Int. j. inf. tecnol.
https://doi.org/10.1007/s41870-020-00484-y

ORIGINAL RESEARCH

Data analysis of COVID-2019 epidemic using machine learning methods: a case study of India

Ramjeet Singh Yadav[1]

Received: 25 April 2020 / Accepted: 19 May 2020
© Bharati Vidyapeeth's Institute of Computer Applications and Management 2020

Abstract At this time, COVID-2019 is spreading its foot in the form of a huge epidemic for the world. This epidemic is spreading its foot very fast in India too. One of the World Health Organization states that COVID-2019 is a serious disease that spreads from one person to another at very fast speed through contact routes and respiratory drops. On this day, India and the world should rise to an effective step to analyze this disease and eliminate the effects of this epidemic. In this paper presented, the growing database of COVID-2019 has been analyzed from March 1, 2020, to April 11, 2020, and the next one is predicted for the number of patients suffering from the rising COVID-2019. Different regression analysis models have been utilized for data analysis of COVID-2019 of India based on data stored by Kaggle in between 1 March 2020 to 11 April 2020. In this study, we have been utilized six regression analysis based models namely quadratic, third degree, fourth degree, fifth degree, sixth degree, and exponential polynomial respectively for the COVID-2019 dataset. We have calculated the root mean square of these six regression analysis models. In these six models, the root mean square error of sixth degree polynomial is very less in compared other like quadratic, third degree, fourth degree, fifth degree, and exponential polynomial. Therefore the sixth degree polynomial regression model is very good models for forecasting the next 6 days for COVID-2019 data analysis in India. In this study, we have found that the sixth degree polynomial regression models will help Indian doctors and the Government in preparing their plans in the next 7 days. Based on further regression analysis study, this model can be tuned for forecasting over long term intervals.

Keywords Regression analysis models · Machine learning · Deep learning · RMSE · COVID-19 · Corona virus · Spread exposed

1 Introduction

The full name of COVID-2019 is the Coronavirus disease of 2019, which has created panic in the whole world today [1, 2]. Novel COVID-2019 has been reported to be the most harmful and dangerous in the world since the 1918 H1N1 influenza epidemic. Based on the report of the World Health Organization, by April 10, 2020, a total of 15,225,252 case reports were filed and a total of 100,075 deaths occurred. Thus, it can be said that COVID-2019 has been spreading very fast since the first December 2020 to till date. Till now COVID-2019 has spread in 172 countries. At present, the highest number of cases has been found in the United States of America (USA). COVID-2019 is a terrible contagious disease that results in very rapid movement from one person to another people. The COVID-2019 epidemic is a member of the family of Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2). Thus here it can be said that coronavirus is a contagious disease.

The invention of the coronavirus was first discovered in 2002 and 2012 from China and Saudi Arabia respectively. Corona is a family of viruses that is responsible for diseases ranging from cold, cough, respiratory diseases and life-threatening diseases such as Middle East Respiratory

✉ Ramjeet Singh Yadav
ramjeetsinghy@gmail.com

[1] Department of Computer Science and Engineering, Ashoka Instutute of Technology and Management, Varanasi, Uttar Pradesh 221007, India

Published online: 26 May 2020

Syndrome (MERS) and Severe Acute Respiratory Syndrome (SARS). COVID-2019 was first invented by China in mid-December 2020. It was first found in the city of Wuhan, China [2]. According to some media reports, this COVID-2019 was found in China in mid-19 November. Therefore, we can say here that China did not reveal the correct information about this virus to the countries of the world. This is a serious matter.

In a study with Jiang and his colleagues, it was found that the fatality rate for COVID-2019 is around 7.5% [3]. These people have also found in their study that the fatality rate for persons in the age group of 70–79 is 8.0% whereas, for those above the age of 80 years, the fatality rate is around 14.8% [3]. This study considers individuals above the age of 50 years with the highest risk of underlying illnesses such as diabetes, Parkinson's disease, and cardiovascular.

A person suffering from COVID-2019 starts showing symptoms in 2–14 days. Due to this virus, the patient suffers from diseases like fever, cough, breathlessness, pneumonia, kidney failure, etc. [1]. The coronavirus spreads very rapidly from one human to another by respiratory drops. This virus does not live long in the air. The virus does not spread through the air because it is not alive for long in the air [3].

Machine learning is an automated machine used to analyze various types of data. Regression analysis is a part of machine learning. Machine learning is a subset of Artificial Intelligence. Today, most of the data analyzers and scientists are using machine learning for data analysis in different domains. In this proposed study, we have proposed regression analysis based quadratic, third, fourth, fifth, sixth degree, and exponential polynomial for COVID-2019 prediction in the next 7 days for Indian doctors and Indian Government. These regression analysis based models help us for doctors and the Indian government for the next 7 days plans.

In recent times, machine algorithms have proved it to be efficient in predicting healthcare data [4–7]. Nsoesie and his colleagues have provided a systematic way to predict influenza pandemic dynamics [8]. They have also studied most of the research paper regarding prediction such as regression analysis, mass action based deterministic models, prediction rules, deterministic mass action models, regression models, prediction rules, Bayesian network, SEIR model, ARIMA forecasting. The full form of SEIR is susceptible (S), exposed (E), infected (I), and resistant (R) and ARIMA forecasting is Auto-Regressive Integrated Moving Average. The study of the solution by researchers on COVID-2019 has revealed that only exploratory analysis of limited data has been done on it [8–11]. No country has yet invented any medicine to reduce the effects of the COVID-2019 epidemic and to cure the disease completely [12]. Therefore, we can say that an important part of the management of this epidemic is to reduce the peak of the epidemic. Lowering this peak of the epidemic is also called leveling the epidemic curve. Data mining researchers and data scientists are very important to explain the characteristics of COVID-2019 and to collect technology and related data for the role of this virus [13–16]. This type of study can help in making the right decision of this epidemic and make a concrete plan of its actions. So, in the end, this study shows that in the future we will be able to properly treat and reinforce the infrastructure, wellbeing, vaccine development, and such epidemics. This type of study also shows that How can we get rid of diseases in the future.

The objectives of these studies are given below:

1. Finding the rate of spread of the disease in the next 7 days with the help of regression analysis models.
2. We have developed a machine learning-based regression analysis models for exposed COVID-2019.
3. Forecast of COVID-2019 in India with the next 7 days for better management for doctors and various government organizations.

2 Machine learning

Machine learning is an automated method for data analysis in various domains like medical engineering, financial sector, business sector, educational domains other related sectors. It comes under Artificial Intelligence which teaches machines from training datasets. Through machine learning, we can identify patterns, analyze data, and make correct decisions with no human intervention or less human intervention. Machine learning is broadly categorized into three parts which are given below:

1. Supervised learning.
2. Unsupervised learning.
3. Reinforcement learning.

Superior learning means that a machine or model teaches the teacher, or in other words, we can say that the machine or model learns through a training dataset. In supervised learning, class-level information is available in the training datasets.

Whereas unsupervised learning means-learning without a teacher or in other words learning algorithms learn dynamically with help partitioning or clustering algorithm. Most of the clustering algorithms are available in literature such as K-Means, Fuzzy C-Means, hierarchical clustering methods, and so on. Reinforcement learning is a combination of supervised and unsupervised learning methods.

3 Regression analysis

Regression analysis is a part of machine learning or in other words, regression analysis is a subset of machine learning algorithms [17, 18]. It is the first machine learning algorithm. Regression analysis inventor says that "Regression analysis consists of a set of machine learning methods that allow us to predict a continuous outcome variable (Y) based on the value of one or multiple predictor variables (X). It assumes a linear relationship between the outcome and the predictor variables". Let us consider equation straight line connecting any two variables X and Y can be stated algebraically as:

$$Y = aX + b \tag{1}$$

where b is called the intercept on the y-axis and a is called the slope of the line. Here a and b are also called the parameters of regression analysis. These parameters should learn through proper learning methods.

In this proposed, we have proposed six regression analysis based models known as exponential, quadratic, third degree fourth degree, fifth degree polynomial. The description of these models is given below:

$$\mathrm{Y} = \mathrm{a}e^{bx} \tag{2}$$

$$Y = aX^2 + bX + c \tag{3}$$

$$Y = aX^3 + bX^2 + cX + d \tag{4}$$

$$Y = aX^4 + bX^3 + cX^2 + dX + e \tag{5}$$

$$Y = aX^5 + bX^4 + cX^3 + dX^2 + eX + f \tag{6}$$

$$Y = aX^6 + bX^5 + cX^4 + dX^3 + eX^2 + f + g \tag{7}$$

where a, b, c, d, e, f and g are called the parameters of regression analysis.

3.1 Correlation coefficients

The strength of a linear relationship between two variables is known as the Correlation coefficient means. According to Karl Pearson, the coefficient of correlation is a measure or degree of the linear relationship between two variables. Karl Pearson has been developed a formula known as Correlation Coefficient. The correlation coefficient between two random variables X and Y, usually denoted by is a numerical measure of the linear relationship between them and is defined as:

$$r(X, Y) = \frac{Cov(X, Y)}{\sigma_X \sigma_Y} \tag{8}$$

where $Cov(X, Y), \sigma_X$ and σ_Y is defined by the following formulae:

$$Cov(X, Y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}) \tag{9}$$

$$\sigma_X = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 \tag{10}$$

$$\sigma_Y = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2 \tag{11}$$

Here $(x_i, y_i), i = 1, 2, 3, 4 \ldots, N$, is the set of input and output variables. Here there are some prediction is given below:

1. If the value of the correlation coefficient is zero, it means there is no correlation between input variables X and output variable Y.
2. If the value of the correlation coefficient is equal to positive one. It means there is a strong relation between the input variable and the output variable. In other words, if the input variable is increased then the output variable is also increased.
3. If the value of correlation coefficient is equal to negative. It means the input variable is increased then the output variable is also decreased and vice versa.

Two variables that have a small or no linear correlation might have a strong nonlinear relationship. However, calculating linear correlation before fitting a model is a useful way to identify variables that have a simple relationship. In this proposed study, first of all, we have calculated the correlation coefficient between date and number of confirmed cases of COVID-2019 spread up of India between 1st March 2020 to 11 April 2020 [19, 20]. The correlation coefficient between date and number of confirmed cases are as follows:

$$r = \begin{bmatrix} 1.0000 & 0.8157 \\ 0.8157 & 1.0000 \end{bmatrix}$$

In the matrix, the diagonal elements represent the perfect correlation of the input variable (Date) and the output variables (confirmed cases) of COVID-2019 spared up of India with itself and are equal to 1. The off-diagonal elements are very close to 1, indicating that there is a strong statistical correlation between the variables Date and number of confirmed case of COVI-2019 spread up datasets in India.

3.2 Residuals and goodness of fit

The difference between the observed value of the response variable *(Y)* and the value of proposed model is known as residuals. This is the measure of goodness of fit a straight lines of the proposed models. Residuals are the difference between observed values and values of the proposed

models. The following formula is used for the calculation of residuals:

$$Residuals = Y_{Observed} - Y_{ModelesValues} \quad (12)$$

Another formula also measure the goodness of fit is known as R^2 and defined as the following formula:

$$R^2 = 1 - \frac{SS_{Residual}}{SS_{Total}} \quad (13)$$

where $SS_{Residual}$ represent the sum of the squared residual from the regression analysis and SS_{Total} represent the sum of the squared difference from the mean of the dependent variables. The sum of the squared residuals from the regression and the sum of the squared differences from the mean of the dependent variables both are positive.

3.3 Adjusted R^2 for polynomial regressions analysis

In the proposed study, the two or more polynomials have been used for data analysis of COVID-2019. Therefore, in this study, the residuals in a model can be reduced by fitting a high degree of polynomial. Adjusted R^2 for polynomial regression is defined as the following formula:

$$R^2_{Adjusted} = 1 - \left(\frac{SS_{Residual}}{SS_{Total}}\right) * \left(\frac{(n-1)}{(n-d-1)}\right) \quad (14)$$

where n is the number of observations in COVID-2019 data training datasets and d is the degree of polynomials of proposed regression analysis models. In this proposed study, we have compute the both simple and adjusted R^2 to evaluate whether the extra terms n and d terms improve the predictive power of proposed methods.

4 Experimental results and discussion

In this proposed study, we have taken the COVID-2019 outbreak dataset from India. The first case of the COVID-2019 epidemic was found in Kerala state of India in January 2020. At that time, three the COVIDs cases in Kerala were infected with the the COVID-2019 epidemic. All three patients came from the city of Wuhan in China at that time. However, things escalated in March, after several cases were reported all over the country, most of whom had a travel history to other countries. The first outbreak of the COVID-2019 epidemic was to begin in India in early March and by 20 March the number had risen to about 282. For the first time, the Prime Minister of India, Shri Narendra Modi, addressed the nation about the COVID-2019 pandemic on 19 March 2020 and announced a public curfew on 22 March 2020. After this, the Prime Minister of India again addressed the name of the nation for the second time about the COVID-2019 on March 22 and locked India down from March 25 to April 14, 2020. The growth of the COVID-2019 epidemic in India is going on in exponential form from 20 March 2020 to 10 April 2020. Even today, the outbreak of this epidemic is happening in exponential form.

We can do a machine learning based regression analysis methods for data analysis to create a model based on regression analysis that helps in the forecast next 7 days for the COVID-2019 outbreak in India. The whole dataset of the COVID-2019 outbreak of India is available on Kaggle and World Health Organization (WHO) website [19, 20]. In this study, we have used MATLAB software for programming and data analysis. Figure 1 shows the number of cases detected from 1st March 2020 11th April 2020. Different regression analysis models have been utilized for

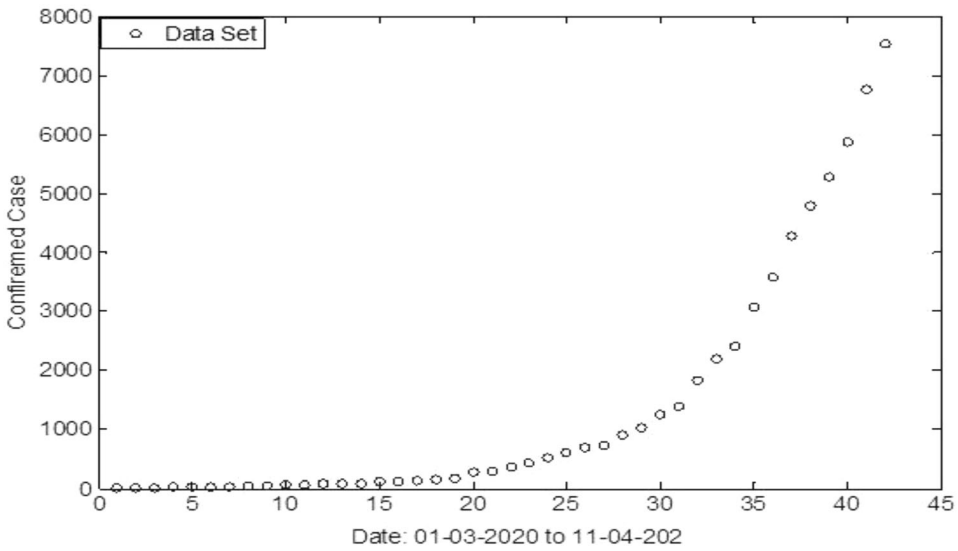

Fig. 1 Scatter plot of traning dataset of the COVID-2019 in India

Table 1 Training dataset of COVID-2019 of India during 1st March 2020 to 11th April 2020

Date	Confirmed cases
1-Mar	3
2-Mar	5
3-Mar	6
4-Mar	28
5-Mar	30
6-Mar	31
7-Mar	34
8-Mar	39
9-Mar	46
10-Mar	58
11-Mar	60
12-Mar	74
13-Mar	81
14-Mar	84
15-Mar	110
16-Mar	114
17-Mar	137
18-Mar	151
19-Mar	173
20-Mar	273
21-Mar	283
22-Mar	360
23-Mar	433
24-Mar	519
25-Mar	606
26-Mar	694
27-Mar	724
28-Mar	909
29-Mar	1024
30-Mar	1251
31-Mar	1397
1-Apr	1834
2-Apr	2201
3-Apr	2415
4-Apr	3072
5-Apr	3577
6-Apr	4281
7-Apr	4789
8-Apr	5274
9-Apr	5865
10-Apr	6761
11-Apr	7529

data analysis of the COVID-2019 of India based on data stored by Kaggle in between 1 March 2020 to 11 April 2020. In this study, we have been utilized six regression analysis based models namely quadratic, third degree, fourth degree, fifth degree, sixth degree and exponential polynomial respectively for the COVID-2019 dataset.

Table 2 Testing dataset of the COVID-2019 of India during 12th April 2020 to 19th April 2020

S.No.	Date	Confirmed case (actual result)
1	12-April	8447
2	13-April	9352
3	14-April	10,815
4	15-April	11,933
5	16-April	12,759
6	17-April	13,835
7	18-April	17,792
8	19-April	16,116

Table 1 and Table 2 shows the training and datasets of the COVID-2019 outbreak of India during 1st March 2020 to 11th April 2020 and testing dataset during 12 April 2020 to 19 April 2020.

The analysis is based on the date data and confirmed cased data of whole India as presented in Table 1. In this regard, the regression calculations between date and confirmed cases parameter have been done for dataset from 1st March 2020 to 11th April 2020 [19, 20]. For the purpose of experimental results, we have used linear regression models like quadratic to sixth degree polynomial and exponential polynomial. In these proposed regression models, we have used date (say X) as independent variable and number of confirmed cases consider as dependent variable (say Y) or predictor. The proposed linear regression models equation are given below:

$$\mathrm{Y} = 18.74 * e^{0.14x} \tag{15}$$

$$Y = 8.1572 * X^2 - 214.7599 * X + 1013.4 \tag{16}$$

$$Y = 144.4802 * X^3 + 597.7748 * X^2 + 865.6646 * X + 618.8247 \tag{17}$$

$$Y = 144.4802 * X^4 + 597.7748 * X^3 + 865.6646 * X^2 + 618.8247 * X + 272.4252 \tag{18}$$

$$Y = -52.17 * X^5 + 144.48 * X^4 + 766.97 * X^3 + 865.66 * X^2 + 512.09 * X + 272.43 \tag{19}$$

$$Y = -90.92 * X^6 \pm 52.17 * X^5 + 505.65 * X^4 + 766.97 * X^3 + 515.23 * X^2 + 513.09 * X + 320.96 \tag{20}$$

Equation 15–20 represent the exponenetial, quadratic, third dgree, fourth degree, fifth degree and sixth degree polynomial equations.

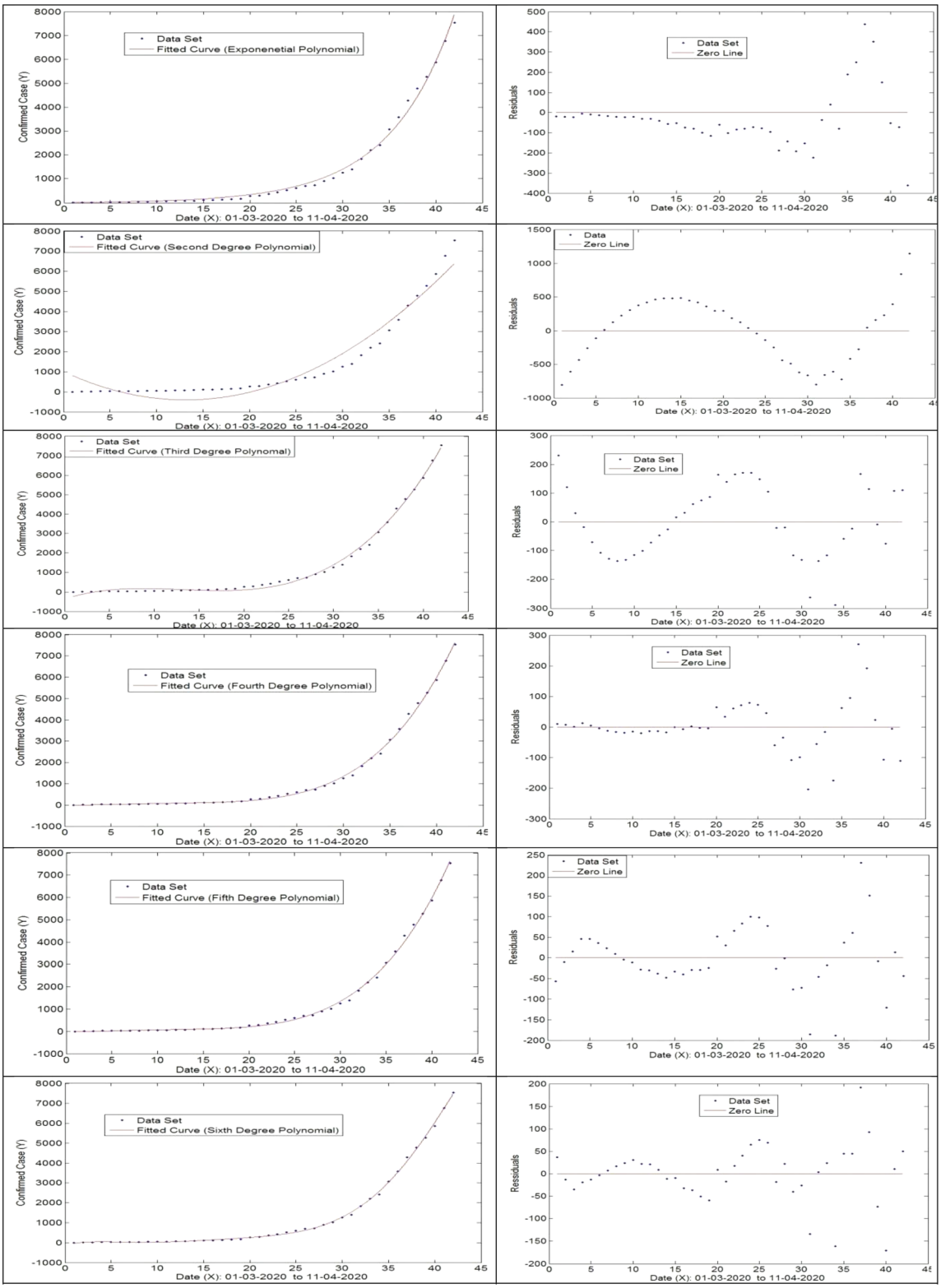

Fig. 2 **a** Fitted curve with data of exponential, quadratic, third degree, fourth degree, fifth degree polynomial. **b** Residuals with dataset of exponential, quadratic, third degree, fourth degree, fifth degree polynomial

Table 3 The results of SSE, R^2, DFE and adjusted R^2

S. no.	Model	SSE	R^2	DFE	Adjusted R^2
1	Exponential polynomial	845,000	0.9951	40	0.9950
2	Quadratic polynomial	9,209,100	0.9463	39	0.9436
3	Third degree polynomial	6,466,400	0.9962	38	0.9959
4	Fourth degree polynomial	2,777,700	0.9984	37	0.9982
5	fifth degree polynomial	2,426,900	0.9986	36	0.9984
6	Sixth degree polynomial	1,656,800	0.9990	35	0.9989

Figure 2a shows the results of confirmed cases of the proposed fitted regression analysis based models namely exponential, quadratic, third degree, fourth degree, fifth degree and sixth degree polynomials for the training datasets. In the proposed study, we have plotted all the calculated residuals of the proposed models namely exponential, quadratic, third degree, fourth degree, fifth degree and sixth degree of polynomial. In regression analysis, residuals play an important role the COVID-2019 outbreak data analysis in India. Figure 2a shows the residuals for the proposed methods namely exponential, quadratic, third degree, fourth degree, fifth degree and sixth degree polynomials. These Fig. 2a, b also shows that the sixth degree polynomial for fitted result and residuals, respectively and gives better results in comparison to other like quadratic, third degree, fourth degree and fifth degree polynomial fitted results and residual. In this study, the quadratic polynomial gives unsatisfactory results, while exponential, third degree, fourth degree, fifth degree and six degree polynomials give better and satisfactory results on how to fit the dataset of the COVID-2019 in India (Fig. 2a). Figure 2b shows the residual of exponential polynomial fit gives the strongly pattern behavior while other polynomial fit residuals still strongly patterned.

Regarding the best fit of the proposed models, we have calculate the R^2 and adjusted R^2 for the proposed models. Table 3 shows the calculated results of the Sum of Square Errors (SSR), R^2, Degree of Freedom for Error (DFE) and adjusted R^2. Table 3 also shows that sixth degree polynomial based regression analysis model has lowest values of SSE, R^2, DFE and adjusted R^2 in comparison to other models. It means proposed regression analysis based sixth degree polynomial gives better results for the prediction or forecast the COVID-2019 outbreak in India in comparison to other models like exponential, quadratic, third degree, fourth degree, fifth degree polynomial regression models.

Tables 4 show the results of the COVID-2019 outbreak training datasets of India during 1st March 2020 to 11 April 2020. The last column of this table shows that results of proposed sixth degree polynomial. Because, according to above discussion here we have found that regression analysis based sixth degree polynomial gives better result for predicting the outbreak of the COVID-2019 in India to next 7 days. Table 5 shows the results of the COVID-2019 outbreak testing datasets of Indian during 11th April 2020 to 19 April 2020. In the last column of Table 5 also shows the predicted or fitted results of proposed sixth degree polynomial.

Figure 3 shows Comparison of Confirmed Case (Actual Result) and Results of the Proposed Model Sixth Degree Polynomial (predicted results) for training dataset of the COVID-2019. This figure is also shows that the result of the proposed sixth degree polynomial method is very close to confirmed cases (actual results).

The above Fig. 4 shows a comparison of the confirmed case (actual result) and results of the proposed model sixth degree polynomial (predicted results) for the training dataset of the COVID-2019. This figure also shows that the result of the proposed sixth degree polynomial method is very close to confirmed cases (actual results). Therefore the proposed method is very useful for future prediction of the COVID-2019 outbreak to the next 7 days from the current date.

5 Conclusion

In this paper, we have proposed six regression analysis based machine learning models for prediction of the COVID-2019 outbreak datasets of India. These models

Table 4 Training datasets analysis of the COVID-2019 of India during 1st March 2020 to 11th April 2020

S. no.	Date	Confirmed case (actual result)	Results of the proposed model sixth degree polynomial (predicted results)
1	1-Mar	3	-33.6
2	2-Mar	5	17.8
3	3-Mar	6	41.3
4	4-Mar	28	47.0
5	5-Mar	30	42.9
6	6-Mar	31	34.9
7	7-Mar	34	27.2
8	8-Mar	39	22.5
9	9-Mar	46	22.5
10	10-Mar	58	27.5
11	11-Mar	60	37.9
12	12-Mar	74	53.2
13	13-Mar	81	72.5
14	14-Mar	84	94.9
15	15-Mar	110	120
16	16-Mar	114	146
17	17-Mar	137	174
18	18-Mar	151	202
19	19-Mar	173	232
20	20-Mar	273	265
21	21-Mar	283	301
22	22-Mar	360	343
23	23-Mar	433	393
24	24-Mar	519	454
25	25-Mar	606	530
26	26-Mar	694	625
27	27-Mar	724	742
28	28-Mar	909	887
29	29-Mar	1024	1060
30	30-Mar	1251	1280
31	31-Mar	1397	1530
32	1-Apr	1834	1830
33	2-Apr	2201	2180
34	3-Apr	2415	2580
35	4-Apr	3072	3030
36	5-Apr	3577	3530
37	6-Apr	4281	4090
38	7-Apr	4789	4700
39	8-Apr	5274	5350
40	9-Apr	5865	6040
40	10-Apr	6761	6750
42	11-Apr	7529	7480

Table 5 Testing datasets analysis of the COVID-2019 of India during 12th April 2020 to 19th April 2020

S. no.	Date	Confirmed case (actual result)	Results of proposed model sixth degree polynomial (predicted results)
1	12-April	8447	8301
2	13-April	9352	9254
3	14-April	10,815	9960
4	15-April	11,933	11,600
5	16-April	12,759	12,953
6	17-April	13,835	13,975
7	18-April	17,792	17,490

basically regression analysis based exponential, quadratic, third degree, fourth degree, fifth degree and sixth degree polynomials. These models also predict the outbreak of the COVID-2019 in India for the next 7 days. After analyzing the COVID-2019 outbreak datasets on India between 1st March 2020 to 11th April 2020 and predict the results to the next 7 days with the help testing datasets from 12th April 2020 to 19 April 2020. Here, we have find out that the value of for proposed models namely sixth degree polynomial is very close to the confirmed case or actual results regarding training dataset of the COVID-2019. According to Table 3, the value residuals of sixth degree polynomial are higher in comparison to the residual of other proposed models. It means this model achieved best fitted results for COVID-2019 datasets of India. Therefore, here we can says that the proposed regression analysis based sixth degree polynomial gives better results of the COVID-2019 outbreak training and testing datasets of India. Table 5 shows the prediction results of the COVID-2019 outbreak results of the next 7 days. This table also shows that the very little difference between confirmed results and predicted results for the COVID-2019 outbreak of India. In the last, this proposed study is very useful for Indian doctors and the Indian government for managing the COVID-2019 outbreak for the next 7 days. In the future, we will develop a regression analysis based on artificial neural networks that can be developed to obtain data at regular intervals. This model will automatically estimate the number of cases of weekly and bi-weekly data. Therefore, we can say that the Indian government and

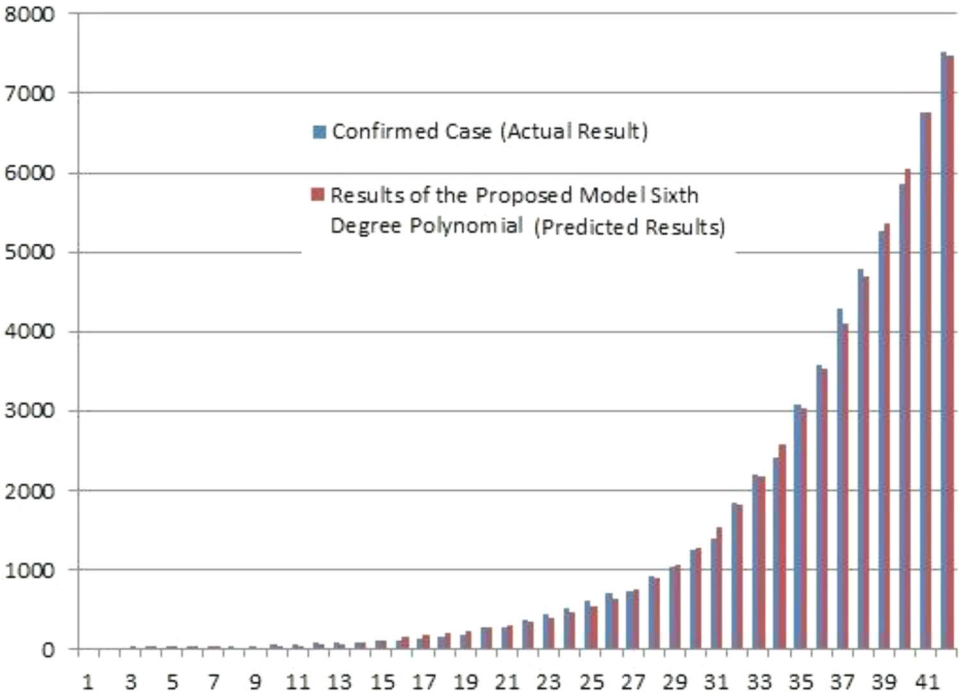

Fig. 3 Comparison of confirmed case (actual result) and results of the proposed model sixth degree polynomial (predicted results) for training dataset of the COVID-2019

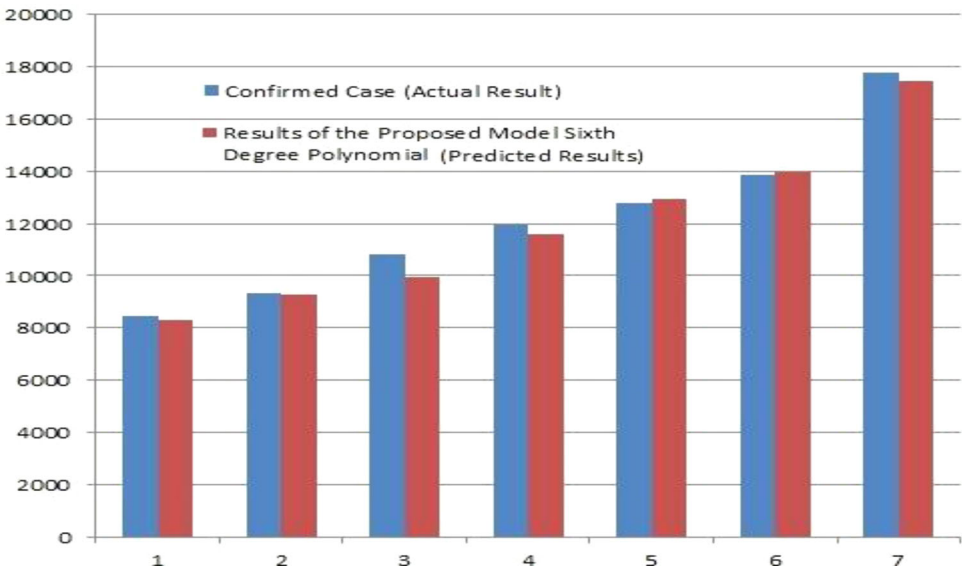

Fig. 4 Comparison of confirmed case (actual result) and the results of the proposed model sixth degree polynomial (predicted results) for testing dataset of the COVID-2019

doctors can maintain a check on hospital facilities, necessary supplies for new patients, medical aid, and isolation for next week or in the future.

References

1. World Health Organization (2020) Coronavirus disease 2019 (COVID19): situation report, p 67
2. Stoecklin SB, Patrick R, Yassoungo S, Alexandra M, Christine C, Anne S, Matthieu M et al (2020) First cases of coronavirus disease 2019 (COVID-19) in France: surveillance, investigations and control measures. Eurosurveillance 25:6
3. Wu Z, McGoogan JM (2020) Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China: summary of a report of 72 314 cases from the Chinese Center for Disease Control and Prevention. JAMA
4. Jiang F, Deng L, Zhang L, Cai Y, Cheung CW, Xia Z (2020) Review of the clinical characteristics of coronavirus disease 2019 (COVID-19). J Gen Int Med 35:1545–1549. https://doi.org/10.1007/s11606-020-05762-w
5. Ye QH, Qin LX, Forgues M, He P, Kim JW, Peng AC, Simon R, Li Y, Robles AI, Chen Y, Ma ZC (2003) Predicting hepatitis B virus-positive metastatic hepatocellular carcinomas using gene expression profiling and supervised machine learning. Nat Med 9(4):416–423
6. Mai MV, Krauthammer M (2016) Controlling testing volume for respiratory viruses using machine learning and text mining. In:

AMIA annual symposium proceedings, vol 2016. American Medical Informatics Association, p 1910
7. Purcaro G, Rees CA, Wieland-Alter WF, Schneider MJ, Wang X, Stefanuto PH, Wright PF, Enelow RI, Hill JE (2018) Volatile fingerprinting of human respiratory viruses from cell culture. J Breath Res 12(2):026015
8. Kalipe G, Gautham V, Behera RK (2018) Predicting malarial outbreak using machine learning and deep learning approach: a review and analysis. In: 2018 International conference on information technology (ICIT). IEEE, pp 33–38
9. Nsoesie EO, Brownstein JS, Ramakrishnan N, Marathe MV (2014) A systematic review of studies on forecasting the dynamics of influenza outbreaks. Influenza Respir Viruses 8(3):309–316
10. Pirouz B, ShaffieeHaghshenas S, Shaffiee Haghshenas S, Piro P (2020) Investigating a serious challenge in the sustainable development process: analysis of confirmed cases of COVID-19 (new type of coronavirus) through a binary classification using artificial intelligence and regression analysis. Sustainability 12(6):2427
11. More GD, Dunowska M, Acke E, Cave NJ (2020) A serological survey of canine respiratory coronavirus in New Zealand. N Z Vet J 68(1):54–59
12. Wu C, Chen X, Cai Y, Zhou X, Xu S, Huang H, Zhang L, Zhou X, Du C, Zhang Y, Song J (2020) Risk factors associated with acute respiratory distress syndrome and death in patients with coronavirus disease 2019 pneumonia in Wuhan. JAMA Internal Medicine, Wuhan
13. Deb S, Manidipa M (2020) A time series method to analyze incidence pattern and estimate reproduction number of COVID-19. ARXiv preprint ARXiv: 2003.10655
14. Mandal S, Bhatnagar T, Arinaminpathy N, Agarwal A, Chowdhury A, Murhekar M, Gangakhedkar RR, Sarkar S (2020) Prudent public health intervention strategies to control the coronavirus disease 2019 transmission in India: a mathematical model based approach. Indian J Med Res 151:190
15. Dong E, Du H, Gardner L (2020) An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect Dis. https://doi.org/10.1016/S1473-3099(20)30120-1
16. Gupta R, Pal SK (2020) Trend analysis and forecasting of COVID-19 outbreak in India. Published online 30/03/2020 in preprint archive
17. Ting DSW, Lawrence C, Victor D, Wong TY (2020) Digital technology and COVID-19. Nat Med 26:459–461. https://doi.org/10.1038/s41591-020-0824-5
18. Benvenuto D, Marta G, Lazzaro V, Silvia A, Massimo C (2020) Application of the ARIMA model on the COVID-2019 epidemic dataset. Data in brief 105340. MedRxiv. https://www.medrxiv.org/content/10.1101/2020.03.26.20044511v1
19. India Coronaviruses (COVID-19) Datasets (2020). Retrieved 13 Apr 2020. https://www.kaggle.com/sudalairajkumar/covid19-in-india/data
20. WHO Coronaviruses (COVID-19) (2020). Retrieved 30 Mar 2020. https://www.who.int/emergencies/diseases/novel-coronavirus-2019