Diabetes & Metabolic Syndrome: Clinical Research & Reviews 14 (2020) 1467–1474

Contents lists available at ScienceDirect

Diabetes & Metabolic Syndrome: Clinical Research & Reviews

journal homepage: www.elsevier.com/locate/dsx

Prediction of new active cases of coronavirus disease (COVID-19) pandemic using multiple linear regression model

Smita Rath [a,*], Alakananda Tripathy [a], Alok Ranjan Tripathy [b]

[a] Department of Computer Science and Engineering, Siksha 'O' Anusandhan Deemed to be University, Odisha, India
[b] Department of Computer Science,Ravenshaw University, Cuttack, India

ARTICLE INFO

Article history:
Received 16 July 2020
Received in revised form
26 July 2020
Accepted 28 July 2020

Keywords:
Coronavirus
India
Odisha
Correlation coefficient
Linear regression
Multiple linear regression

ABSTRACT

Introduction and Aims: The COVID-19 pandemic originated from the city of Wuhan of China has highly affected the health, socio-economic and financial matters of the different countries of the world. India is one of the countries which is affected by the disease and thousands of people on daily basis are getting infected. In this paper, an analysis of daily statistics of people affected by the disease are taken into account to predict the next days trend in the active cases in Odisha as well as India.

Material and methods: A valid global data set is collected from the WHO daily statistics and correlation among the total confirmed, active, deceased, positive cases are stated in this paper. Regression model such as Linear and Multiple Linear Regression techniques are applied to the data set to visualize the trend of the affected cases.

Results: Here a comparison of Linear Regression and Multiple Linear Regression model is performed where the score of the model R^2 tends to be 0.99 and 1.0 which indicates a strong prediction model to forecast the next coming days active cases. Using the Multiple Linear Regression model as on July month, the forecast value of 52,290 active cases are predicted towards the next month of 15th August in India and 9,358 active cases in Odisha if situation continues like this way.

Conclusion: These models acquired remarkable accuracy in COVID-19 recognition. A strong correlation factor determines the relationship among the dependent (active) with the independent variables (positive, deceased, recovered).

1. Introduction

In the beginning of 2020, the first case of COVID-19 pandemic in India was reported on January 30, 2020. COVID-19 is corona virus disease caused by SARS-CoV 2 (severe acute respiratory syndrome coronavirus 2). The novel corona virus was first originated from the wet market of Wuhan a city in China [1]. These plays a havoc on human by claiming 523,011 lives worldwide according to the world health organization [2]. The virus spread among the people more often through small droplets released by coughing, sneezing and talking in the close contact [3]. Instead of moving long distance through air the droplet falls onto the ground or surface. The basic symptoms of the COVID-19 are fever, cough, shortness of breath, loss of sense and fatigue [4]. Other symptoms include breathing difficulty and chest pain. To prevent the spreading of virus number of measures are being carried out like personal hygiene, washing hand frequently with soap and water, using face mask and making social distancing. In order to prevent the transmission of virus many countries impose shutdown and lockdown. The first case of COVID-19 is detected in January 30, 2020. In India, the COVID-19 has huge impact. According to the world health organization report [2] the number of cases in India is 793, 802 confirmed cases of the virus as on 11 July the total number of samples tested so far is more than one crore. The number of fatalities due to COVID-19 pandemic is 21,604 till date.

As per the government of India report [5] the worst effected states and union territory are Maharashtra with 2,30,599 cases and the number of deaths is 9667, Tamil Nadu with 1,26,581 cases followed by Delhi with 1,07, 051. India declares nationwide lockdown to stop the exponentially growth of infection that affected in other countries like Italy [6]. The nationwide lock down is made in order to flatten the infection curve in India. The focus of the paper lies in finding out each daily active cases or new confirmed COVID-19 cases using a regression model that will be helpful in forecasting the next day's scenario of the country. The objective was to identify

* Corresponding author.
E-mail addresses: smitarath@soa.ac.in (S. Rath), alakanandatripathy@soa.ac.in (A. Tripathy), tripathyalok@gmail.com (A.R. Tripathy).

https://doi.org/10.1016/j.dsx.2020.07.045
1871-4021/

Table 1
Correlation table of Odisha daily Covid-19 cases.

	Positive	*Active*	*Recovered*	*Deceased*
Positive	1			
Active	0.997739	1		
Recovered	0.977684	0.964648	1	
Deceased	0.979267	0.978835	0.957287	1

the relation among the data collected from each day and thus could make a significant contribution to a reliable data, and project a forecasting model for India as well as Odisha. We felt this was of utmost importance, because it would help to clarify the potential plan of action as well as prepare it.

2. Materials and methods used

The whole data set was collected from WHO site https://covid19india.org and https://covidindia.org/odisha/for the daily new positive cases, active cases, deceased and recovered cases in a csv file from March 22, 2020 to July 4, 2020. The daily data of COVID- 19 cases of Odisha and India are in a form of continuous set where the active cases are dependent on the other variables as confirmed from the correlation values in Table 1 and Table 2. Correlation research aims at calculating and understanding the impact of a linear or nonlinear relationship between two continuous variables. Coefficients of association assume values ranging from negative correlations (−1) to uncorrelated (0) to positive correlations (+1). The sign of the coefficient of correlation (i.e., positive or negative) determines the direction of relation. The absolute value shows the strength of the linear relationship (Tables 1 and 2) which is very close to +1.

Initially the data cleaning process is performed on the two data set to remove any missing values. Then a correlation analysis is performed on the data sets using Python programming through Spyder of Anaconda Navigator App. Then Linear regression model is used to evaluate the relative impact of active cases due to daily positive cases in Odisha as well as in case of India data. The key goal of linear regression is to fit a straight line with the data forecasts Y

Table 2
Correlation table of India daily Covid-19 cases.

	Positive	*Recovered*	*Deceased*	*Active*
Positive	1			
Recovered	0.985837589	1		
Deceased	0.988199348	0.950408963	1	
Active	0.988313465	0.948769793	0.9985863	1

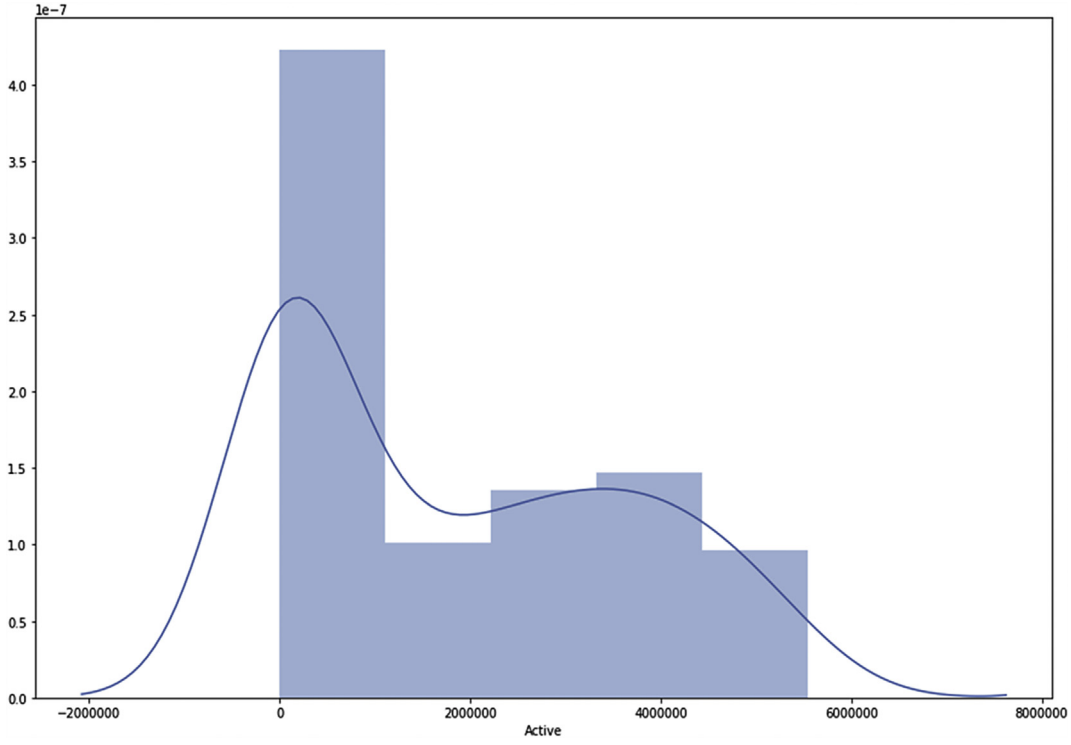

Fig. 1. Daily Active Cases of India showing the average values in curves.

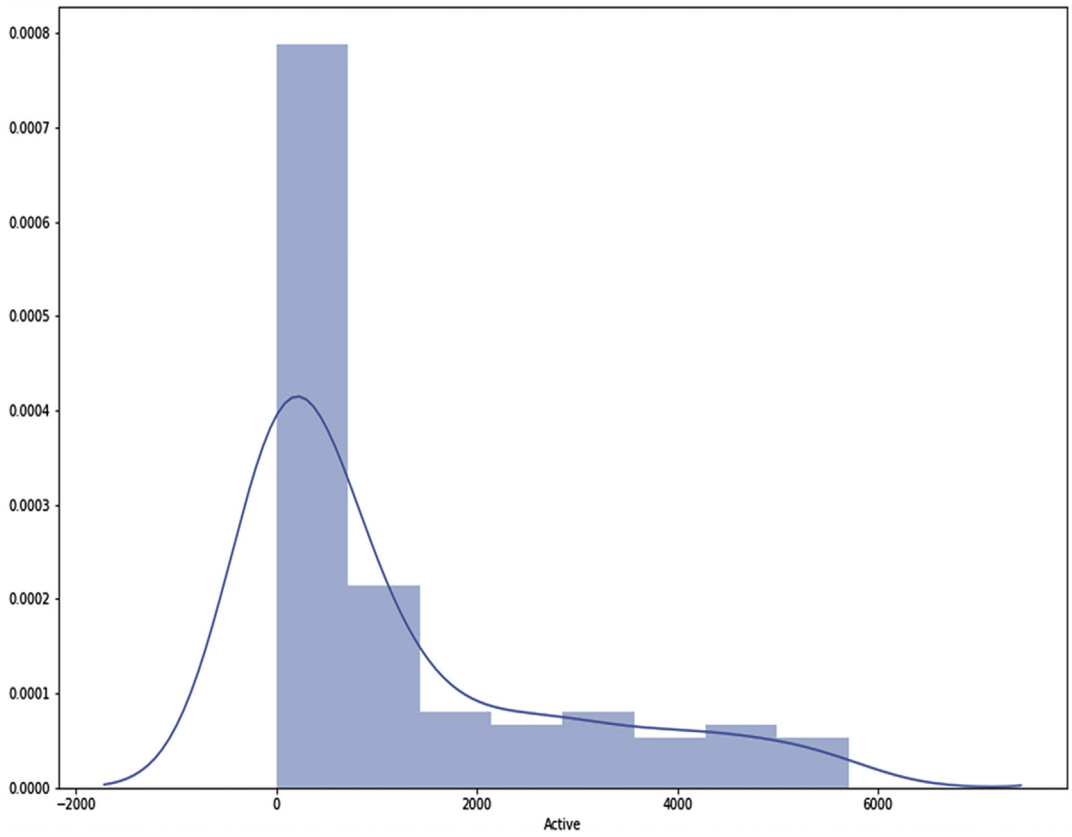

Fig. 2. Daily Active Cases of Odisha showing the peak with average value.

Table 3
Values obtained after training with linear regression prediction model.

Data set	Intercept	Coefficient	Score (R^2)	Mean Absolute Error (MAE) $\times 10^{-6}$	Mean Squared Error (MSE) $\times 10^{-6}$	Root Mean Squared Error (RMSE) $\times 10^{-6}$
Odisha	−31.97450729	0.6876260	**0.995588**	73.8606	11320.1564	106.3962238
India	202497.14638	0.5714703	**0.974855**	245838.76	74851765386.71	273590.5067

Bold indicates R-squared or Coefficient of Multiple Determination.

for X where Y is the total number of daily active cases and X is the total number of positive cases. The least squares method is commonly used to estimate the intercept and slope regression parameters which define the line. The below Figs. 1 and 2 shows the average peak values of active cases in part of Odisha as well as India.

The model can be expressed as in Eq. (1) where *Y and X* are dependent and independent variable, $\propto$ is the intercept and *b* is the regression parameter as slope and $\in$ is the random error respectively.

$$Y = \propto + bX + \in \quad (1)$$

The limitations of Linear regression are that it often explores a relation between the mean of the input variables and output variables. Just as the mean is not a full description of a single variable, linear regression is just not a clear understanding of variable relationships. Therefore, an analysis of the various factors is done using Multiple Linear Regression (MLR) model. The dependent variable (target variable) is dependent on many independent variables, in this case. You can describe a regression equation involving multiple variables as:

$$Y = \beta + \beta\_o\ x1 + \beta\_1\ x2 + \beta\_2\ x3 + \varepsilon \quad (2)$$

where Y is the predictor or target variable and $x1, x2, x3$ are the independent variables. β is the y-intercept and β_o,β_1,β_2 and ε are the coefficients and error term respectively.

3. Results

Case 1. The two data sets are first spilt into eighty percent training set and twenty percent testing set and then Linear Regression is performed to train the first 80% set. Here the number of daily active cases are predicted based on daily positive cases. Here the model generates the coefficients to find the next active cases number on the test set as shown in Table 3. As we explained, the regression line effectively selects the right value for the intercept and slope resulting in a line that fits the criteria best.

So, it can be said that for every one-unit of positive cases increase there is an increase of 68% in case of active cases in Odisha and for every one-unit increase in positive cases of India shows an increase of 57% in active cases. The score represents the value of R^2 as 0.995588 and 0.974855 value which indicate it as a strong predictor model. A comparison of actual and predicted values in both data sets can be visualized through bar graphs by taking 25 records

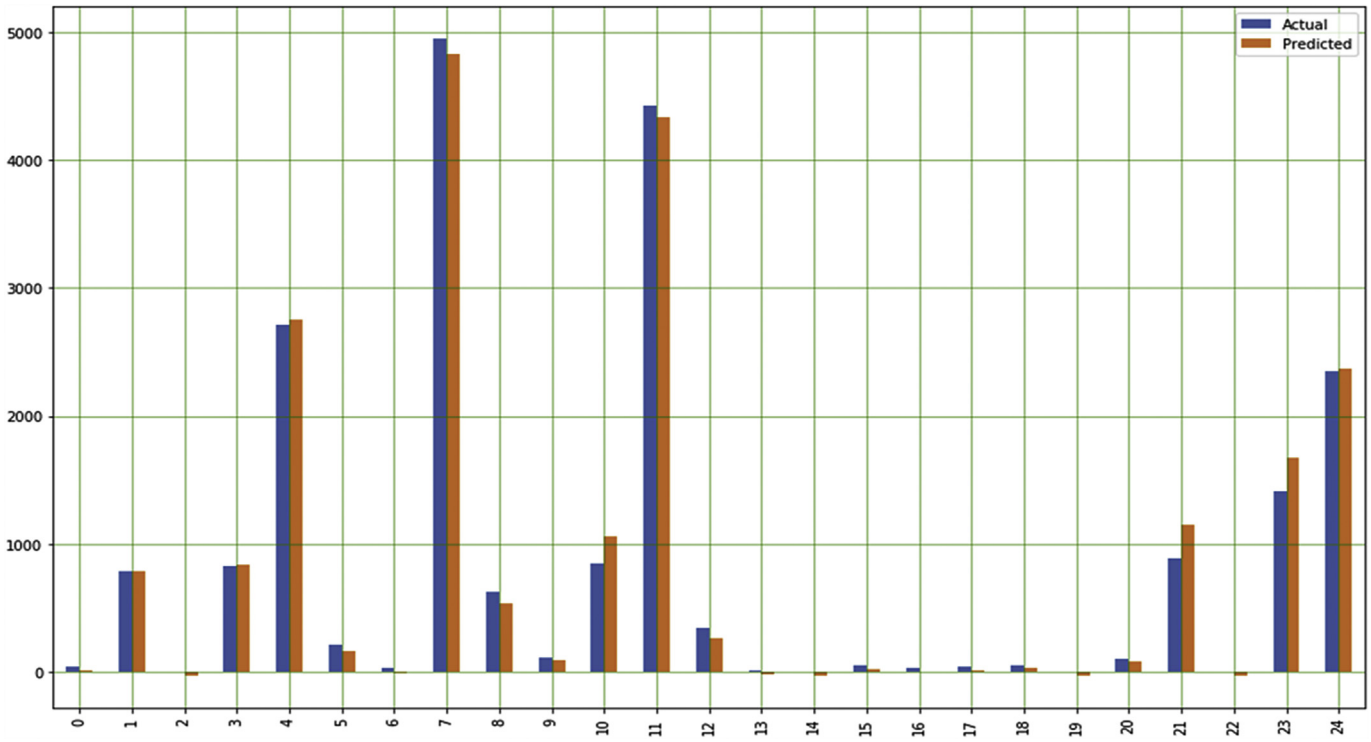

Fig. 3. Visualization of Actual vs predicted values using Linear Regression Model in Odisha COVID-19 Cases.

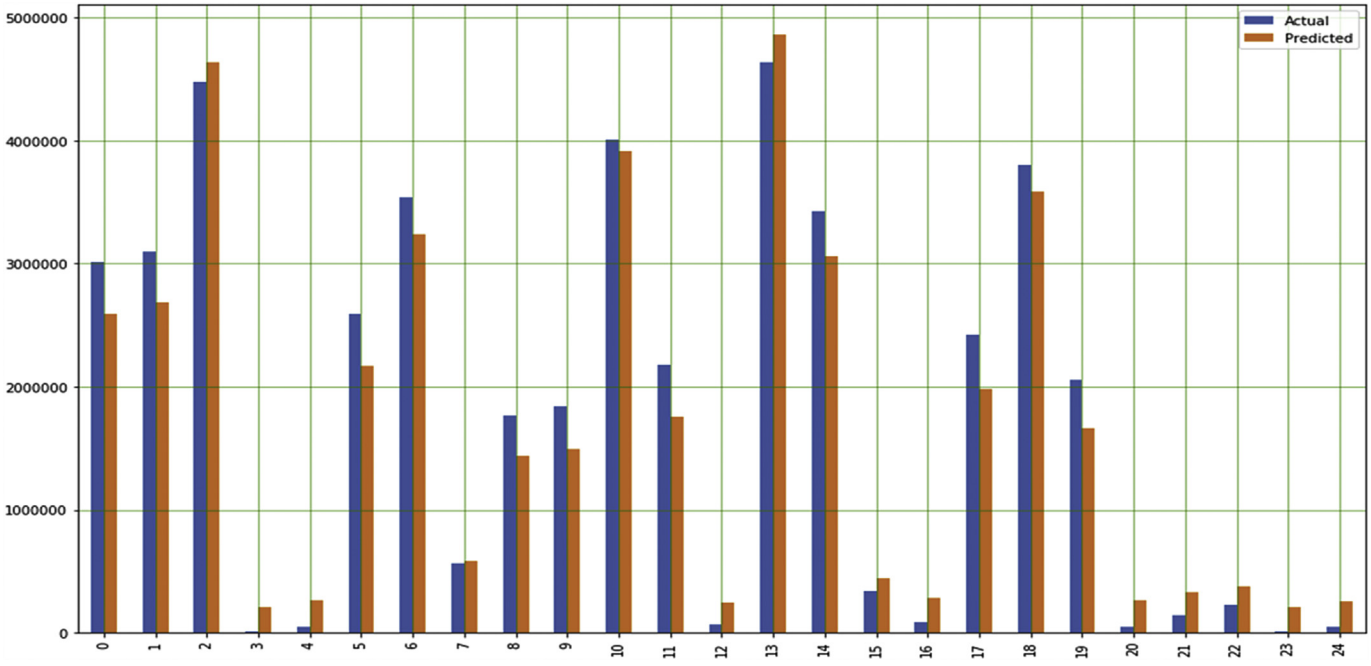

Fig. 4. Visualization of Actual vs predicted values using Linear Regression Model in India COVID-19 Cases.

from the data in the below Figs. 3 and 4.

Case 2. Linear regression comprising various variables is named linear multiple regression. The steps for multiple linear regression are nearly similar to those for simple linear regression. The distinction lies with estimation. You can use this to find out how factor does have the maximum impact on the output forecasted and how independent factors are interrelated. Here again the whole data is spilt into training and test data set to perform multiple linear regression. The inputs are daily positive cases, recovered and deceased cases to predict the daily number of active cases. We can derive a relation between the above variables using the correlation factor from the above Tables 1 and 2 Table 4 represents the MAE, MSE, RMSE, intercept, Score and Coefficient for the predictor model.

Table 4
Values obtained after training with multiple linear regression prediction model.

Data set	Intercept	Coefficient (β_o, β_1, β_2)	Score (R^2)	Mean Absolute Error (MAE)	Mean Squared Error (MSE)	Root Mean Squared Error (RMSE)
Odisha	−21.972463	[0.78035157, −0.45427124, 13.57489689]	**0.9985**	45.5734759e-06	3826.5455326712213e-06	61.85907801342679e-06
India	−3.259629011154175e-09	[-1,1,1]	**1.0**	2.6540334374658414e-09	7.735857208018085e-18	2.7813409010795647e-09

Bold indicates R-squared or Coefficient of Multiple Determination.

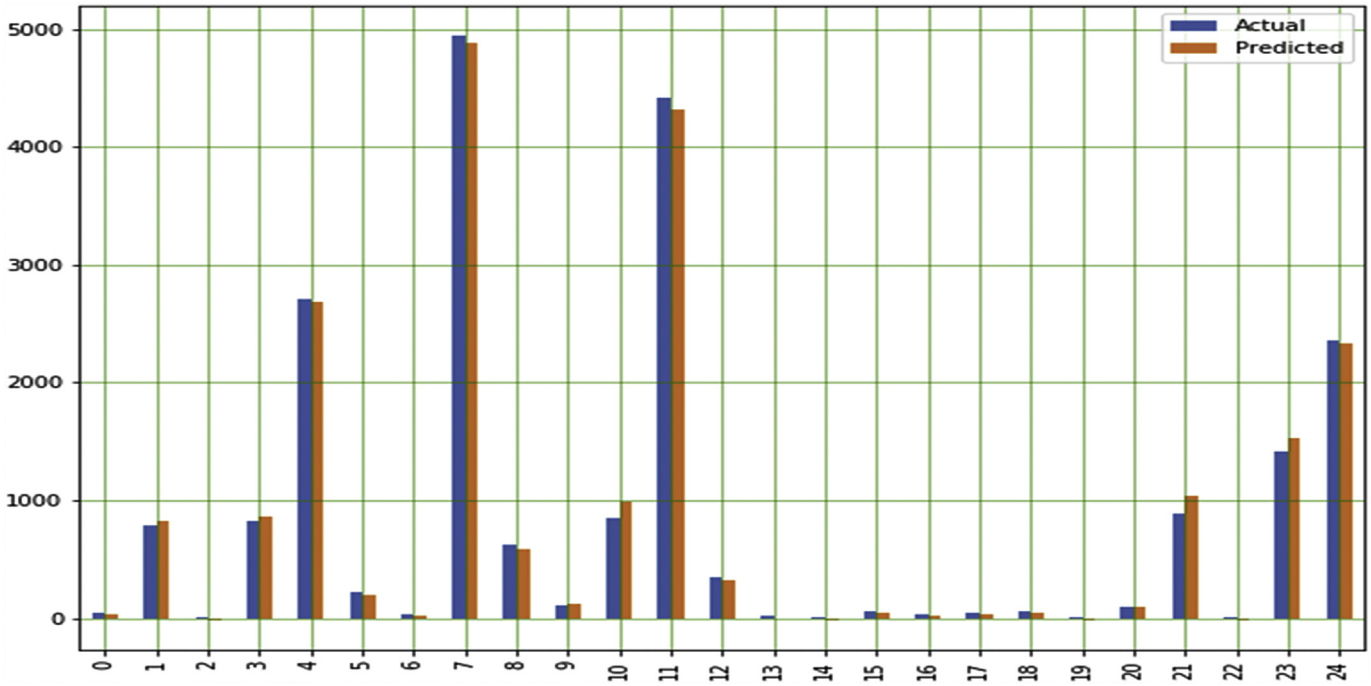

Fig. 5. Visualization of Actual vs predicted values using Multiple Linear Regression Model in Odisha COVID-19 Cases.

The R^2 value shows the predictor Multiple Linear Regression model to be more accurate as compared to the results obtained using Linear Regression model. From the above Table 4, an equation is established as follows for predicting the next day active cases as

$$Y\_odisha = -21.97 + 0.78 \times x1 \pm 0.45x2 + 13.57x3 \quad (3)$$

and

$$Y_{india} = -3.259 + 1 \times x1 \pm 1 \times x2 + 1 \times x3 \quad (4)$$

A visualization of 25 records in terms of actual vs predicted values are shown in the below bar graph in Figs. 5 and 6 which shows the closeness between them.

The forecast values for the next few days are shown in Figs. 7 and 8 using the above prediction model.

4. Discussion

India as well as Odisha as one of its state is now at a critical reaction point as shown in the above Figs. 7 and 8. As in the case of polio, surveillance plays a key role in the battle against COVID-19 too. Accordingly, at the government's decision, WHO has raised support for enhancing effectiveness of the control and response at the federal, district and block levels; cluster confinement operations; reinforcing data gathering actions in real time.

India has already taken effective steps like full initiatives.

Statistical models are important techniques for evaluating infectious disease data analyses in real time. We used the Linear and Multiple Linear regression model in this paper to evaluate the epidemic data of the region of India and India as a country through a detail investigation into the different applications of the models in history. Syazali et al. [7] examined the influence of volume, quality of goods and the brand name on buying value from consumers using Multiple Linear Regression analysis, Multiple Linear regression (MLR) is discussed by Salleh et al. [8] to infer GRN from data relating to gene expression and prevent inferring indirect interaction as a direct interaction and shows the effectiveness of MLR in dealing cascade error. Uyanık and Güler [9] discusses on Multiple Linear Regression and examined the values using the assumptions like normality, linearity, no extreme values. Three data models like Artificial Neural network, adaptive neuro fuzzy inference system and multiple linear regression were used by Khademi et al. [10] to predict the overall strength of recycled aggregator concrete. Hosseinzadeh et al. [11] uses Artificial Neural Network and Multiple Linear Regression to forecast the recovery of nutrients from solid waste under different treatments with vermicompost. Multiple linear regression -TOPSIS is studied by Luu, von Meding, and Mojtahedi [12] for predicting disaster from data. Similarly, the relationship between the mechanical properties of the tea stem and their impact factor has been studied by Du, Hu, and Buttar [13] to improve the picking efficiency of the tea plucking machine using MLR technique. Kadam et al. [14] uses Artificial Neural Network and

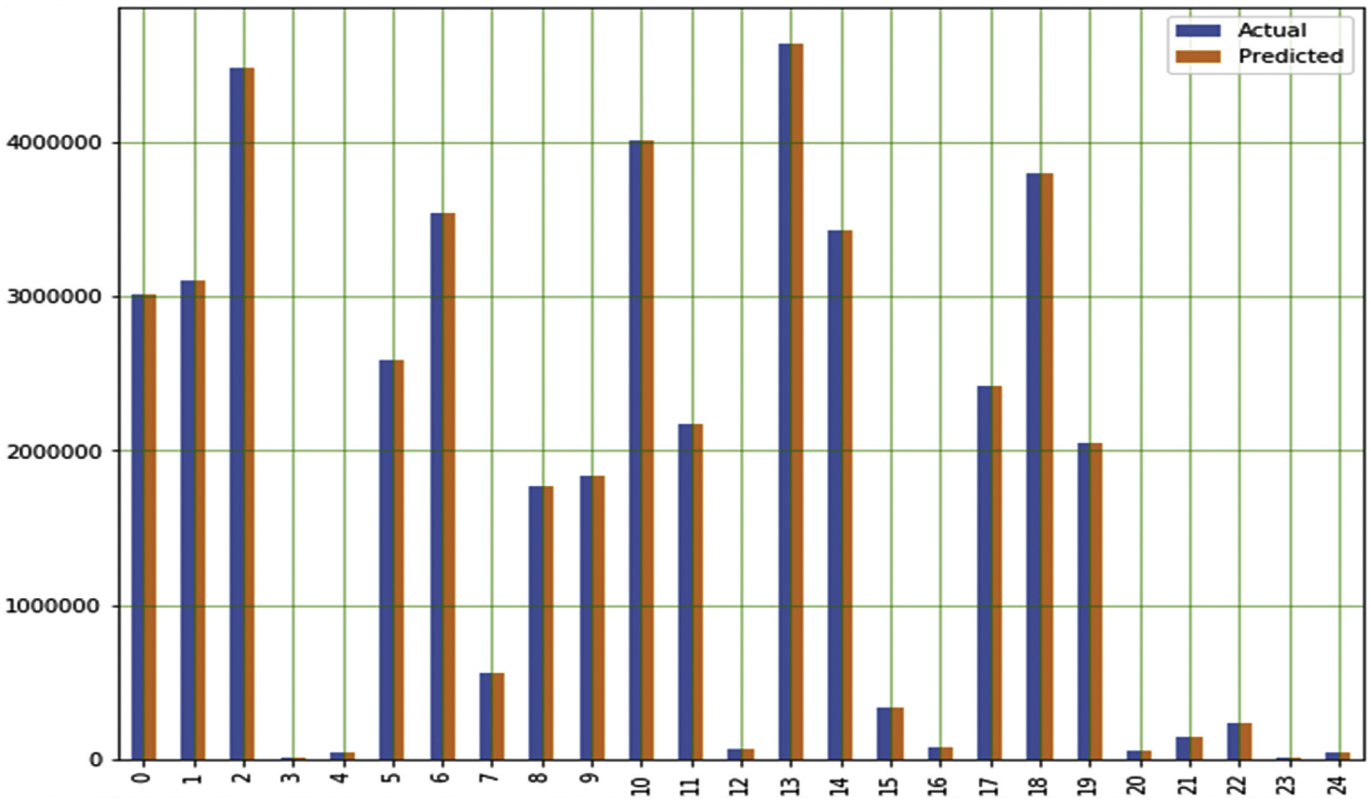

Fig. 6. Visualization of Actual vs predicted values using Multiple Linear Regression Model in India COVID-19 Cases.

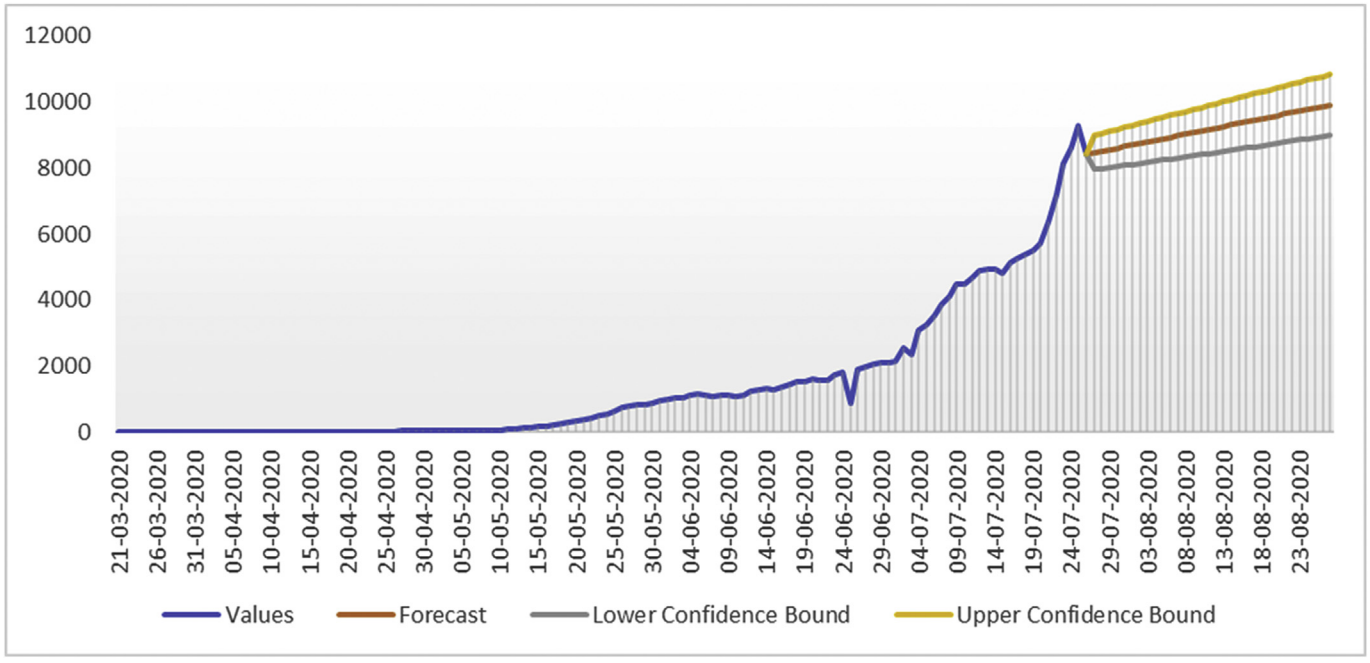

Fig. 7. Forecast of next days of Odisha COVID-19 cases.

Multiple Linear Regression to predicting ground water quality fitness for drinking from Shivganga river basin located in the eastern slope of the western Ghats, India. Quasi-Monte Carlo combined with multiple linear regression (QMC-MLR) is suggested by Xu and Yan [15] to solve the calculation of probabilistic load flow (PLF).

To reduce the number of accidents Jomnonkwao, Uttra, and Ratanavaraha [16] in their paper provide plan which required future forecast data using regression models. Yuchi et al. [17] uses Multiple Linear Regression and random forest model to predict the

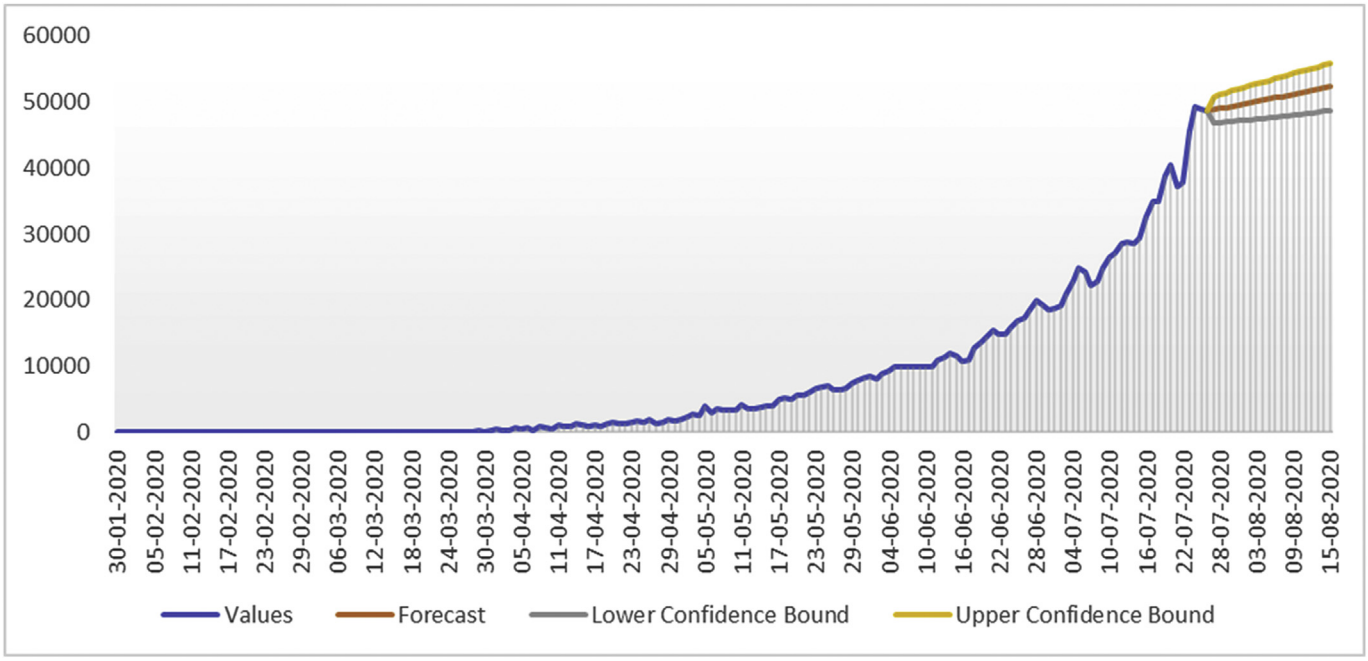

Fig. 8. Forecast of next days of India COVID-19 cases.

Table 5
Statistical ANOVA measure of Multiple Linear Regression model.

	Df	SS	MS	F	Significance F
Regression	3	5.19E+14	1.73E+14	18.64554641	0.00000015
Residual	156	1.45E-15	9.31E-18		
Total	159	5.19E+14			

particulate matter increasing the death and diseases.

This prediction model will speculate the advance situation that is coming in days and effective measures are to be more enhanced to flatten the curve. The forecast value of 9358 active cases with a lower bound of 8582 cases and upper bound of 10,134 in Odisha and 52,290 active cases with a lower bound of 48,711 and upper bound value of 55,868 in India as shown in Figs. 7 and 8 shows the growth in upward direction in COVID-19 cases. The similarity of using the model lies in the behaviour of future forecast using MLR technique and how it influences the different factors of the daily analysis of positive, recovered and deceased cases.

Strength of the Model: The strength of the model is its R^2 value came to be 1.0 which shows a strong predictor model taking into consideration of all the factors as shown below in Table 5 as the Statistical ANOVA measure. Variance Analysis (ANOVA) comprises of simulations which provide knowledge on levels of variation within a regression model and form the basis for meaningful tests. The significance F value is 0.00000015 which derives the P value to check the null hypothesis that all-group data are derived from groups with the same means. P value is greater than 0.05 is a chance that the null hypothesis is true (see Table 6).

Limitations of the Model: Limitations of the model can be thought of in terms of gathering more independent variables or information, ways to find the number of contact tracing cases. If the number of contact tracing cases are been reduce, it will indirectly reduce the number of daily active cases.

5. Conclusion

From the above training and testing of the prediction models, it was found to be an effective way to forecast the next number daily active cases during second week of August as we can see the forecast figure shows the active number of cases will tend to be around upper confidence value to be 10,134 cases and lower confidence value of 8582 cases in Odisha and similarly the upper confidence value of forecast is around 48,711 and lower confidence value as 55,868 in case of India. These models acquired remarkable accuracy in COVID-19 recognition. Bearing in mind these projected active results, the current estimated for COVID-19 containment needs to be reinforced or updated. Our framework could assist and protect healthcare professionals, government officials in making plans appropriate to cope with the influx of future COVID-19 patients.

Table 6
Summary Output: ANOVA showing the Significance of p-value to validate the model for prediction of daily Active cases.

	Coefficients	Standard Error	T Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	2E-09	4.12E-10	4.164241	0.6734	9.02616E-10	2.53E-09	9.02616E-10	2.53164E-09
Positive	1	2.52E-15	3.97E+14	0.6633	1	1	1	1
Recovered	−1	2.56E-15	−3.9E+14	0.6533	−1	−1	−1	−1
Deceased	1	2.41E-14	4.16E+13	0.6753	1	1	1	1

Funding

No Funding

Declaration of competing interest

None to declare.

References

[1] Wang W, Tang J, Wei F. Updated understanding of the outbreak of 2019 novel coronavirus (2019-nCoV) in Wuhan, China. J Med Virol 2020;92(4):441–7.
[2] Coronavirus disease 2019 (COVID-19): situation report. World Health Organization; 2020. p. 70.
[3] Modes of transmission of virus causing COVID-19: implications for IPC precaution recommendations: scientific brief. World Health Organization. [Online] Available at : 27 March 2020 (No. WHO/2019-nCoV/Sci_Brief/Transmission_modes/2020.1).
[4] Centers for Disease Control and Prevention. Symptoms of coronavirus [online] Available at: https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/symptoms; 2020. Accessed on: April 21, 2020.
[5] India COVID-19 TRACKER. 2020 [online]. Available at, https://www.covid19india.org/. Accessed on: 11th July 2020.
[6] Barkur G, Vibha GBK. Sentiment analysis of nationwide lockdown due to COVID 19 outbreak: evidence from India. Asian J Psychiatr 2020. https://doi.org/10.1016/j.ajp.2020.102089. Jun; 51: 102089. Published online 2020 Apr 12.
[7] Syazali M, Putra F, Rinaldi A, Utami L, Widayanti W, Umam R, Jermsittiparsert K. Partial correlation analysis using multiple linear regression: impact on business environment of digital marketing interest in the era of industrial revolution 4.0. Manag Sci Lett 2019;9(11):1875–86. 2019.
[8] Salleh FHM, Zainudin S, Arif SM. Multiple linear regression for reconstruction of gene regulatory networks in solving cascade error problems. Adv Bioinformat 2017:1–15. https://doi.org/10.1155/2017/4827171. 4827171.
[9] Uyanık GK, Güler N. A study on multiple linear regression analysis. Procedia Soc Behav Sci 2013;106:234–40.
[10] Khademi F, Jamal SM, Deshpande N, Londhe S. Predicting strength of recycled aggregate concrete using artificial neural network, adaptive neuro-fuzzy inference system and multiple linear regression. Int J Sustain Built Environ 2016;5:355–69.
[11] Hosseinzadeh A, Baziar M, Alidadi H, Zhou JL, Altaee A, Najafpoor AA, Jafarpour S. Application of artificial neural network and multiple linear regression in modeling nutrient recovery in vermicompost under different conditions. Bioresour Technol 2020;303:122926.
[12] Luu C, von Meding J, Mojtahedi M. Analyzing Vietnam's national disaster loss database for flood risk assessment using multiple linear regression-TOPSIS. Int J Disast Risk Re 2019;40:101153.
[13] Du Z, Hu Y, Buttar NA. Analysis of mechanical properties for tea stem using grey relational analysis coupled with multiple linear regression. Sci Hortic 2020;260:108886.
[14] Kadam AK, Wagh VM, Muley AA, Umrikar BN, Sankhua RN. Prediction of water quality index using artificial neural network and multiple linear regression modelling approach in Shivganga River basin, India. Model Earth Syst Environ 2019;5:951–62.
[15] Xu X, Yan Z. Probabilistic load flow calculation with quasi-Monte Carlo and multiple linear regression. Int J Electr Power Energy Syst 2017;88:1–12.
[16] Jomnonkwao S, Uttra S, Ratanavaraha V. Forecasting road traffic deaths in Thailand: applications of time-series, curve estimation, multiple linear regression, and path analysis models. Sustainability 2020;12:395.
[17] Yuchi W, Gombojav E, Boldbaatar B, Galsuren J, Enkhmaa S, Beejin B, Barn P. Evaluation of random forest regression and multiple linear regression for predicting indoor fine particulate matter concentrations in a highly polluted city. Environ Pollut 2019;245:746–53.