ORIGINAL RESEARCH ARTICLE

A Study to Analyse Covid-19 Outbreak Using Multiple Linear Regression: A Supervised Machine Learning Approach

Jayanti Semwal[1], Abhinav Bahuguna[2*], Akanksha Uniyal[3], Shaili Vyas[4]

[1,2,3,4]Himalayan Institute of Medical Sciences, Swami Rama Himalayan University, Dehradun, India

DOI: 10.55489/njcm.140220232656

ABSTRACT

Introduction: Globally, COVID-19 have impacted people's quality of life. Machine learning have recently become popular for making predictions because of their precision and adaptability in identifying diseases. This study aims to identify significant predictors for daily active cases and to visualise trends in daily active, positive cases, and immunisations.

Material and methods: This paper utilized secondary data from Covid-19 health bulletin of Uttarakhand and multiple linear regression as a part of supervised machine learning is performed to analyse dataset.

Results: Multiple Linear Regression model is more accurate in terms of greater score of R^2 (=0.90) as compared to Linear Regression model with R^2=0.88. The daily number of positive, cured, deceased cases are significant predictors for daily active cases ($p < 0.001$). Using time series linear regression approach, cumulative number of active cases is forecasted to be 6695 (95% CI: 6259 - 7131) on 93rd day since 18 Sep 2022, if similar trend continues in upcoming 3 weeks in Uttarakhand.

Conclusion: Regression models are useful for forecasting COVID-19 instances, which will help governments and health organisations to address this pandemic in future and establish appropriate policies and recommendations for regular prevention.

Key words: Covid-19, Supervised machine learning, Multiple Linear Regression, Forecast, Uttarakhand

ARTICLE INFO

Financial Support: None declared

Conflict of Interest: None declared

Received: 16-12-2022, **Accepted**: 02-02-2023, **Published**: 28-02-2023

***Correspondence:** Abhinav Bahuguna (Email: bahuguna.abhinav.17@gmail.com)

How to cite this article:

Semwal J, Bahuguna A, Uniyal A, Vyas S. A Study to Analyse Covid-19 Outbreak Using Multiple Linear Regression: A Supervised Machine Learning Approach. Natl J Community Med 2023;14(2):82-89. DOI: 10.55489/njcm.140220232656

www.njcmindia.com│pISSN09763325│eISSN22296816│Published by Medsci Publications

Introduction

As of April 15, 2020, the World Health Organization (WHO) reported 1,878,489 confirmed cases (119,044 confirmed fatalities) of the novel coronavirus disease 2019, including 11,439 confirmed cases (377 deaths) from India.[1] According to a study from the Indian government[2], the worst-affected states and union territories include Delhi with 1, 07,051 cases, followed by Tamil Nadu with 1, 26,581 cases, and Maharashtra with 2,30,599 cases. The pandemic gradually expanded to other states and union territories, including Uttarakhand which is nestled in the lap of the Himalayas and crossed by the sacred River Ganges.[3] According to the 2011 Indian Population Census, the state currently has a population of 11.9 million, making it home to more people, than countries like Hong Kong, Switzerland, Israel, etc. The rough terrains of Uttarakhand, with impeding healthcare infrastructure, have further aggravated the pandemic. Overall, Uttarakhand reports 334,024 confirmed cases and 66,699 deaths during the writing of the document. COVID-19 nowadays represents a true global health crisis for humanity and an important new challenge that must be met. A method for predicting COVID-19, using pictures and a machine learning (ML) algorithm has been provided in various studies that have shown an accuracy rate of 86%.[4]

Machine learning is categorized under unsupervised learning and supervised learning. Without being programmed, computers can learn with the aid of machine learning. In supervised learning, we map fresh instances by examining the training data's input-output relationship. Covid-19 outbreaks are being predicted with the help of models and a seasonal periodic regression model.[5-7]

Suspected-Infected-Recovered-Dead (SIRD) Model[8] makes estimates for epidemiological variables like infection, reproduction, and mortality rates. Recovered, death cases, negative cases, and confirmed cases have all been predicted utilizing long short-term memory and Gated Recurrent Unit employing Recurrent Neural Network.[9]

The number of confirmed cases, fatalities and recovered cases are used to develop linear and polynomial regression models. The estimated cases and fatalities over the next few days can be forecasted with the use of these models. However, limited studies exist on the analysis of COVID-19 outbreak using the Supervised Machine Learning Approach in Uttarakhand. Hence, the present study is an attempt to correlate the underlying factors and improve preparedness if such a pandemic devastates the state in the future.

Methodology

The applied methodology in this paper is described with the help of a flow chart. [Fig 1]

Dataset Description: The dataset for this research study was retrieved from the daily COVID-19 health bulletin, which can be publicly assessed from the official website of Medical Health and Family Welfare (MoHFW), is managed by the Uttarakhand Government.[10] The daily records on positive cases, active cases, deceased cases, cured cases, sample testing and vaccinations were obtained from 01 March 2022 to 30 November 2022 for the state of Uttarakhand.

Statistical Analysis: The whole dataset was cleaned and maintained by removing missing and duplicate values and analysis was performed using R software version 4.2.2 with the help of required packages and libraries. A Supervised Machine Learning approach was used by apportioning the whole dataset where 75% of it belongs to training set and remaining 25% to testing set. The procedure is described into following three steps.

Step 1: A Linear Regression (LR) analysis is performed on training set to evaluate the influence of active cases due to daily positive cases in Uttarakhand whose purpose is to predict the response variable (Y=daily active cases) for a given value of explanatory variable (X=daily positive cases) with the help of the following equation:

$$Y = \beta_0 + \beta X + €$$

Where β_0 is the intercept, β is the slope and € is the error term in the above LR model.

Step 2: Multiple linear regression (MLR) is used to evaluate the influence of response variable (Y) by taking a set of explanatory variables (X_1=positive cases, X_2=deceased, X_3=cured cases, X_4= daily vaccinations) with the help of the following equation:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + €$$

Where β_1, β_2, β_3, β_4 are slope coefficients and € is the error term in the above MLR model.

Step 3: Further, with the help of dataset from 18 September 2022 to 30 November 2022, cumulative number of active cases are forecasted for the next 3 weeks using time series based linear regression model by taking trend as an independent variable.

Results

On average, 391 active cases were observed in the month of March 2022 and this average decreased to 101 in next month. Again, a sudden increase in active cases were observed at the end of July and beginning of August and thereafter cases were noticed to be declined. [Fig 2] [Fig 3] [Fig 4] shows the trend of daily active cases, positive cases and vaccinations since the first day of March 2022.

A scatterplot matrix shows the correlation values amongst the variables as depicted in [Fig 5]. The correlation between response variable (daily active cases) and explanatory variable (daily positive cases) is

0.912 (p-value <0.001), which shows a strong positive linear relationship between them and hence motivates us to further establish a mathematical relationship between them.

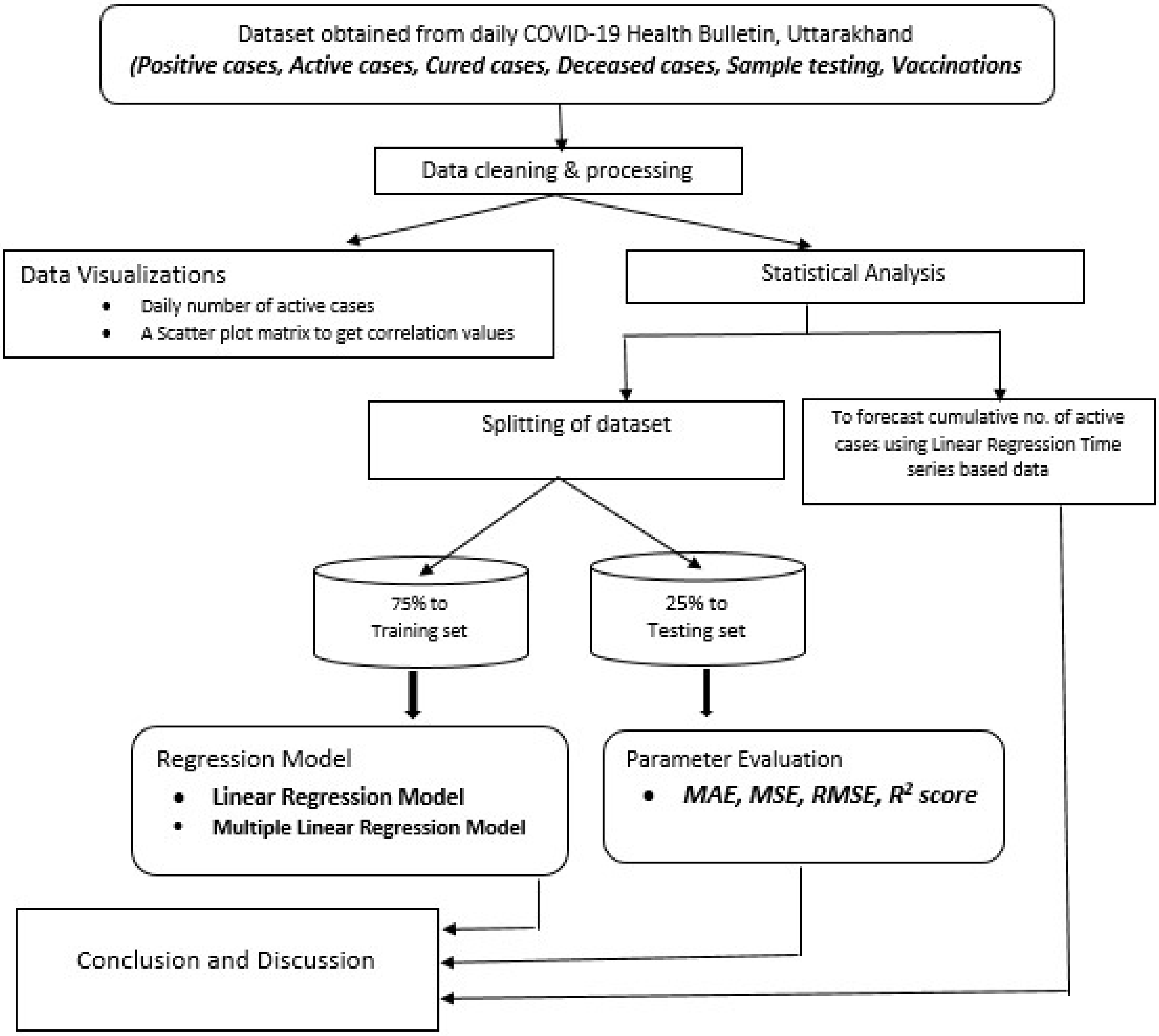

Dataset obtained from daily COVID-19 Health Bulletin, Uttarakhand
(Positive cases, Active cases, Cured cases, Deceased cases, Sample testing, Vaccinations
Data cleaning & processing
Data Visualizations
Daily number of active cases
A Scatter plot matrix to get correlation values
Statistical Analysis
Splitting of dataset
To forecast cumulative no. of active cases using Linear Regression Time series based data
75% to Training set
25% to Testing set
Regression Model
Linear Regression Model
Multiple Linear Regression Model
Parameter Evaluation
MAE, MSE, RMSE, R² score
Conclusion and Discussion

Figure 1: Flowchart for the proposed methodology

3.1: The fitted linear regression model after applying it to 75% of training set is as follows:

$Y_{\text{(daily no. of active cases)}} = 82.41 + 5.37\, X_{\text{(daily no. of positive cases)}}$

------- Equation (1)

Equation (1) indicates that for every one unit increase in daily number of positive cases, there is an increase of 5 cases in daily active cases in the state of Uttarakhand [Table 3.1] and F-statistic for the fitted model is 923.4 (p-value <0.001) showing the evidence of linear relationship between the variables.

3.2: The fitted multiple linear regression model is explained with the help of Equation (2):

The daily sample testing is weakly correlated (r= 0.263) with the daily active cases [Fig 5], so the MLR model is defined with the help of only four explanatory variables (X_1=daily no. of positive cases, X_2=cured cases, X_3=deceased cases, and X_4=vaccinations).

$Y_{\text{(daily active cases)}} = 74.59 + 4.16\, X_1 + 1.46\, X_2 + 66.63\, X_3 - 0.0013\, X_4$ -------- Equation (2)

Equation (2) indicating the increase of approximate 4 cases to daily number of active cases for each one unit increase in daily number of positive cases in Uttarakhand while keeping the other explanatory variables fixed and F-statistic for the fitted model comes out to be 299.8 (p <0.001) showing that at least one explanatory variable in fitted MLR model has significant linear relationship with the response variable. In the present study, daily positive cases, cured cases, and deceased cases are significant predictors for response variable in the model [Table 3.2].

Prediction is done with testing set and the evaluation parameters have been evaluated. Mean Absolute Error (MAE) in the MLR model comes out to be 105.54 with R^2 =0.90 which explains it as a strong predictor model as compared to the LR model having MAE=114.90 with R^2 =0.88 as represented in [Table 3.3].

The actual vs. predicted values for the LR and the MLR model in the form of scatter diagram are shown in [Fig 6] and [Fig 7] respectively which shows that differences of most of the values are close to fitted regression line.

Further, we used a time series based linear model by taking a cumulative number of active cases as a response and trend as explanatory variables, whose purpose is to fit a straight line following a trend pattern only [Fig 8]. The forecasted value comes out to be 6695 (95% CI: 6259 - 7131) on 93rd day since 18 Sep 2022 [Table 4] and this visualization of cumulative number of active cases for the next 3 weeks is depicted in [Fig 9].

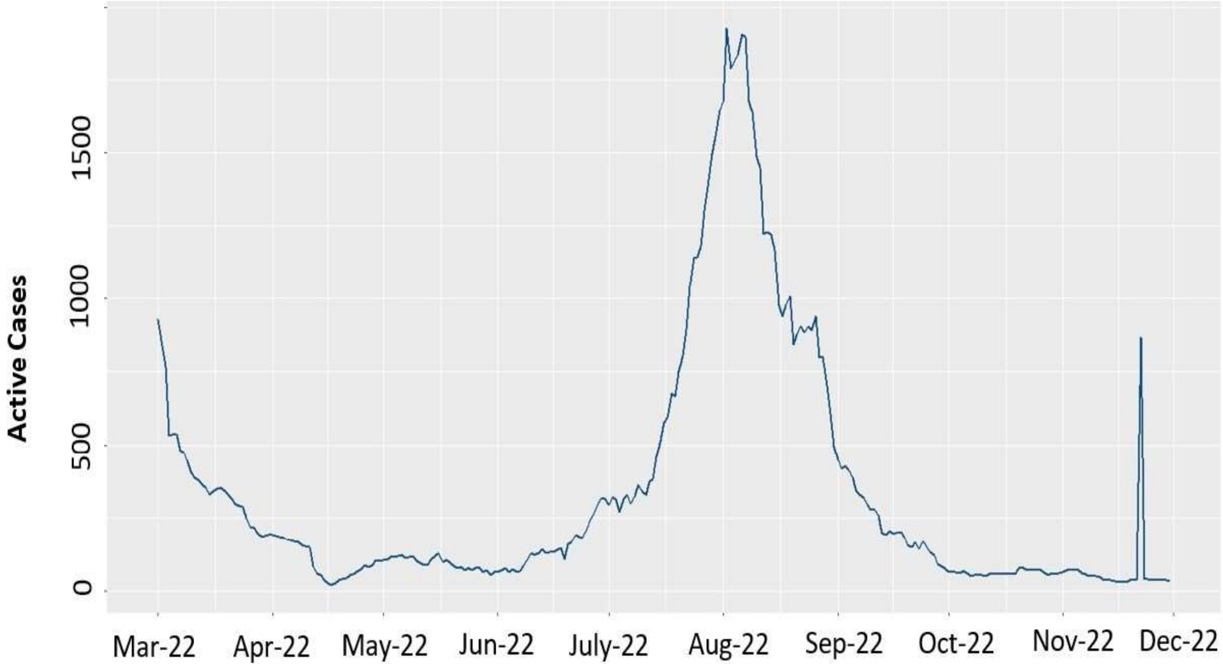


Figure 2: Daily active cases since March 2022

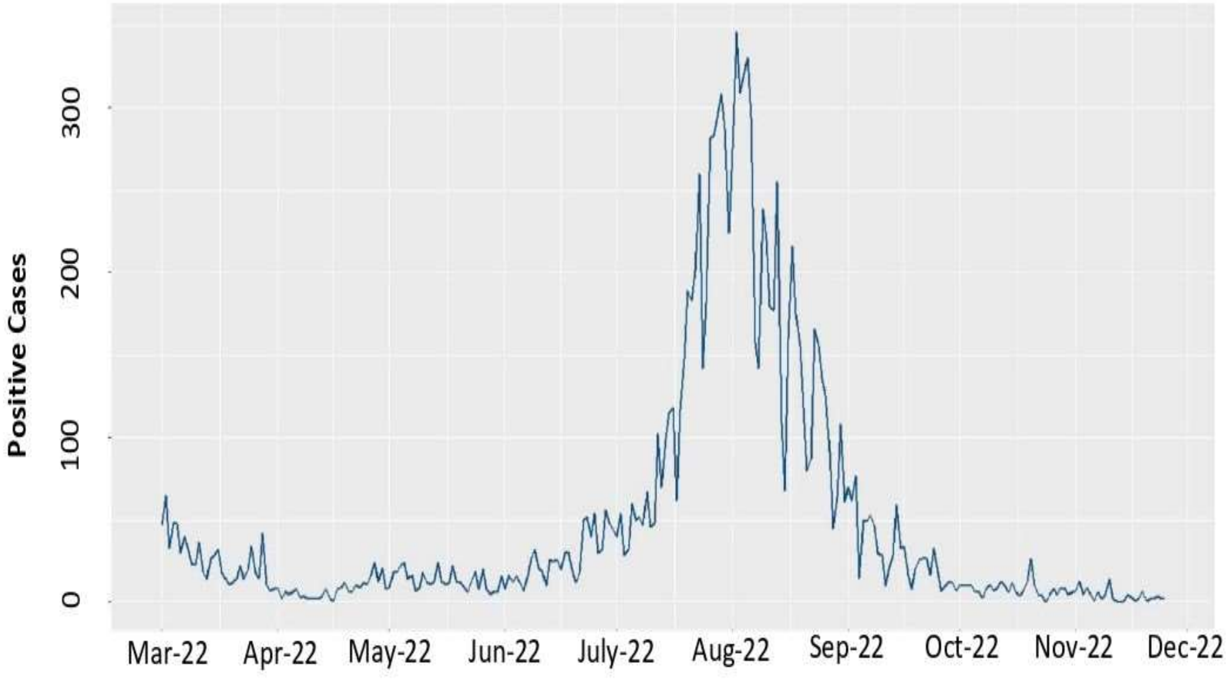


Figure 3: Daily positive cases since March 2022

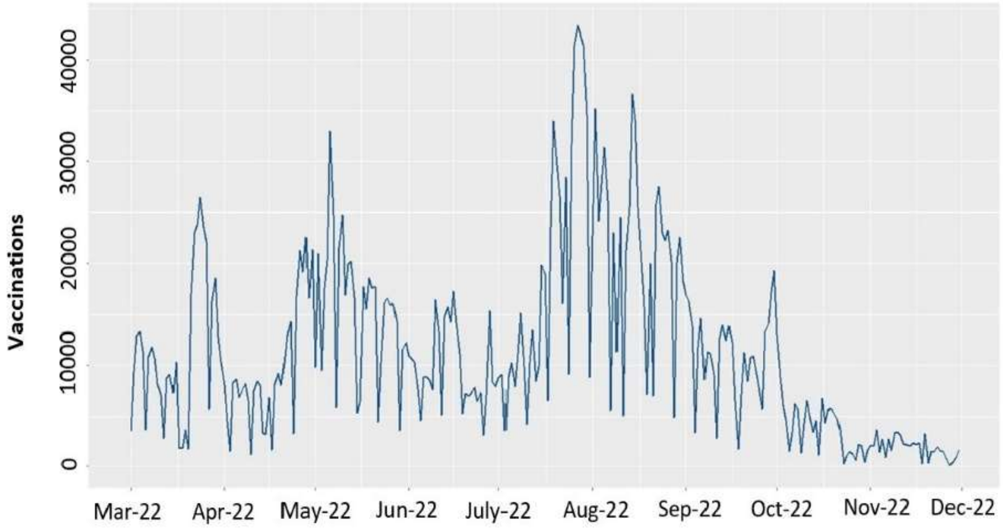


Figure 4: Daily vaccinations since March 2022

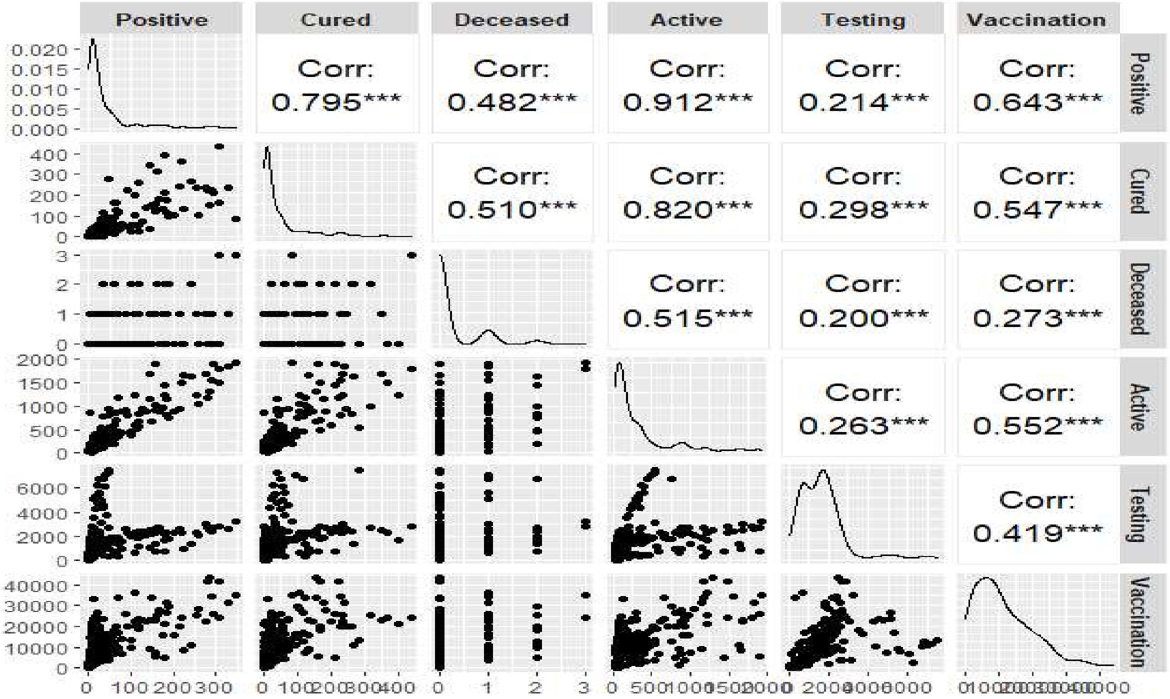
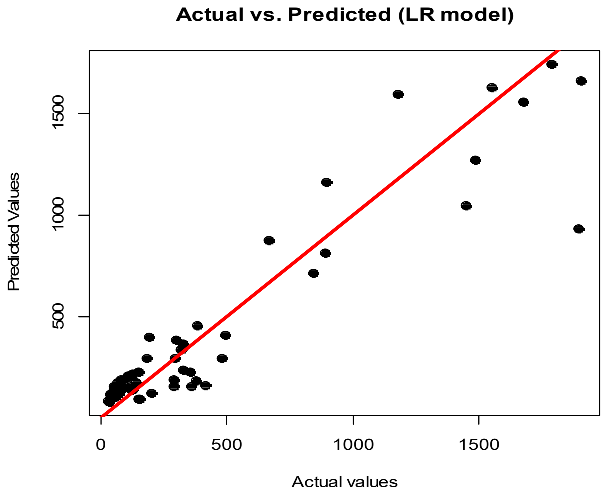


Figure 5: The Correlation matrix to show correlation values



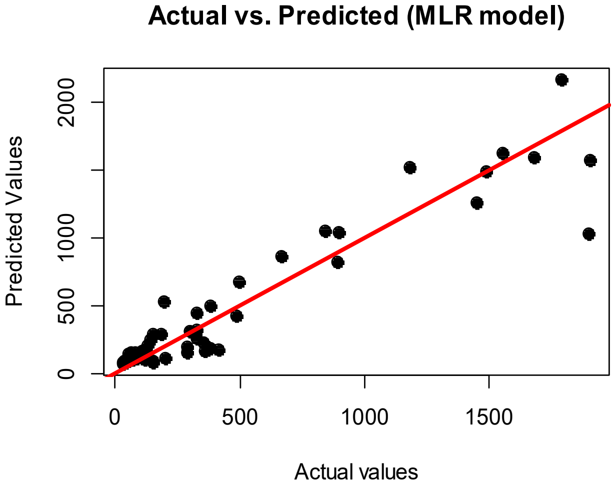


Figure 6: Actual vs. Predicted values for Linear Regression Model

Figure 7: Actual vs. Predicted values for Multiple Linear Regression Model

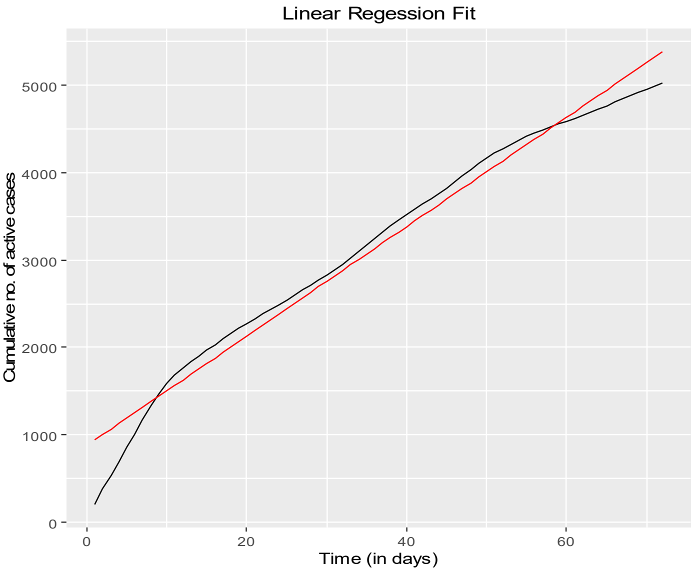


Figure 8: Best Linear fit on cumulative number of active cases

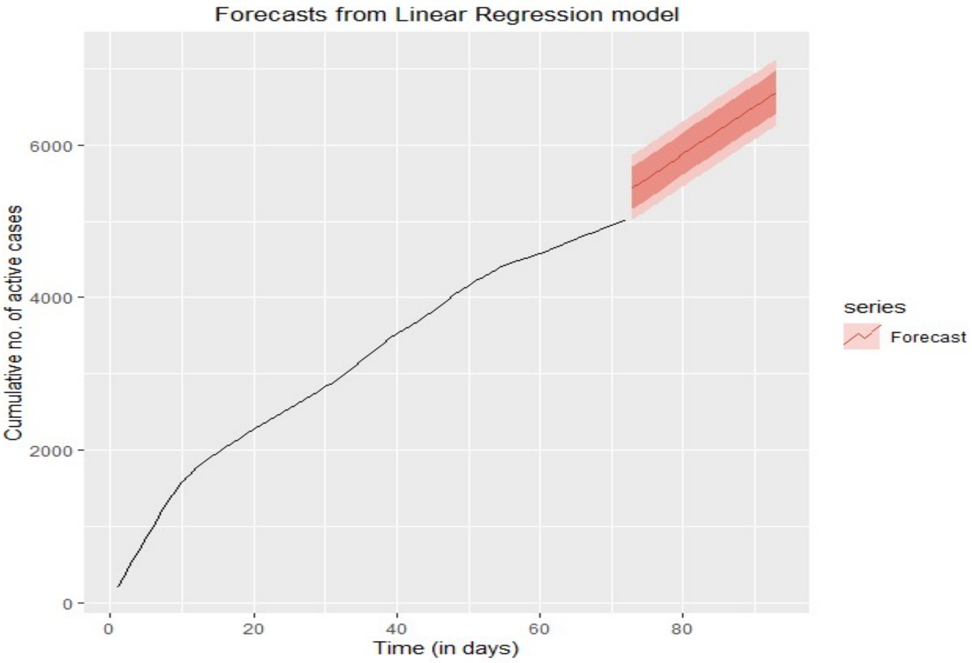


Figure 9: Forecast for cumulative number of active cases for next 3 weeks

Table 3.1: Linear regression coefficients from training set, active cases as a response variable

	Estimate	Standard Error	T value	P-value	95% Confidence Interval
Intercept	82.41	13.85	5.94	0.000***	(55.08, 109.73)
Positive	5.37	0.17	30.38	0.000***	(5.02, 5.72)

***p <0.001 and R^2= 0.822, F-statistic: 923.4, p-value: <0.001

Table 3.2: Multiple Linear regression coefficients from training set, active cases as a response variable

	Estimate	Standard Error	T value	p-value	95% Confidence Interval
Intercept	74.59	17.50	4.26	0.000***	(40.07, 1.091076e+02)
Positive cases	4.16	0.27	15.36	0.000***	(3.62, 4.695661e+00)
Cured cases	1.46	0.23	6.26	0.000***	(1.00, 1.931812e+00)
Deceased cases	66.63	24.78	2.68	0.007*	(17.75, 1.155123e+02)
Vaccinations	-0.0013	0.0016	-0.85	0.244	(-0.004, 1.825015e-03)

***p <0.001, * p <0.01 and Adjusted R^2= 0.856, F-statistic: 299.8, p-value < 0.001

Table 3.3: Evaluation parameters of the prediction model

Model	R^2 score	Mean absolute error (MAE)	Mean Squared Error (MSE)	Root Mean Squared Error (RMSE)
LR model	0.88	114.90	30,046.75	173.34
MLR model	0.90	105.54	26,536.41	162.90

Forecasts using Time Series based Linear Regression Model:

Table 4: Forecasting cumulative no. of active cases for the next 3 weeks.

Days (Since 18-Sep-2022)	Cumulative number of forecasts	95% Confidence Interval
73* (01-Dec-2022)	5444	(5020, 5868)
-----------	------------	---------------
-----------	------------	----------------
92* (20-Dec-2022)	6633	(6197, 7068)
93* (21-Dec-2022)	6695	(6259, 7131)

Discussion

India has been already taken the initiatives for the management of COVID-19 pandemic. All the governments around the world applied severe actions for managing the rapid spread of COVID-19 infection among people.[11] Every area of life is being rapidly impacted by technological advancements; the medical field is one of the important departments which is directly related to people's daily lives. Due to the greater accuracy of data processing and precise decision-making, artificial intelligence (AI) has recently been applied to the medical profession and has demonstrated promising results in healthcare. For the present study, we applied Linear and multiple linear regression model as a part of Supervised Machine Learning (ML) approach in order to predict the daily COVID-19 active cases and also aimed to forecast the cumulative number of active cases for the next three weeks in Uttarakhand.

The dataset for this research study was extracted from the daily COVID-19 Health Bulletin which reported cumulative number of a total of 104497 positive cases, 100328 recovered cases and 333 deceased cases as of 25th November, 2022 since the beginning of the year 2022 and moreover, there were a total of 36 active cases in the state of Uttarakhand as per the Health Bulletin reported on 25th November, 2022.[10] Suganya et al.[12] discussed that multiple linear regression is a suitable technique to forecast the cumulative number of confirmed and deceased cases (R^2=0.992).

A recent study published by Kanday et al.[13] extracted their dataset from GitHub from 212 reports, reporting a total of 1000 cases. They used logistic regression and multinomial Naive Bayes algorithm in their study. However, the findings showed that these algorithms (with an accuracy of 96%) are better than the commonly used algorithms. Another study conducted by Varun et al.[14] in the year 2020, consisted of 184,319 reported cases. They used supervised learning as their method with convolutional neural network CNN. The study focused on developing a ML algorithm for frontline physicians in the emergency department.

Another study conducted by Burdick et al.[15] in the year 2020, conducted on 197 patients by using supervised learning as their main method. They used classification logistic regression and reported that this algorithm displays higher diagnostic odds ratio (12.58) for foreseeing ventilation and effectively triage patients than a comparator early warning system, such as Modified Early Warning Score (MEWS) which showed (0.78) sensitivity, while this algorithm showed (0.90) sensitivity which leads to higher specificity. Based on hospitalization data, a total of 5435 persons were supposed to be hospitalized in month of September 2022 using single exponential smoothing with time series data in a study conducted by Semwal et. al (2022)[16] for the state of Uttarakhand. Painuli et. al (2021)[17] in their study had applied ARIMA modelling on dataset and predicted the confirmed cases for the next 45 days in India as well as its state's Telangana and Maharashtra. The findings of the present study explain weak correlation [Fig 5] of sample testing with daily active cases (r=0.263) and daily positive cases (r=0.214), and hence daily sample testing was not included in MLR model as a predictor, whereas Saswat. S et. al (2021)[18] showed that increase in daily number of tests will result to consecutive increase in daily cases and had reported that a total of 50234 predictive positive cases on 31 July 2020.

Moreover, findings from the present paper reported that the daily positive, cured cases, and deceased cases were statistically significant predictors for the response variable (daily active cases) at a particular level of significance whereas as similar findings of S Rath et al.[19] reported no significant predictors were observed while analyzing data for the state of Odisha and India as well. In addition, we utilized linear regression-based time series model to forecast the cumulative number of active cases as response variable and trend as explanatory variable, whose purpose is to fit a straight line [Fig 8] following a trend pattern only. According to present study, cumulative no. of active cases for the next 3 weeks is visualized in [figure 9] and forecasted to be 6695 (95% CI: 6259 - 7131) on 93rd day since 18 September 2022, considering the pattern being same in upcoming days in the state of Uttarakhand [Table 4]. Researchers can use artificial neural network (ANN) and Autoregressive Integrated Moving Average (ARIMA) for forecasting purposes by considering historical behavior of Covid-19 virus.

Limitation

A more set of explanatory variables such as migration of infectious cases, travelling history, and demographic features of an individual can be added to MLR model as it might have a positive impact in reducing the number of daily active cases. Also, time series based linear regression approach used in forecasting cumulative number of active cases is based on trend only, while seasonality can be considered.

Conclusion

Researchers and health organisations across the world have made great effort in prevention of COVID-19 pandemic and controlling its spread among community by developing various statistical models and increasing rate of testing and vaccinations. In present study, we proposed Linear Regression (LR) and Multiple Linear Regression (MLR) models by employing daily number of active cases as response variable and other as explanatory variable/s, and with the help of R2 we found the MLR model to be more accurate as compared to the LR model. The correlation values show strong positive relationships of positive cases and cured cases with daily active cases whereas vaccinations and deceased cases are moderate positively correlated with daily active cases. The forecast value through Linear regression based time series model is an adequate way to predict the cumulative number of daily active cases for next 21 days showing that cumulative number of active cases on 93rd day will be 6695 with a lower bound of 6259 cases and upper bound of 7131 cases. The findings through these models will be helpful for government, local administration and hospital staff to make strong policies against Covid-19 and take necessary steps in its prevention for upcoming days.

Acknowledgement

The authors acknowledge Swami Rama Himalayan University for their support and Dr. Kapil Lakhwara, Dr. Mansi Pandey and Dr. Swapnil Tiwari, postgraduate students of the Department of Physiotherapy, Himalayan Institute of Medical Sciences, Dehradun, for their valuable help during the pre-processing phase of the data. Thanks for the help.

References

1. Coronavirus disease (COVID-19) outbreak situation. Available at https://www.who.int/emergencies/diseases/novelcorona virus-2019 (2020).
2. India COVID-19 TRACKER. 2020 [online]. Available at, https://www. covid19india.org/. Accessed on: 11th July 2020
3. Negi SS. Uttarakhand: Land and people. Delhi, India: MD Publications Pvt. Ltd; 1995.
4. Band SS et al., A Survey on Machine Learning and Internet of Medical Things-Based Approaches for Handling COVID-19: Meta-Analysis. Frontiers in Public Health. 2022 Vol-10;2296-65.
5. Ghosal S, Sengupta S, Majumder M, Sinha B. Linear Regression Analysis to predict the number of deaths in India due to SARS-CoV-2 at 6 weeks from day 0 (100 cases - March 14th 2020). Diabetes Metab Syndr. 2020;14(4):311-315. doi:10.1016/j.dsx.2020.03.017
6. New York City Department of Health and Mental Hygiene (DOHMH) COVID-19 Response Team. Preliminary Estimate of Excess Mortality During the COVID-19 Outbreak - New York City, March 11-May 2, 2020. MMWR Morb Mortal Wkly Rep. 2020;69(19):603- 605. Published 2020 May 15. doi:10.15585/mmwr.mm6919e5
7. Pandey G , Chaudhary P , Gupta R , Pal S, SEIR and Regression Model based COVID-19 outbreak predictions in India, medRxiv 2020.04.01.20049825;
8. Anastassopoulou C, Russo L, Tsakris A, Siettos C Data-based analysis, modelling and forecasting of the novel coronavirus (2019-Ncov) outbreak. medRxiv preprint 10.1101/2020.02.11.20022186
9. Dutta S, Bandyopadhyay SK. Machine Learning Approach for Confirmation of COVID19 Cases: Positive, Negative, Death and Release, medRxiv 2020: https://doi.org/10.1101/2020.03.25.20043505
10. Data Sources: Daily Covid-19 Health Bulletin (March 2022 - November 2022): Ministry of Health & Family Welfare (MoHFW), Government of Uttarakhand; URL: https://health.uk.gov.in
11. Alimadadi A , Aryal S, Manandhar , Munroe PB, Joe B , Cheng X . Artificial intelligence and machine learning to fight COVID-19. Physiol Genomics. 2020 ; 52(4): 200–202.
12. Suganya R, Arunadevi R, Buhari SM. COVID-19 forecasting using multivariate linear regression.Research Square.2020:1-17.
13. Khanday AMUD et al ,Rabani ST, Khan QR, Rouf N, Mohiuddin .Machine learning based approaches for detecting COVID-19 using clinical text data. Int J Inf Technol.2020; 12(3):731–739
14. Arvind V , Kim JS, Cho BH, GengE, Cho SK. Development of a machine learning algorithm to predict intubation among hospitalized patients with COVID-19. J Crit Care 2020;62:25–30
15. Burdick H et al.Prediction of respiratory decompensation in Covid-19 patients using machine learning: the READY trial. Comput Biol Med.2020; 124:103949
16. Semwal, J., Bahuguna, A., Sharma, N., Dikshit, R. K., Bijalwan, R., & Augustine, P. Time Series Analysis of COVID-19 Data-A study from Northern India. Indian Journal of Community Health. 2020;34(2):202-206.
17. Painuli, D., Mishra, D., Bhardwaj, S., & Aggarwal, M. Forecast and prediction of COVID-19 using machine learning. Data Science for COVID-19 . 2021:381-397).
18. Singh, S., Chowdhury, C., Panja, A. K., & Neogy, S. Time series analysis of COVID-19 data to study the effect of lockdown and unlock in India. Journal of The Institution of Engineers (India): Series B. 2021;102(6):1275-1281.
19. Rath et al. Diabetes & Metabolic Syndrome: Clinical Research & Reviews 14 (2020) 1467e1474