Prediction of COVID-19 Cases Using the ARIMA Model and Machine Learning

Akash Pal, Garima Jain, Ishita Roy, and Sumit Sharma

Abstract In both the first and second waves of COVID-19, India suffered many casualties at the hands of the infectious virus in the ongoing pandemic. As the days passed, the number of daily cases steadily increased, with most cases reported from large cities and outbreaks in rural areas. In this proposed paper, we tried to explore the effects of the previous data on the number of daily new cases using the ARIMA Model in machine learning. To forecast COVID-19 spread, we analyzed the effectiveness of multiple machine learning approaches. The accuracy of models was compared using the root mean squared error (RMSE), mean absolute error (MAE), R2 coefficient of determination (R2), and represent fundamental percentage error (MAPE) parameters. The primary goal of this research is to figure out how these new COVID-19 instances will influence the rest of the globe and how many people will be affected by the pandemic. This paper works with the ARIMA Model to predict the accurate growth rate and COVID-19 cases, which can help the government and various organizations plan their systematic strategies. In this paper, we use the ARIMA Model to analyze the growth rate and different factors related to the COVID-19. The model deployed accurately predicts the confirmed cases with an accuracy of 98.97%.

Keywords COVID-19 · Pandemic · Machine learning · ARIMA model · Global pandemic

1 Introduction

In today's world, almost every work sector produces some data. Various industries can use this data to get proper knowledge related to it. The data analysis is done

A. Pal (✉) · G. Jain · I. Roy · S. Sharma
Department of Computer Science and Engineering, Noida Institute of Engineering and Technology, Gr. Noida 201310, India
e-mail: apal62214@gmail.com

G. Jain
e-mail: garimajain@niet.co.in

J. K. Mandal and D. De (eds.), *Frontiers of ICT in Healthcare*, Lecture Notes in Networks and Systems 519, https://doi.org/10.1007/978-981-19-5191-6_37

on multiple levels to obtain appropriate and desirable results. For example, a businessman is interested in finding out the likelihood of sales in the years to come. He can use this predicted data to adjust the production factors in his business to profit. It could also lead to the avoidance of unsold products and good production quality.

When statistical data is collected over time at regular intervals, this kind of data is called a “time-series”. When we observe data at different points in time, the set of observations is called a “time-series”. Time is not fixed here permanently; it can be a set of intervals which can be a year, months, days, hours, minutes, or even seconds [1]. Time is one of the essential factors in the time-series analysis as it is directly connected to a single variable.

1.1 The Utility of Time Series

- By observing the data, we can analyze the transformation that happened in the past.
- It helps us form plans for future work.
- It helps in evaluating the current events.

1.2 Covid-19

At the end of 2019, a sudden outburst of a new disease with symptoms similar to the common cold was noted in a particular part of China. The number of victims of this disease has increased rapidly in the leading country and its surroundings. In the following three months, the disease was well known to be the harbinger of death and was feared across the world [2, 3]. Even today, in 2021, we are witnessing the various variants of the COVID-19 taking their toll on multiple countries. Though vaccines and proper medications have been discovered for COVID-19, there is still no cure available to assure 100% safety of the patient. Recently, the Omicron variant has been feared more than the previous variants of COVID-19 as it is a mutation of the previous ones. Due to a lack of accurate information, it is irrelevant to say anything about the current variant of COVID-19.

1.3 Research Objective

In this research, we are trying to accurately predict the increase in the number of cases of COVID-19, i.e., daily confirmed cases, using machine learning techniques. We will be analyzing the data through the ARIMA Model. Through this predicted data, hundreds of lives can be saved.

1.4 Research Motivation

The motivation for this research was to develop efficient techniques to accurately predict the data, so that we can help the general public in terms of providing sufficient supplies in the COVID-19 era.

1.5 Research Contribution

The contribution in this paper are as follows:

- The paper contributes towards the implementation of the ARIMA Model to the dataset of COVID-19 cases.
- The paper effectively predicts the future values of the dataset with accuracy of 98.97%.

1.6 Road Map

The rest of the chapter is organized as follows. The first section introduces the time-series analysis and its usage, and COVID-19 pandemic to the paper. Section II gives the literature review and previous works that has been done in the prediction and analysis of the COVID-19 cases. The introduction to the ARIMA Model and its general characteristics with the algorithm is described in section III. The results of this study are explained in the section IV. Finally, section V concludes the paper.

2 Literature Review

2.1 Genetic Programming

The suggested prediction models are provided with precise formulae, and the importance of prediction variables is investigated. The developed models were analyzed and tested using statistical parameters and metrics. As per the findings, the suggested GEP-based models for time-series prediction of COVID-19 cases in India employ simple linkage functions and are highly reliable [4].

The ERP models used in the proposed paper are highly consistent in the prediction of both daily COVID-19 cases and the daily death cases in India, satisfying all the external conditions [5]. The solution in the proposed paper is highly dependable as the RSME and R-value of all cases are higher and are close to the factor of 1 [5]. The prediction variables that are used in the model only depend on one or two variables, thus making the model less reliable than the prediction models. The essential characteristic of GEP models is their ability to deal with fewer time-series data while still generating consistent results. In the early days of the pandemic, it was instrumental in analyzing and forecasting the course of cases. The problem develops when the volume of data increases. An optimization of GEP models is usually required. They are optimized using highly effective algorithms. This model can be used to forecast the spread of a pandemic in its early stages.

2.2 ARIMA-WBF Models

The main objective of using this model in the paper "Real-time forecasts and risk assessment of novel coronavirus (COVID-19) cases: A Data-driven analysis" [6] is to counter two main issues.

(a) Short-term real-time forecasting of new COVID-19 cases and deaths for various nations.
(b) To predict the disease's fatality rate using the demographic characteristics of the country's topography and the characteristics of COVID-19 diseases.

In the second edition, we use an optimal regression tree algorithm to identify variable root causes that significantly impact death rates in different countries. Various constraints are taken on the dataset used to analyze COVID-19 cases in problem (b) and projection of the future possibilities in the problem (a). These constraints or assumptions are:

1. The virus's mutation rate is the same across all countries.
2. When a person recovers from the virus, the recovered person will attain permanent immunity.
3. The effects of climate change are not taken into consideration. Since this model can be presented as a real-time forecasts system, thus the actual data can be updated regularly. This model can be used efficiently to adjust the lockdown period according to the actual time predictions.

2.3 Two-Piece Scale Mixture Normal (TP-SMN)

Two-piece scale mixture standard (TP-SMN) distribution model. The suggested time-series models surpassed standard Gaussian and symmetry models and were first adjusted to past COVID-19 datasets. [1] After that, the time series with the best match to each dataset is picked. Finally, the models chosen will be used to forecast the number of confirmed cases and the global death rate from COVID-19. The proposed algorithm model showed us that it could closely predict the potential forthcoming confirmed cases based on the various constraints and the historical data is present to be used.

2.4 Dynamic Statistical Techniques

In the proposed paper "Investigating the cases of novel coronavirus disease (COVID-19) in China using dynamic statistical techniques" uses the dynamic statistical techniques and use of the formation of the time-series model and data panel models to make and investigate the relationship between death cases and confirmed cases and recovered cases and the confirmed cases [7]. The study on the paper reveals that the effect of the confirmed COVID -19 cases on the accountable deaths due to the COVID -19 is heterogeneous. The relationships between the confirmed fatalities and the confirmed cases across mainland China and Wuhan are linear, while the same for the confirmed recovered cases and the confirmed cases are nonlinear. The models also show that the increase in the confirmed cases by 1% also results in the rise of confirmed deaths due to the COVID-19 by approximately 0.1–0.7%.

3 Methodology

ARIMA Model or Autoregressive Integrated Moving Average Model is a statistical model used to analyze univariate and multivariate time-series data to enhance the understanding of the datasets and predict the future outcomes of a problem. It is used to predict future values based on the previous values. 'An autoregressive (AR) model is a sort of random model used to describe predictable time-varying phenomena in nature, among other things. The future value in an AR(p) model is predicted to be a linear mixture of p prior observations, a constant, and a random error term. The AR(p) model may be stated mathematically as process:'

The moving average (MA) model is a technique for modeling univariate time series in time-series analysis, as shown by Eq (1). According to the moving-average model, the output variable is influenced by the current and numerous previous values of incorrect parameters. The MA(q) model may be expressed mathematically as process:

An AR (p) model is often delineated as:

$$Y_t = c + \varphi_1 Y_1 + \varphi_2 Y_2 \ldots + \varphi_p Y_p + Z_t, \tag{1}$$

where $Z_t \sim$ (0, 2),

Equation (2) shows that c associate unknown constant term, and φ_i, $i = 1 \ldots p$, are the parameters of the AR Model. A MA(q) Model is often described as:

$$Y_t = c + Z_t + \theta_1 Z_{t-1} + \ldots + \theta_q Z_{t-q}, \tag{2}$$

where $\{Z_t\}$ denotes a only random process with zero mean and variance σ2z.

3.1 Understanding ARIMA Model and Its Components

An autoregressive integrated moving average model is a statistical analysis that measures the strength of one dependent variable concerning other variables that change their values.

- Autoregression (AR) Model: A model in which a changing variable reverts to its previous values.
- Integrated (*I*) Model: An integrated model represents the differentiation of raw observations for the time series to become stationary.
- Moving Average (MA) Model: A moving average model integrates an observation's dependence on an estimation error from a moving average model applied to delayed readings.

3.2 Stationarity of Data in ARIMA Model

Seasonality is where data exhibits consistent and predictable patterns that reoccur for a calendar year and may damage the regression model. Many of the computations during the process cannot be made with great accuracy if a trend arises and stationarity is not visible.

The data in an autoregressive integrated moving average model are differenced to make it stationary. A stationarity model demonstrates that the data is consistent throughout time.

Equations (3), (4), and (5) show an ARIMA (p, d, q) model of the nonstationary random method $Y(t)$ is expressed as

$$\Phi(B)(1-B)^{\wedge}dX^t = \theta(B)Z^t \tag{3}$$

with an AR operator

$$\Phi p(B) = 1 - \phi 1B - \ldots - \phi^{\mathrm{p}}\mathrm{Bp} \tag{4}$$

and a MA operator

$$\Theta q(B) = 1 - \theta 1B - \ldots - \theta^{\mathrm{q}}\mathrm{Bq} \tag{5}$$

where ϕ^p *is* pth AR coefficient, θ^q the qth MA coefficient, Z^t noise, X^t Rain attenuation. [8–10]. For stationary processes, Eq. (6) shows define autocorrelation between any two observations which can depend on the Time Lag h between them.

The autocorrelation for lag *h* is given by:

$$\rho_k = \mathrm{Corr}\big(y_t, y_{t-h}\big) = \frac{y_h}{y_0}, \tag{6}$$

3.3 ARIMA Model Parameters and Notations

Each component in the ARIMA Model function is a component of the observed data in the time series. In ARIMA Model [11], a standard notation would be *p*, *d*, and *q* to denote the numeric values that are substituted in place of argument to indicate the type of ARIMA Model used in the prediction.

The Notation is:

- The notation, *p*, denotes the number of lagged readings in the model; this is generally referred to as the lag order.
- The notation, *d*, denotes the count of the differenced raw readings, also known as the degree of difference.
- The notation, *q*, denotes the size of the moving average window; also referred to as the order of the moving average.

From Eq. (6), we define a (0) as a lag 0 covariance, i.e., the unconditional variance of the process. Confounding can explain as the distortion in the estimated measure of association where two variables can result from a mutual linear dependence on other variables.

3.4 Forecasting with ARIMA Model

ARIMA forecasting is accomplished by entering time-series data for the variable of interest. Statistical model and program will determine the proper number of delays or amount of differencing to the data and verify for stationarity. It will then produce the findings, which are frequently interpreted in the same way as a multiple linear regression model [10, 12, 13].

3.5 Dataset

In the ongoing pandemic, nation-wise data for confirmed cases for COVID-19 were obtained for India. The daily confirmed cases denote the total number of persons tested positive for COVID-19 in a single day. The dataset that is used in the section

Table 1 Dataset information

Variables in dataset	Variable type	Mean (in thousands)	Variance (in millions)	Min. value	Max. value
Daily confirmed	Numerical	7.79	132.44	0	50072
Total confirmed	Numerical	213.52	113404.08	1	1387088
Daily recovered	Numerical	4.98	62.59	0	37125
Total recovered	Numerical	123.24	44885.93	0	887124
Daily deceased	Numerical	0.18	0.07	0	2004
Total deceased	Numerical	5.88	75.79	0	32123

can be found at "https://www.kaggle.com/imdevskp/covid19-corona-virus-india-dataset?select=nation_level_daily.csv" for reproducibility of this work. The dataset contains variables such as Daily Confirmed Cases, Total Confirmed Cases, Total Recovered cases, Daily Confirmed Cases, Total Deceased Cases, and Daily Deceased Cases, which can be found in the mentioned dataset [14] (Table 1).

Algorithm

```
# IMPORT All the statistical libraries used for ARIMA Model and Machine Learning
# IMPORT Visualization Libraries
# Read the data set "MAIN_DATASET.CSV" and CONVERT it in form of DATAFRAME NAMED "DF"
# PRINT THE FIRST 5 ROWS OF DATAFRAME DF
# PRINT THE LAST 5 ROWS OF DATAFRAME DF
# CREATE A NEW DATAFRAME TO GET COLUMN "Daily Confirmed" for ANALYSIS
# PLOT THE GRAPH OF NEW DATAFRAME "NEW_DF" USING VISUALISATION LIBRARIES
# USE ML LIBRARIES TO GET THE DIFFERENCED VALUE OF THE DATASET "NEW_DF"
# DO THE ABOVE STEP AGAIN TO GET SECOND DIFFERENCED VALUE OF THE DATASET
# PLOT THE BOTH THE "DIFF1" AND "DIFF2" TO CHECK FOR STATIONARITY OF DATASET
# USE THE ARIMA FUNCTION OF THE ML LIBRARIES TO GET THE P, R, Q VALUES OF THE ARIMA MODEL ANALYSIS DONE ON THE DATASET
# PREDICT THE VALUE OF THE FUTURE COVID-19 CONFIRMED CASES USING THE ARIMA MODEL ANALYSIS
# PLOT THE GRAPH OF THE DATASET AND THE PREDICTED VALUE TO COMPARE THEM TO CHECK ACCURACY OF THE PREDICTION.
```

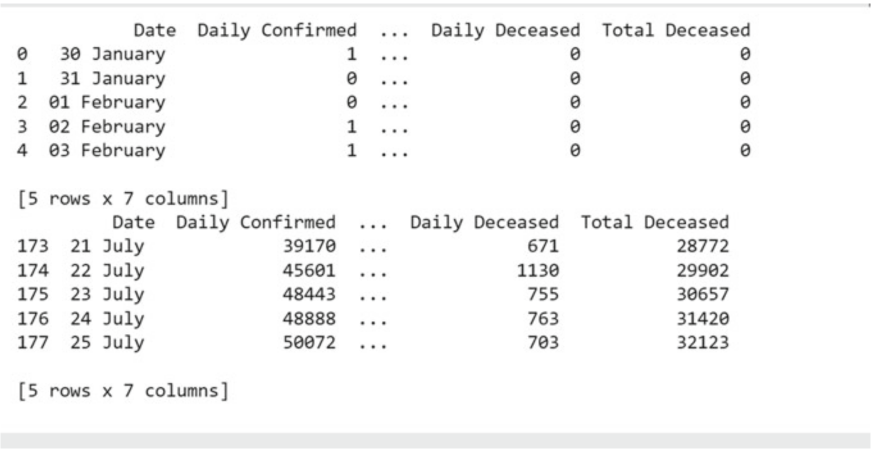
In this paper, we use various matplotlib, scikit-learn, stats, ARIMA libraries in the code part of the project to the apply the above algorithm using the python programming language. We analyze the data, get the required coefficients for the ARIMA Model. And try to predict the future outcome using the same.

4 Result

From the above code and visualization, we can see that the forecast of the number of daily cases in the COVID-19 crisis is nearly correct with a minor error of ± 500 points, which is not significant given the large dataset on which the work is in processing. The graph and values of the predicted data show that the forecast is nearly accurate, with a prediction rate of 98.97%.

Figure 1 shows the initial dataset we were working on, it shows the first five and last five entries of the dataset.

Graphs are one of the best visualization tools to show the trends of a variable. We used Python matplotlib library to successfully plot the trends of COVID-19 cases in form of graph (Fig. 2).



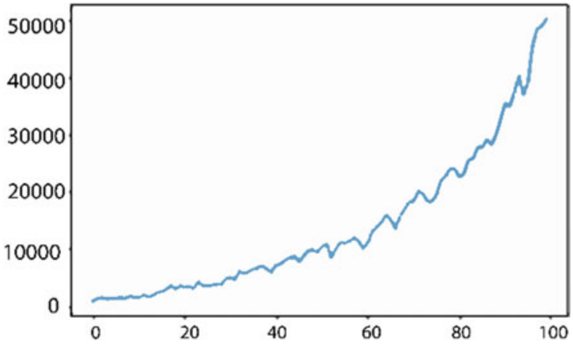
```
           Date  Daily Confirmed  ...  Daily Deceased  Total Deceased
0    30 January                1  ...               0               0
1    31 January                0  ...               0               0
2   01 February                0  ...               0               0
3   02 February                1  ...               0               0
4   03 February                1  ...               0               0

[5 rows x 7 columns]
         Date  Daily Confirmed  ...  Daily Deceased  Total Deceased
173   21 July            39170  ...             671           28772
174   22 July            45601  ...            1130           29902
175   23 July            48443  ...             755           30657
176   24 July            48888  ...             763           31420
177   25 July            50072  ...             703           32123

[5 rows x 7 columns]
```

Fig. 1 Shows the start and end of the main dataset to begin with

Fig. 2 This figure shows the dataset used in analyzing with its visualisation



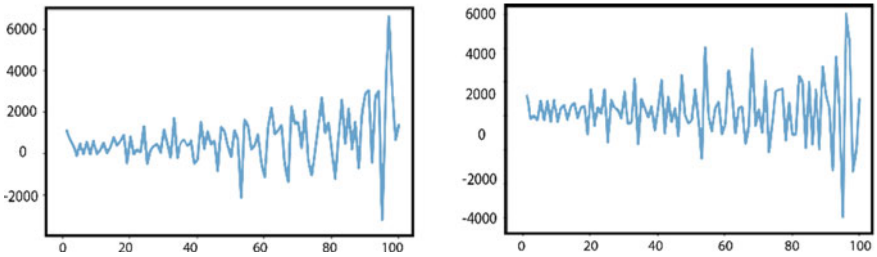


Fig. 3 This graph shows the difference between previous and current values for the ARIMA Model

The given dataset is checked for stationarity. The univariate time-series data of COVID-19 cases is checked for stationarity by differencing the current value from the previous values until the data is approximately stationary [15, 16]. The following figures show the differenced values of the univariate data for first time and second time, respectively (Fig. 3).

The following line graph shows the autocorrelated values of the dataset used in the prediction. This autocorrelated value is then used as a parameter in the ARIMA Model.

Autocorrelation refers to how much the value of dataset is correlated with its previous values. Partially Autocorrelated value is the statistical relationship between the readings of the dataset with the previous values of the dataset, where lagged observations are removed [17] This Autocorrelation plot shows the correlation between the various points of the dataset (Fig. 4).

The following table shows the result of the ARIMA Model obtained after applying the various functions. It returns values such as p-value, AIC value, BIC value, and HQIC values of the dataset. It is further used to predict the accurate results in ARIMA Model Analysis (Fig. 5).

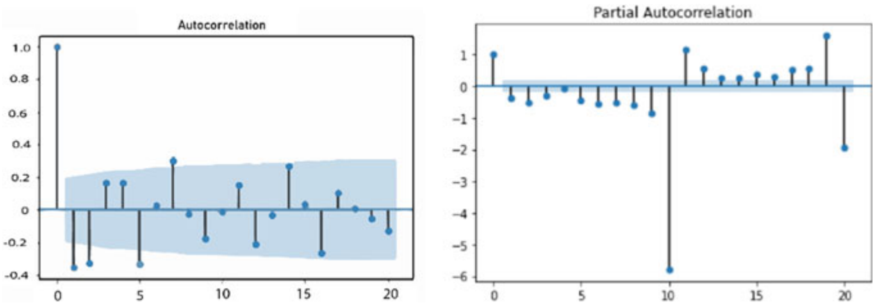
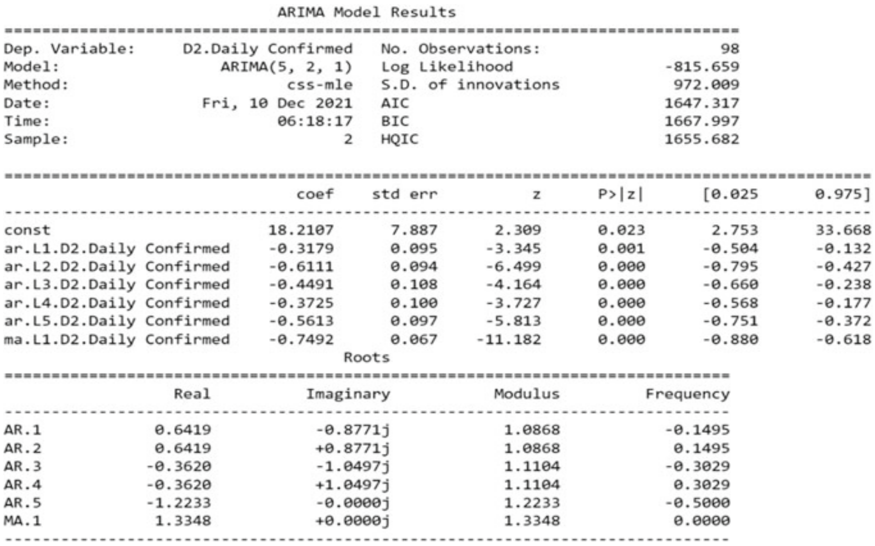


Fig. 4 Autocorrelated value and partially autocorrelated values of the dataset



```
                              ARIMA Model Results
==============================================================================
Dep. Variable:     D2.Daily Confirmed   No. Observations:                   98
Model:                 ARIMA(5, 2, 1)   Log Likelihood                -815.659
Method:                       css-mle   S.D. of innovations            972.009
Date:                Fri, 10 Dec 2021   AIC                           1647.317
Time:                        06:18:17   BIC                           1667.997
Sample:                             2   HQIC                          1655.682

============================================================================================
                               coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------------------
const                       18.2107      7.887      2.309      0.023       2.753      33.668
ar.L1.D2.Daily Confirmed    -0.3179      0.095     -3.345      0.001      -0.504      -0.132
ar.L2.D2.Daily Confirmed    -0.6111      0.094     -6.499      0.000      -0.795      -0.427
ar.L3.D2.Daily Confirmed    -0.4491      0.108     -4.164      0.000      -0.660      -0.238
ar.L4.D2.Daily Confirmed    -0.3725      0.100     -3.727      0.000      -0.568      -0.177
ar.L5.D2.Daily Confirmed    -0.5613      0.097     -5.813      0.000      -0.751      -0.372
ma.L1.D2.Daily Confirmed    -0.7492      0.067    -11.182      0.000      -0.880      -0.618
                                    Roots
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
AR.1            0.6419            -0.8771j            1.0868           -0.1495
AR.2            0.6419            +0.8771j            1.0868            0.1495
AR.3           -0.3620            -1.0497j            1.1104           -0.3029
AR.4           -0.3620            +1.0497j            1.1104            0.3029
AR.5           -1.2233            -0.0000j            1.2233           -0.5000
MA.1            1.3348            +0.0000j            1.3348            0.0000
-----------------------------------------------------------------------------
```

Fig. 5 ARIMA model result

The following Fig. 6 shows the proper value of the dataset of COVID-19 as compared with the predicted value generated by the ARIMA Model in the form of graph. It shows that the data are approximately similar with minor difference within the range of ± 500. The ARIMA Model generates the predicted data at an accuracy of approximately 98.97%.

5 Conclusion

The study's conclusion was that the daily growth of COVID-19 cases in the country could be predicted reliably. The forecast accuracy of 98.97% is considered satisfactory. This forecast can be used to take a variety of precautionary measures to prevent the virus from spreading. It can also be used to prepare hospital beds and to begin the production of various medications that can be used to combat the infection.

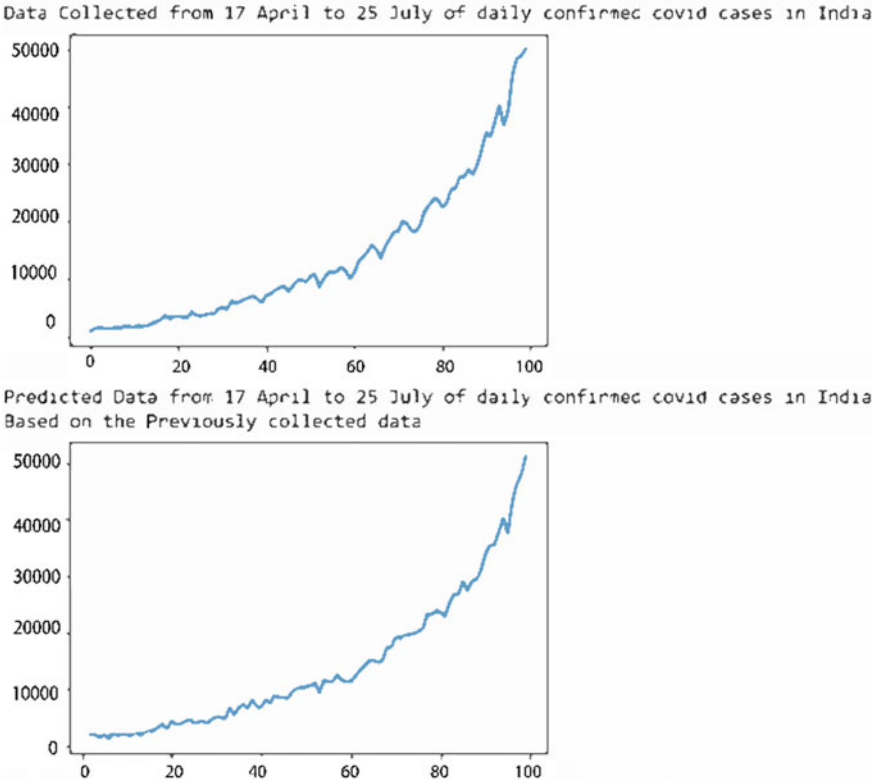


Fig. 6 Graphs of both the actual data and predicted data are approximately similar

References

1. Athiyarath S, Paul M, Krishnaswamy S (2020) A comparative study and analysis of time series forecasting techniques. SN Comput Sci 1:1–7
2. Tian J, et al (2020) Modeling analysis of COVID-19 based on morbidity data in Anhui, China. Math Biosci Eng 17(4):2842–2852
3. Zhuang Z, Zhao S, Lin Q, Cao P, Lou Y, Yang L, He D (2020) Preliminary estimation of the novel coronavirus disease (COVID-19) cases in Iran: a modelling analysis based on overseas cases and air travel data. Int J Infect Dis 94:29–31
4. Singh S, et al (2021) Time series analysis of COVID-19 data to study the effect of lockdown and unlock in India. J Institut Eng (India): Series B
5. Qi H, et al (2020) COVID-19 transmission in Mainland China is associated with temperature and humidity: a time-series analysis. Sci Total Environ 728:138778
6. Chakraborty T, Ghosh I (2020) Real-time forecasts and risk assessment of novel coronavirus (COVID-19) cases: A data-driven analysis. Chaos, Solitons Fractals 135:109850
7. Sarkodie SA, Owusu PA (2020) Investigating the cases of novel coronavirus disease (COVID-19) in China using dynamic statistical techniques. Heliyon 6(4):e03747
8. Jain G, Mallick B (2017) A study of time series models ARIMA and ETS. Available at SSRN 2898968
9. Jain G (2018) Time-Series analysis for wind speed forecasting. Malaya J Matematik 1:55–61

10. Garima J, Bhawna M (2016) A review on weather forecasting techniques. Int J Adv Res Comput Commun Eng 5(12):177–180
11. ArunKumar KE et al (2021) Forecasting the dynamics of cumulative COVID-19 cases (confirmed, recovered and deaths) for top-16 countries using statistical machine learning models: auto-regressive integrated moving average (ARIMA) and seasonal auto-regressive integrated moving average (SARIMA). Appl Soft Comput 103:107161
12. Chyon FA et al (2021) Time series analysis and predicting COVID-19 affected patients by ARIMA model using machine learning. J Virolog Meth 114433
13. Rahimi I, Chen F, Gandomi AH (2021) A review on COVID-19 forecasting models. Neural Comput Appl 1–11
14. The dataset used in the paper can be found publicly at https://www.kaggle.com/imdevskp/covid19-corona-virus-india-dataset?select=nation_level_daily.csv
15. Jain G, Prasad RR (2020) Machine learning prophet and XGBoost algorithm: analysis of traffic forecasting in telecom networks with time series data. In: 2020 8th international conference on reliability infocom technologies and optimization (trends and future directions) (ICRITO), pp 893–897
16. Ghafouri-Fard S, et al (2021) Application of machine learning in the prediction of COVID-19 daily new cases: a scoping review. Heliyon 7(10):e08143
17. Maleki M, et al. (2020) Modeling and forecasting the spread and death rate of coronavirus (COVID-19) in the world using time series models. Chaos, Solitons Fractals 140:110151
18. Benvenuto D, et al (2020) Application of the ARIMA model on the COVID-2019 epidemic dataset. Data Brief 29:105340
19. Kucharski AJ, et al (202) Early dynamics of transmission and control of COVID-19: a mathematical modelling study. Lancet Infect Diseases 20(5):553–558
20. Nishiura H, Linton NM, Akhmetzhanov AR (2020) Serial interval of novel coronavirus (COVID-19) infections. Int J Infect Dis 93:284–286
21. Kamarudin ANA, et al (2021) Prediction of COVID-19 Cases in Malaysia by using machine learning: a preliminary testing. In: 2021 international conference of women in data science at Taif University (WiDSTaif). IEEE