International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT228134

Analysis and Prediction of COVID-19 Pandemic in India

K. M. Ravikumar[1], D. Chandrasekhar[2]
[1]PG Scholar, Department of CSE, Sarada Institute of Science Technology and Management, Ampolu road, Srikakulam, Andhra Pradesh, India
[2]Professor, Department of CSE, Sarada Institute of Science Technology and Management, Ampolu road, Srikakulam, Andhra Pradesh, India

Article Info
Volume 8, Issue 1
Page Number : 229-235

Publication Issue :
January-February-2022

Article History
Accepted : 05 Feb 2022
Published :18 Feb 2022

ABSTRACT

The real-time data has become a dominant aspect for understanding past, present, and future situations. Machine Learning (ML) is one platform that uses a variety of algorithms to provide the correlation between the given data, visualize the current scenario, and predict the future forecast, which is the most crucial part. The entire world is currently experiencing a devastating situation due to the outbreak of a novel coronavirus known as COVID19. The COVID19 at present has proved that it is a potential threat to human life. To contribute to controlling the spread and rising number of active cases in India, this study demonstrates the future forecasting of the total number of active cases in India in the upcoming days. Future Forecasting is performed using the ARIMA model (autoregressive Integrated moving average by combining Facebook) A prophet who gives us the highest precision. Real-time data Collection is done from different sources depending on the data preprocessing and data wrangling is done. The record is that it is divided into a training set and a test set. Finally, the model was trained and checked for accuracy. After the test with training, the model is ready to predict future predictions. The model also records predicted and actual values help him achieve higher accuracy in the future.

Index Term:- COVID19, ARIMA Model, Machine Learning, Time Series Analysis, Forecasting, R-Square score, root mean square error, mean squared error.

I. INTRODUCTION

Machine learning (ML), considered one of the most important courses in computer science in recent years, has been recognized for solving many problems in real time, including image processing ., medical diagnosis, financial analysis, etc. Many high-level applications such as Autonomous Vehicles (AV), intelligent robots, machine translation, product recommendations, and climate modeling use ML algorithms to give them the highest accuracy. Reinforcement learning, which comes into play when building ML models, not only avoids the use of traditional step-by-step coding instructions based on logic and if-then rules, but also improves their performance over time. Forecasting or predicting

future trends is the best place for ML to show off its skills. A variety of forecasts such as weather, national stock markets and many others use ML algorithms to predict the future so that necessary actions are taken. The rate at which the coronavirus infects humans is rapidly catching up with higher numbers. The need to reduce deaths and stabilize the country's economy has now become a top priority. The virus is transmitted from an infected person to the normal person through droplets from the mouth or nose of an infected person, as well as when a healthy person comes into contact with an infected surface.

Every individual must adhere to safety and disinfection measures to get through this COVID-19 pandemic. The chain should be broken by staying indoors and avoiding places affected by the virus. A terrifying situation like this forces progress in the field of research and development. Therefore, a number of diligent researchers from all scientific fields have tried to provide all possible solutions. To contribute to the current disaster, we have attempted to provide a future forecast for the COVID-19 pandemic. Forecast of an increase in the number of COVID19 cases over the upcoming days in India. AR (autoregression) and MA (moving average) models using time series analysis. The AR model restores the values of a variable belonging to the first periods as input to the regression equation which then predicts the output for the upcoming period. The MA model is a time series model that takes into account the possibility of a relationship between a variable and residuals from previous periods. The use of AR and MA models for prediction is not sufficient due to lack of accuracy. This shows that the ARIMA model is only valid if the variables are immovable. As of this day, time series performs conversion of the mobile variables into immobile variables using methods such as detrending or differencing for a convincing time series modelling. The conversion now becomes the first initiative to introduce the ARIMA model. ARIMA (p, q, d) denotes the ARMA model with p autoregressive lags, q moving average lags, and the variation in the order of d.

Helpful observations made in this study are listed below:

- The data set was taken from a real-time web, making the code for this study dynamic.
- Data visualization is done to better understand the current situation in India.
- ML algorithms need large amounts of data to make better predictions. As the size of the training dataset increases, the performance of the model increases. Therefore, for each prediction, the data from the previous day is added to the training data.
- Predictions based on ML algorithms can be very helpful in implementing protective measures and guiding action plans during pandemics like COVID19.

II. MATERIALS AND METHODS

A. Dataset

The purpose of this study is to predict the future of COVID19 Contagion focuses on the total number of active cases in India. The data set used for the study was taken from the official report government website, providing total active cases across India in upcomming days. Furthermore, data stored locally in the system using web scraping, the art of extracting a large volume of data from any web pages and data can then be stored in your local files computer or database. Tables I, II and III include samples data sets.

TABLE I
COVID-19 TIME-SERIES OF INDIA ON DAILY BASIS

S. No.	Date	Confirmed Cases
1	12-03-2020	74
2	13-03-2020	75
3	14-03-2020	84
.	.	.
24864	24-01-2022	349641119
24865	25-01-2022	353106672
24866	26-01-2022	356955803

TABLE II
COVID-19 PATIENT ACTIVE, CURED, AND DEATH STATEWISE

Date	Region	Confirmed Cases	Active Cases	Cured/Discharged	Death
12-03-2020	Andhra Pradesh	1	1	0	0
13-03-2020	Andhra Pradesh	1	1	0	0
14-03-2020	Andhra Pradesh	1	1	0	0
15-03-2020	Andhra Pradesh	1	1	0	0
16-03-2020	Andhra Pradesh	1	1	0	0
.	.	.	.	.	.
25-01-2022	West Bengal	1969791	94535	1854881	20375
26-01-2022	West Bengal	1974285	80168	1873706	20411
27-01-2022	West Bengal	1979254	67369	1891440	20445

TABLE III
COVID-19 VACCINATIONS STATUS

location	date	total_vaccinations	people_vaccinated	fully_vaccinated	total_boosters
India	15-01-2021	0	0	0	
India	16-01-2021	191181	191181	0	
India	17-01-2021	224301	224301	0	
India	18-01-2021	454049	454049	0	
India	19-01-2021	674835	674835	0	
India	20-01-2021	806484	806484	0	

B. Time-Series analysis with Auto-regressive Integrated Moving Average (ARIMA)

Auto-regressive Integrated Moving Average is conceivably a class related to fashions which get to the bottom of a specified census sustained with the aid of using its preceding values, its lags, and also lingered estimate errors. Any 'un-seasonal' statistic which showcases styles and now no longer belong to the non-linear noise desires to be formed primarily based totally on the prevailing version. This version is characterised with the aid of using three terms:

p is the order of the AR expression

q is the order of the MA expression

d mentioning the amount of difference had to shape the statistic immobility.

AR(p) Autoregression is a regression model that takes advantage of the dependant relationship between a current observation and prior data. The usage of past values in the regression equation for the time series is referred to as an auto regressive (AR(p)) component.

I(d) Integration - makes the time series stationary by differencing observations (subtracting an observation from an observation from the previous time step). Differencing is the process of subtracting a series' current values from its prior values d times.

MA(q) Moving Average - a model that takes advantage of the relationship between an observation and the residual error from a moving average model applied to lagged data. A moving average component represents the model's error as a sum of prior error terms.

The combined equation for ARIMA is given as follows:

$$y_t = C + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \ldots - \theta_q \varepsilon_{t-q}$$

Where,

Y = Value at time t *t*

C = Bias

φ = Auto -Regressive parameter

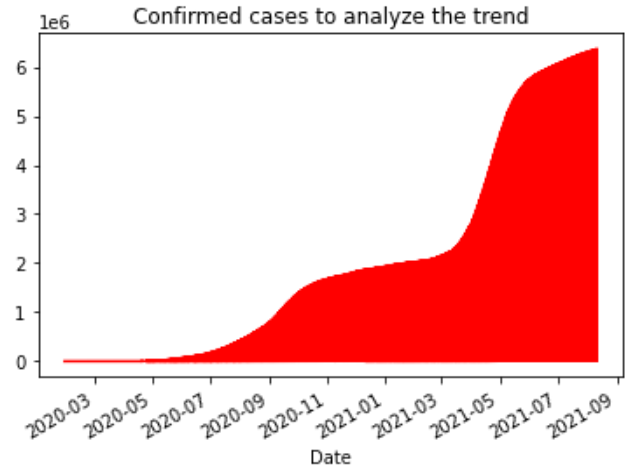
ε = Random error at t *t*

θ = Moving Average

Parameters Parameter estimation, model identification, and diagnostic checking are the phases in creating an ARIMA predictive model

III. METHODOLOGY

A study of the new coronavirus formerly known as COVID19 future predictions have attracted special attention from all over world. Help control the spread and increase the numbers active cases in India, this study attempts to demonstrate future prediction of the total number of active cases in India in the upcoming days. According to the research done for this study, we find that the ARIMA . model could be the right choice. To start with the dataset, we took from a real-time web. Extraction has been done possible with the use of web scraping. With the completion of web scraping, the data set undergoes data wrangling and data preprocessing after which the data gets stored in the local drive. Now the preprocessed data is visualized for a perfect overview of the data set. The splitting of the data set begins at this moment where it gets partitioned into train data of 85% along with test data of 15%. The 15% test data, taken from the same dataset, is unrevealed to the model during the training period. By hiding a part of the dataset helps in finding out whether the model has overfit or underfit, which are few of the biggest complications while training any model. The ARIMA Model gets trained by giving in the training data set. After training, the model is finally ready to go through trial phase. But the biggest complications in training any model. before the model goes through the testing phase, Facebook Prophet (Facebook Prophet is an online open source for time series analysis and future prediction) give in the upcomming days with the expected timestamp. Finally, the test dataset is added with the date and time stamps donated by Facebook Prophet. Currently, the model is trained on a total of active scenario models. ARIMA The model has been evaluated based on important indicators such as like RSquared, MAE, MSE and RMSE scores and reports in the results. Ship datasets are uploaded daily total number of active cases for better training model each time predict the forecast. Sample results



Date	Sno	Time	State/UnionTerritory	ConfirmedIndianNational	ConfirmedForeignNational	Cured	Deaths	Confirmed
2021-08-03	17787	8:00 AM	Andaman and Nicobar Islands	-	-	7404	129	7539
2021-08-03	17788	8:00 AM	Andhra Pradesh	-	-	1936016	13410	1970008
2021-08-03	17789	8:00 AM	Arunachal Pradesh	-	-	44823	234	48565
2021-08-03	17790	8:00 AM	Assam	-	-	550534	5294	568257
2021-08-03	17791	8:00 AM	Bihar	-	-	714872	9644	724917
2021-08-03	17792	8:00 AM	Chandigarh	-	-	61116	811	61960
2021-08-03	17793	8:00 AM	Chhattisgarh	-	-	987012	13528	1002458
2021-08-03	17794	8:00 AM	Dadra and Nagar Haveli and Daman and Diu	-	-	10631	4	10650
2021-08-03	17795	8:00 AM	Delhi	-	-	1410809	25054	1436401
2021-08-03	17796	8:00 AM	Goa	-	-	167118	3150	171295
2021-08-03	17797	8:00 AM	Gujarat	-	-	814595	10076	824922

Fig 1. Data Preprocessing

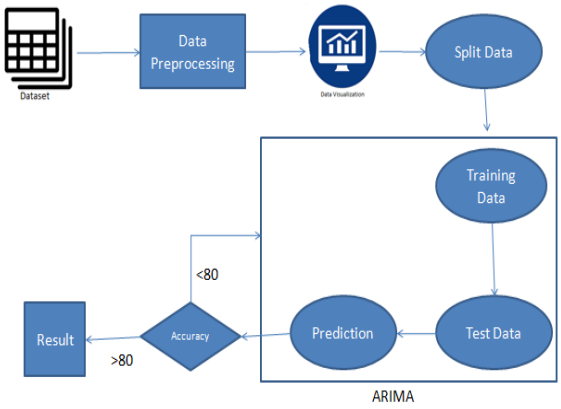


Fig 2. Black diagram

IV. RESULTS

Predict the total number of confirmed cases related to COVID19 in the upcoming days is main goal of this study. The data set includes the total number of confirmed, active cases and deaths reported in Figure 1.

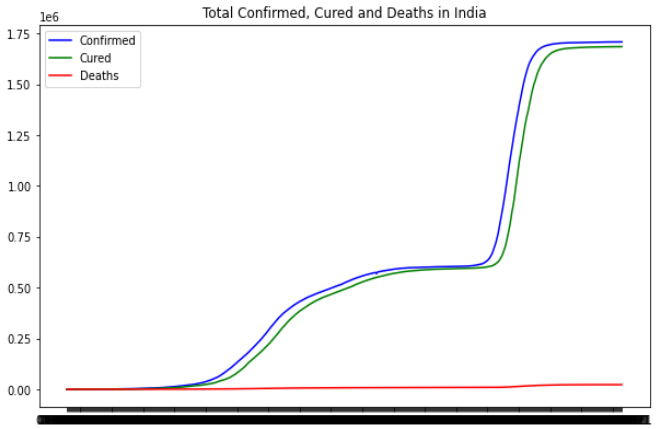


Fig. 3. Total confirmed, Active, Death in India

Total number of confirmed cases for future forecast Research makes predictions about the number of confirmed cases. According to R2 score, ARIMA model gave the best result for this situation as shown below in Figure 3.

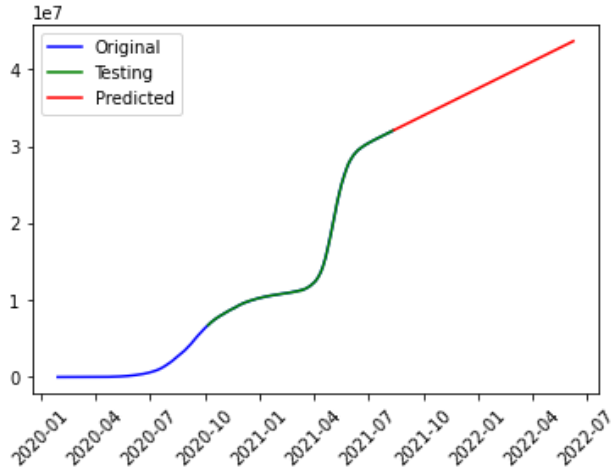


Fig 4. Predicted value for upcoming days

V. CONCLUSION

Because the number of victims and deaths is continually increasing, the COVID-19 continues to pose a potential threat around the world. The COVID-19 has clearly demonstrated its participation in the country's recent economic downfall on a massive scale. It has the capability of infecting anyone. There are serious concerns that COVID-19's economic consequences will be similar to those of the Great Depression. Using machine learning, this study forecasts the total number of active cases in India for

the following 15 days. The ARIMA model is best matched to this case, according to the findings of this study. The model's projection of the current condition will be useful in predicting the future. Overall, this research can assist authorities in gaining caution, which could aid in the containment of the COVID-19 situation. This research will be improved over time as the course progresses. The model might also be used to forecast the economy of a country in the event of a pandemic.

VI. REFERENCES

[1]. Dong E., Du H., Gardner L. An interactive web-based dashboard to track COVID-19 in real time Lancet Infect Dis (2020), 10.1016/S1473-3099(20)30120-1Online].

[2]. W. H. Organization E. Coronavirus disease (COVID-19) pandemic(2020)

[3]. Guorong Ding, Xinru Li, Yang Shen, Brief Analysis of the ARIMA model on the COVID-19 in Italy, medRxiv 2020.04.08.20058636.

[4]. Singh R.K., Rani M., Bhagavathula A.S., Sah R., Rodriguez-Morales A.J., Kalita H., Nanda C., Sharma S., Sharma Y.D., Rabaan A.A., Rahmani J., Kumar P.

[5]. Prediction of the COVID-19 pandemic for the top 15 affected countries: Advanced autoregressive integrated moving average (ARIMA) model JMIR Public Health Surv., 6 (2) (2020), Article e19115, 10.2196/19115 Bertozzi Andrea L., Franco Elisa, Mohler George, Short Martin B., Sledge Daniel

[6]. The challenges of modeling and forecasting the spread of COVID-19 Proc. Natl. Acad. Sci., 117 (29) (2020), pp. 16732-16738, 10.1073/pnas.2006520117 Li Lixiang, Yang Zihang, Dang Zhongkai, Meng Cui, Huang Jingze, Meng Haotian, Wang Deyu, Chen Guanhua, Zhang Jiaxuan, Peng Haipeng, Shao Yiming

[7]. Propagation analysis and prediction of the COVID-19 Infect. Dis. Model., 2468-0427, 5 (2020), pp. 282-292, 10.1016/j.idm.2020.03.002

[8]. S. Taylor and B. Letham, "Forecasting at scale. peerj preprints 5: e3190v2 (2017)."

[9]. "Time series forecasts using facebook's prophet," https://www.analyticsvidhya.com/blog/2018/05/generate-accurate-forecastsfacebook-prophet-python-r/#:~:text=Prophet%20is%20an%20open%20source,of%20custom%20seasonality%20and%20holidays!

[10]. O. Renaud and M.-P. Victoria-Feser, "A robust coefficient of determination for regression," Journal of Statistical Planning and Inference, vol.140, no. 7, pp. 1852–1862, 2010.

[11]. W. Wang and Y. Lu, "Analysis of the mean absolute error (mae) and the root mean square error (rmse) in assessing rounding model," in IOP Conference Series: Materials Science and Engineering, vol. 324, no. 1,2018, p. 012049.

[12]. H. Witzgall and J. Goldstein, "A realizable mean square error estimator applied to rank selection," in Conference Record of the Thirty-Sixth Asilomar Conference on Signals, Systems and Computers, 2002., vol. 1. IEEE, 2002, pp. 881–884.

[13]. "Global impact of new corona virus and population issues — inter press service,"http://www.ipsnews.net/2020/05/global-impact-newcorona-virus-population-issues/.

[14]. S. Azad and N. Poonia, "Short-term forecasts of COVID-19 spread across Indian states until 1 may 2020," Preprints, 2020.

[15]. P. Wang, X. Zheng, J. Li, and B. Zhu, "Prediction of epidemic trends in COVID-19 with logistic model and machine learning technics," Chaos Solitons Fractals, vol. 139, no. 110058, p. 110058, 2020.

[16]. S. J. Taylor and B. Letham, "Forecasting at scale," Am. Stat., vol. 72, no. 1, pp. 37–45, 2018.

[17]. "Prophet: automatic forecasting procedure, (EB/OL])," Online]. Available: https://facebook.github.io/prophet/docs/ or https://github.com/facebook/prophet. Accessed 11 Aug 2020].

ABOUT AUTHORS :

K.M. Ravikumar is currently pursuing her M.Tech (CSE) in Computer Science Department, Sarada Institute of Science Technology and Management, Ampolu road, Srikakulam, A.P. He receiving his M.Tech in CSE from SITM, Srikakulam.

D. Chandrasekhar is currently working as an Assistant Professor in Computer Science Department, Sarada Institute of Science Technology and Management, Srikakulam, A.P. His research includes data mining and Machine learning

Cite this article as :

K. M. Ravikumar, D. Chandrasekhar, "Analysis and Prediction of COVID-19 Pandemic in India ", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 8 Issue 1, pp. 229-235, January-February 2022. Available at doi : https://doi.org/10.32628/CSEIT228134
Journal URL : https://ijsrcseit.com/CSEIT228134