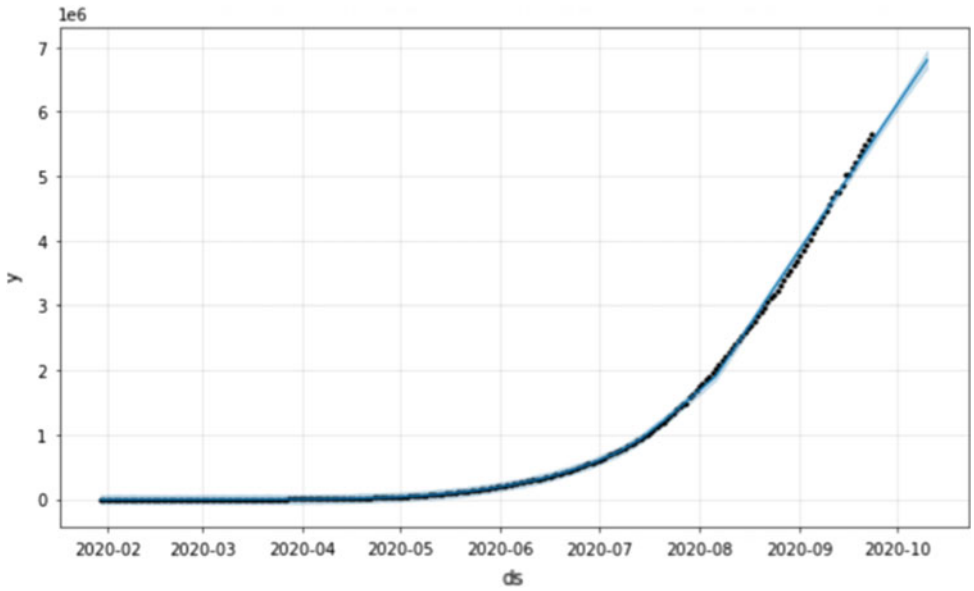

Fig. 9 Facebook's Prophet model

5 Conclusion and Future Work

From the analysis and the prediction, it is evident that corona virus is growing at a very rapid pace. It raises a serious concern in the world as its affects are catastrophic. It is a deadly virus which has taken up many lives and continues to do the sam. It is need of the hour to know about its effects on the whole world. This project helps the government in analyzing the situation and predicting its future outcomes so that preventive measures can be taken to contain and prevent this widely spreading disease.

The following work can be considered as future work:

1. Applying geospatial analysis, i.e., manipulating the data with the help of longitude and latitude of a place which would help in containing a place in which the outbreak has taken place at large scale so that it is localized to that certain area and not allowing it to spread any further.
2. Using a graphical user interface which would help in analyzing the situation in a more user-friendly way.

	Model Name	Root Mean Squared Error
7	Facebook's Prophet Model	36248.027538
0	Polynomial Regression	37850.063460
5	Moving Average Model (MA)	105576.927210
6	ARIMA Model	112111.823352
2	Holt's Linear Model	113382.878426
4	Auto Regressive Model (AR)	134715.533347
3	Holt's Winter Model	224526.107572
1	Support Vector Machine Regresssor	1005399.938112

Fig. 10 Root mean squared error of all models

Date	Polynomial Regression Prediction	SVM Prediction	Holt's Linear Model Prediction	Holt's Winter Model Prediction	AR Model Prediction	MA Model Prediction	ARIMA Model Prediction	Prophet's Prediction	Prophet's Upper Bound	Average of Predictions Models
2020-09-24	5682521	7387905	5909289	6142376	5954376	5900231	5912914	5598438	5666743	6017199
2020-09-25	5760792	7606222	6006013	6284535	6057034	5998567	6012312	5674748	5745622	6127316
2020-09-26	5838125	7830090	6102737	6412858	6160121	6097322	6112132	5751535	5824853	6236641
2020-09-27	5914426	8059624	6199461	6583385	6263644	6196495	6212373	5825041	5897610	6350228
2020-09-28	5989599	8294945	6296186	6740125	6367599	6296088	6313037	5899729	5968758	6462896

Fig. 11 Prediction of all models

References

Alonso, A. K., Bressan, S., O'Shea, A., Sakellarios, M., laosKoresis, N., Solis, A., Santoni, M., & Leonardo. (2020). COVID-19, aftermath, impacts, and hospitality firms: An international perspective. *International Journal of Hospitality Management, 91*, 102654. https://doi.org/10.1016/j.ijhm.2020.102654

Cardil, A., & de-Miguel, S. (2020). COVID-19 jeopardizes the response to coming natural disasters. *Safety Science, 130*, 104861. https://doi.org/10.1016/j.ssci.2020.104861.

Filimonau, V., Derqui, B., & Matute, J. (2020). The COVID-19 pandemic and organizational commitment of senior hotel managers. *International Journal of Hospitality Management, 91*, 102659. https://doi.org/10.1016/j.ijhm.2020.102659. Epub 2020 Aug 31. PMID: 32904709; PMCID: PMC7458044.

Hawas, M. (2020). Generated time-series prediction data of COVID-19's daily infections in Brazil by using recurrent neural networks. *Data Brief, 32*. 106175. https://doi.org/10.1016/j.dib.2020.106175. Epub 2020 Aug 19. PMID: 32839733; PMCID: PMC7437445.

Hayashi, T., Yaegashi, N., Konishi, I. (2020). *COVID-19 era, preventive effect of no going out against co-infection of the seasonal influenza virus and SARS-CoV-2*. Cold Spring Harbor Laboratory Press. 09.27.20202739.

Lee, H. H., & Lin, S. H. (2020). Effects of COVID-19 prevention measures on other common infections, Taiwan. *Emerging Infectious Diseases, 26*(10), 2509–2511. https://doi.org/10.3201/eid2610.203193. Epub 2020 Jul 30. PMID: 32730735; PMCID: PMC7510692.

Leeb, R. T., Price, S., Sliwa, S., Kimball, A., Szucs, L., Caruso, E., Godfred-Cato, S., & Lozier, M. (2020). COVID-19 trends among school-aged children United States, March 1–September 19, 2020.

Lim, W., Teoh, L. Y., Seevalingam, K. K., & Kuppusamy, S. (2020). *COVID-19 pandemic in University Hospital: Is there an effect on the medical interns?* Cold Spring Harbor Laboratory Press. 10.01.20205112.

Sharma, A. K., & Shrivastav, D. (2020). Statistical approach to detect Alzheimer's disease and autism spectrum-related neurological disorder using machine learning. In *Proceedings of SmartCom 2020. Smart Innovation, Systems and Technologies* (Vol. 182). Berlin: Springer.

Sharma, A. K., et al. (2015). Categorization of ICMR Using feature extraction strategy and MIR with ensemble learning. *Procedia Computer Science, 57*(201), 686–694.

Sharma, A. K., Chaurasia, S., & Srivastava, D. K. (2018). Supervised rainfall learning model using machine learning algorithms. In *Intelligent systems and computing book series* (Vol. 723, pp. 275–283).

Stephanie, B., Virginia, B., Nancy, C., Aaron, C., Ryan, G., Aron, H., Michelle, H., Tamara, P., Matthew, R., Katherine, R., Benjamin, S., Tami, S., Preethi, S., Emily, U., Michael, V., Hilary, W., & John, W. (2020). Geographic differences in COVID-19 cases, deaths, and incidence United States, February 12–April 7, 2020.

Wilson, E., Donovan, C. V., Campbell, M., Chai, T., Pittman, K., Sena, A. C., Pettifor, A., Weber, D. J., Mallick, A., Cope, A., Porterfield, D. S., Pettigrew, E., & Moore, Z. (2020). Multiple COVID-19 clusters on a University Campus—North Carolina, August 2020.