RESEARCH ARTICLE

SAMRIDDHI Volume 15, Issue 3, 2023 Print ISSN: 2229-7111 Online ISSN: 2454-5767

Analysis and Prediction of COVID-19 Cases in Pune Using Machine Learning Techniques

Minakshi Bastapure*, Chaitrali Nilvarn, Annapurna Gosavi, Pranjali More, Sandeep Kadam

Anantrao Pawar College of Engineering and Research, Pune, Maharastra, India

Abstract

COVID has caused a major outbreak in the world. It has brought about the breakdown of the economy, significant overburden in the well-being area, a break in the schooling system and lost lives, and so forth. To ensure that the condition of the world comes to a superior spot regarding work, money and well-being, controlling the outbreak has turned into a main concern. In this paper, ARIMA time series forecasting model is utilized to predict and forecast the assessment of the spread of COVID-19 contamination in the following week. We have taken the data of COVID-19 patients of the Pune district in India. The data is visualized in a simple-to-consume and interactive format. By and large, this paper can help experts and authorities control the COVID-19 outbreak by helping them gain alertness.

Keywords: Coronavirus, machine learning, pandemic, ARIMA, prediction, live analysis, time series forecasting, COVID-19.

SAMRIDDHI : A Journal of Physical Sciences, Engineering and Technology (2023); DOI: 10.18090/samriddhi.v15i03.13

Introduction

The COVID-19 pandemic has had a humongous impact on human livelihood. Coronavirus originated in China's Wuhan province. It immediately spread across the globe. The World Health Organization (WHO) announced it as a worldwide pandemic in the wake of considering its spread pace and the infection's nature and conduct.[6]

The data on COVID-19 is for the most part accessible at the national or state level. Locale-level government bodies like Municipal corporations in India are much of the time reliant upon this data to get to the prediction tools to predict the ascent or fall of new COVID-19 cases. Taking into account this situation, making such a prediction tool and making it accessible at the district level government body in itself will give a more prominent effect by adding to the ordered progression of State, Country and World.

This paper utilizes ARIMA time series forecasting model to perform predictions based on the 3 days rolling data given as reference. ARIMA stands for autoregressive integrated moving average. It is a statistical analysis technique which utilizes time series data to better understand a collection of data or forecast future patterns. The dataset is given by www.cessi.com and it contains various daily confirmed, recovered and deceased Coronavirus cases in Pune district.[7] The data is taken from January 1 2021 to December 31 2021. Data visualization is likewise finished by bringing in various libraries, such as matplotlib and seaborn, plotly to analyze the pattern of COVID-19 patients.

To help boost the vaccination drive across pune district, we have added the system of the vaccination center close to the user utilizing python. It will take users straightforwardly to the Co-WIN gateway to book their vaccine.

Corresponding Author: Minakshi Bastapure, Anantrao Pawar College of Engineering and Research, Pune, Maharastra, India, e-mail: studentminakshee@gmail.com

How to cite this article: Bastapure, M., Nilvarn, C., Gosavi, A., More, P., Kadam, S. (2023). Analysis and Prediction of COVID-19 Cases in Pune Using Machine Learning Techniques. *SAMRIDDHI : A Journal of Physical Sciences, Engineering and Technology,* 15(3), 360-363.

Source of support: Nil

Conflict of interest: None

This paper will give a decent representation over the COVID-19 outbreak in Pune district. Near right prediction will assist the managerial office, clinical office and residents with being ready for the approaching 7 days.

Related Work

Several researchers have added to the areas of predicting and forecasting the COVID-19 pandemic. In,[1] the authors used linear and polynomial ML models to predict and forecast the COVID-19 pandemic in India. They assessed the models utilizing R-squared score and error values techniques. They split data from March 12 to October 31, 2020 into 75% for training and 25% for testing. The paper is predicting the number of confirmed, recovered, and death cases of COVID-19. The authors implemented the tableau time series forecasting approach for forecasting the future trend of these cases.

In,[2] the authors took data from March 4 to May 15. They used regression analysis (exponential and polynomial), auto-regressive integrated moving averages (ARIMA) model,

exponential smoothing and Holt-Winters models to examine the development of COVID-19.

In,[3] authors applied the as-of-late evolved eigenvalue decomposition of the Hankel matrix (EVDHM) alongside the ARIMA model to foster a forecasting model for nonstationary time series.

In,[4] a data dashboard is formed with the assistance of data taken from dependable sources to depict it in an interactive and simple-to-consume design with highlights like a chatbot, cases forecast with AI and projection assistance, and data in various formats updated daily.

In,[5] the paper compares different machine learning algorithms to predict the number of positive cases in India. The effect of lockdown is thought about while developing the ML algorithm. The paper additionally advances key measures and ideas about systems and strategies to the policymakers thinking about the impact of the lockdown. The models are based on China's data and validated to India's sample. The created ML model works in real-time and gives near right predictions of positive cases.

Methodology

Data collection

The data is collected from [6] in csv format. The columns of this dataset contain confirmed cases, recovery cases and death cases of COVID-19 patients on an everyday premise from January 1 2021 to December 31 2021. Exploratory information examination (EDA) is directed utilizing Jupyter Notebook to get an understanding of the data.

Data Pre-Processing

The imported data is filtered through data cleaning, duplicate data removal and data formatting. Information is additionally partitioned into two sets: training set and a testing set with the proportion of 80:20.

ARIMA time series model

ARIMA is a simplified form of the autoregressive moving average model. It incorporates autoregressive and moving average models to develop a composite forecasting model.

The AR model utilizes the reliance between the observations and several lagged observations, though the MA model uses the association between the observations and the residual error values by using the MA for the lagged observations. ARIMA uses the order factors p, d, and q. p is the order of the AR expression, q is the order of the MA expression, and d is the order of the differencing.

Mathematical Model

In a pure AR model, Y_t relies just upon its own lags. That is, Y_t is a function of the 'lags of Y_t'.

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \beta_3 Y_{t-3} + \ldots\ldots + \beta_p Y_{t-p} + \varepsilon_1$$

Y_{t-1} is the lag1 of the series, α_1 is the intercept term and β_1 is the coefficient of lag1. β_1 and α_1 are estimated by the model.

In a pure MA model, Y_t relies just upon the lagged forecast errors.

$$Y_t = \alpha + \varepsilon_t + \phi_1 \varepsilon_{t-1} + \phi_2 \varepsilon_{t-2} + \phi_3 \varepsilon_{t-3} + \ldots\ldots + \phi_q \varepsilon_{t-q}$$

The error terms are the errors of the AR models of the respective lags.

The errors E_t and $E_{(t-1)}$ are the errors from the accompanying equations :

$$Y_t = \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \beta_3 Y_{t-3} + \ldots\ldots + \beta_0 Y_0 + \varepsilon_t$$

$$Y_{t-1} = \beta_1 Y_{t-2} + \beta_2 Y_{t-3} + \beta_3 Y_{t-4} + \ldots\ldots + \beta_0 Y_0 + \varepsilon_{t-1}$$

In the ARIMA model, the time series is differenced at least once to make it stationary.

After you combine the AR and the MA expressions, the equation becomes:

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \beta_3 Y_{t-3} + \ldots\ldots + \beta_p Y_{t-p} \varepsilon_t + \phi_1 \varepsilon_{t-1} + \phi_2 \varepsilon_{t-2} + \phi_3 \varepsilon_{t-3} + \ldots\ldots + \phi_q \varepsilon_{t-q}$$

Data Visualization

We used Python libraries for data visualization. Bar and scatter plots are implemented for visualization with the use of Matplotlib, Seaborn, and Plotly.

Error Analysis

Mean absolute percentage error (MAPE) is a significant strategy in statistics that measures the prediction accuracy of forecasting.

R-squared is a mathematical measurement that addresses the degree of fluctuation for a subordinate variable that's

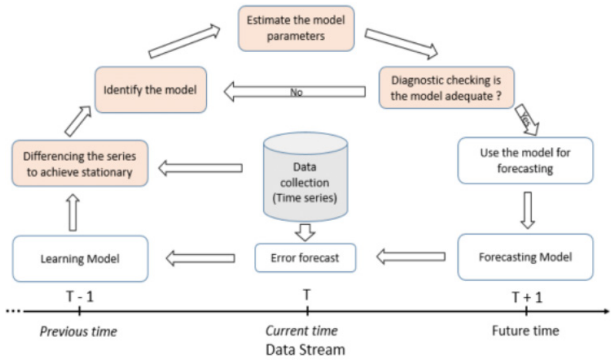

Figure 1: Proposed model

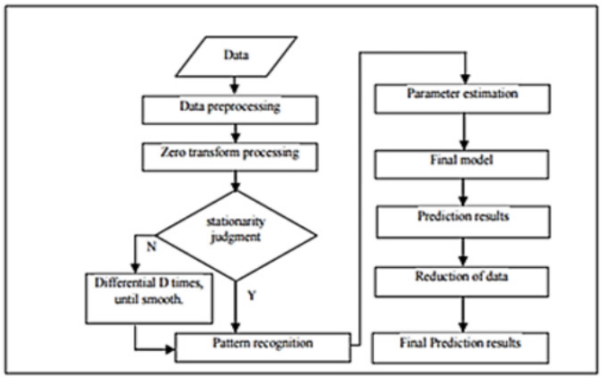

Figure 2 : Algorithm

clarified by a variable or variables in a relapse. The goodness of fit of a model could be estimated utilizing R^2 score.

Using Auto ARIMA the best fit model for this proposed system is SARIMAX.

Results

The website displays a homepage [Figure 3] which shows information about COVID-19, an explore page that analyzes and shows 10 days of COVID's insights for pune district [Figure

Figure 3 : Home Page

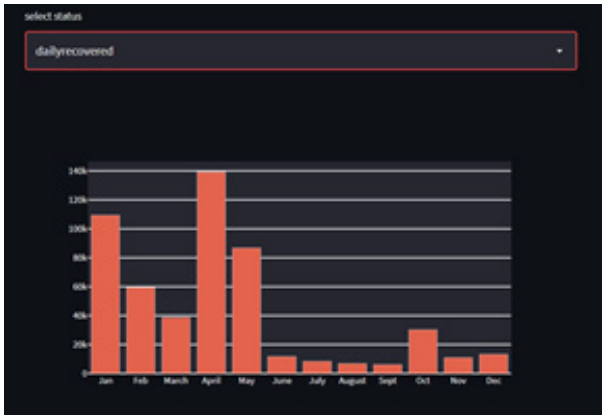

Figure 4 : Monthly trend in confirmed, recovered and deceased COVID-19 cases using bar plot

Figure 5 : Monthly trend in confirmed, recovered and deceased COVID-19 cases using scatter plot

Figure 6 : Variation in COVID-19 confirmed, recovered, deceased cases from last 10 days of 3 days rolling data

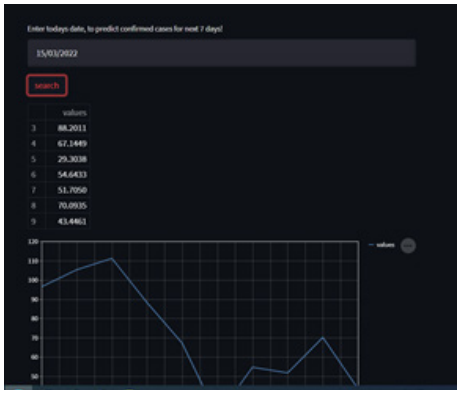

Figure 7 : Prediction of cases for next 7 days

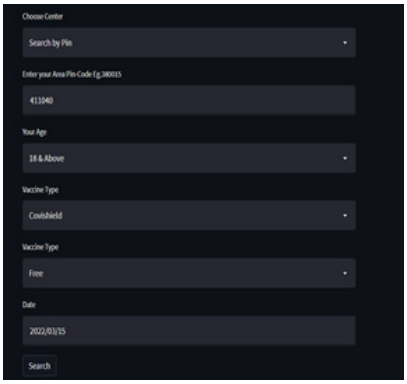

Figure 8 : Vaccination Availability search

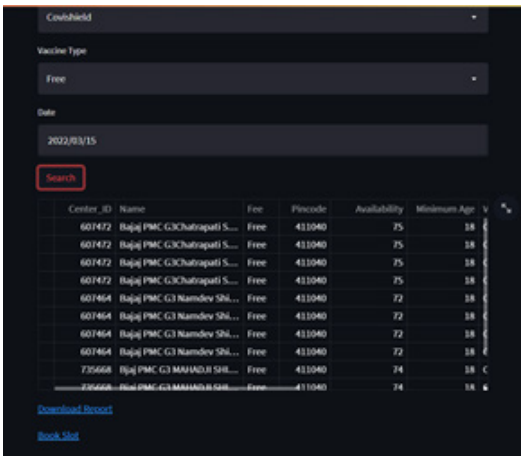

Figure 9 : Vaccination availability result with download option

4-6] and a prediction page that shows the prediction of confirmed COVID-19 cases of the next 7 days using graphs [Figure 7] with 92% accuracy. The website additionally permits clients to look for a vaccination center availability as per pin code [Figure 8] and download the data whenever required [Figure 9].

Conclusion and future scope

The paper successfully shows the analysis of the gathered COVID-19 data of the Pune district. Using the ARIMA forecasting model, successful forecasts of the number of active cases over the following 7 days from the present day is done. The model obtained a prediction accuracy of 92%. The research also aids in booking vaccination slots. This paper will help authorities contain the COVID-19 break by gaining caution.

We wish to expand our project to other districts in Maharashtra state. Live forecasting will be our main focus in the future. More effective algorithms can be developed for better prediction and understanding of the virus' spread with further availability of data. Hope this paper contributes to improving the district government's response to the COVID-19 pandemic and puts forward some references for future research.

References

[1] Shaikh, Saud, et al. "Analysis and Prediction of COVID-19 using Regression Models and Time Series Forecasting." 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence). IEEE, 2021.

[2] Sharma, Vikas Kumar, and Unnati Nigam. "Modeling and Forecasting of COVID-19 growth curve in India." Transactions of the Indian National Academy of Engineering 5.4 (2020): 697-710.

[3] Rishi Sharma, Mohit Kumar, Shishir Maheshwari, Kamla Prasan Ray. EVDHM-ARIMA-Based Time Series Forecasting Model and Its Application for COVID-19 Cases. IEEE.

[4] Jaglan, A., Trehan, D., Megha, U. and Singhal, P., 2020. COVID-19 trend analysis using machine learning techniques. Int J Sci Eng Res, 11(12), pp.1162-1167.

[5] Majhi, Ritanjali, et al. "Analysis and prediction of COVID-19 trajectory: A machine learning approach." Journal of Public Affairs 21.4 (2021): e2537.

[6] World Health Organization (WHO) (2020) Coronavirus https://www.who.int/health-topics/coronavirus Accessed July 13, 2021

[7] www.cessi.com Accessed July 10, 2021