Covid-19 Case Prediction using Machine Learing

Meet Pedhadiya
Information and Communication Technology
School of Engineering and Applied Science, Ahmedabad University
Ahmedabad, India
email: meet.p2@ahduni.edu.in

Ayush Kaneria
Information and Communication Technology
School of Engineering and Applied Science, Ahmedabad University
Ahmedabad, India
email: ayush.k1@ahduni.edu.in

Krunal Pagdar
Information and Communication Technology
School of Engineering and Applied Science, Ahmedabad University
Ahmedabad, India
email: krunal.p@ahduni.edu.in

Mihir Chauhan
Information and Communication Technology
School of Engineering and Applied Science, Ahmedabad University
Ahmedabad, India
email: mihir.c@ahduni.edu.in

***Abstract*— Covid-19 diseases spread all around the world for more than 1 year and WHO declares COVID-19 as global pandemic. Today number of covid-19 cases exceed 100 Million+ across the globe and deaths due to Covid-19 exceed 2.5M + across the globe. In this pandemic situation COVID-19 case prediction might help to the medical management for distributing medical recourses and take care of all the precautions. In this paper we have described our approach and work done for the implementing machine learning (ML) algorithm to predicts the COVID-19 cases in India. This paper contains implement Polynomial regression for degree=3 but has not good accuracy, higher accuracy can be achived at the higer value of the polynomial degree but after literature surevey we moved to time series analysis to predict 'Daily confim case' in India. In time series analysis we check for stationarity of the data and make data stationarity. Using partial Auto correlation (PAC) and Auto Correlation (AC) got the parameter to perfom the ARMA and ARIMA model. Using this models we predict the future 'Daily confirm cases' of covid-19.**

Keywords—Covid-19, Pandemic, Prediction, daily confirmed cases, Active Cases, Daily Tests, Algorithms, Time series analysis, Moving Average, Auto regression, ARMA, ARIMA

I. INTRODUCTION

The SARS-CoV-2 is the cause of Covid-19 diseases which originated in the city Wuhan, China at the end of the year 2019. In the beginning, it affects Wuhan very badly. Gradually it spreads all around the world. It has been a big threat to global health as WHO (World Health Organization) declared COVID-19 diseases as a global pandemic in March 2020.Milions of people have been affected and lakhs of people lost their lives due to the COVID-19 disease. At the end of June 2020 number of COVID-19 cases are exceeded 10 Million+ across the globe. India reported the first case of COVID-19 on 30th January 2020. During that time international traveller, Indian students who studied abroad came back to India, etc. results in an increase in cases of COVID-19 in India. Today India has crossed 10 Million + cases of COVID-19. In a rapidly evolving pandemic, it is important to have proper analysis and prediction of cases in the future. Inefficient distribution of medical resources results in improper medical facility and it affects the recovery rate. If we have a good model that predicts the new cases well, it will result in the efficient distribution of medical resources in arising cases in covid-19.

We found an efficient dataset that contains daily positive cases and deaths due to the COVID-19 disease of India from the [4] www.covid19india.org . The dataset contains daily cases, deaths, tested reports, etc. since the first registered case in India. Dataset also contains State-wise data [3].

II. LITERATURE SURVEY

This Covid-19 pandemic had a major impact on the health and well-being of the global population. Various research works have been done to forecast the cases of COVID-19.

A.Khakharia et al.[1] have developed an outbreak prediction system for COVID-19. For that, they analyze data of the top 10 highly dense and populated countries (India, Bangladesh, the Democratic Republic of Congo, Pakistan, China, Philippines, Germany, Indonesia, Ethiopia, and Nigeria). In the paper, the authors proposed a prediction model to forecast the count of new cases likely to arise in 5 successive days. For the prediction, the author has used 9 different machine learning (ML) algorithms. The 9 different algorithms were Auto-Regressive Moving Average (ARMA), Auto-Regressive Integrated Moving Average (ARIMA), Support Vector Regressor (SVR), Linear Regressor polynomial (LRP), Bayesian Ridge Regression (BRR), Linear Regression (LR), Random Forest Regressor (RFR), Holt-Winter Exponential Smoothing (HW), and Extreme Gradient Boost Regressor (XGB).

The author found that different country has different behaviour of increasing and decreasing in COVID-19 cases. So that not every algorithm could give high accuracy of prediction of cases for each country. They implemented all 9 algorithms to different 10 countries' data. They found that different countries have different accuracy with different models implemented. Some counties have good accuracy with one algorithm while other countries may not have good accuracy with that particular algorithm. Particularly we analyse for Indian behaviour of covid-19 cases then we got a high accuracy in the time series algorithms like ARMA and ARIMA.

Y. Zoabi [2] has implemented a model that predicts COVID-19 diagnosis based on symptoms. They have established a machine learning approach that trained on

records from 51,831 tested individuals. The model is using 8 binary features i.e., sex, age>60 years, known contact with an infected individual, and the appearance of five initial clinical symptoms.

Hyndman, R[3] and Time series analysis[4] has given detailed explanation on the time series analysis and forecasting. They also explaim Moving average, Auto regression , ARMA model and ARIMA model in detailed.

III. IMPLEMENTATION

There are many sources available on the internet that provides a COVID-19 datasheet. Some of the datasets have a large number of errors or very less amount of data. We found a data set from the website "covid19india.org" which has the most appropriate dataset. These datasets contain different kinds of information like all India, state-wise, district-wise, and with their daily information and time-series and in long formation about COVID-19 cases. We have done analytics on the datasets and compare them. We also updated datasets regularly so that we can achieve better implementation of model and for better prediction.

For better furcating we updated dataset with the new column i.e. Daily confirm case, Daily deaths, etc.

Daily confirm cases is calculated by subtracting Total Confirm cases by its previous value.

We have done some data analysis on the datasets and compre the dataset with the different dataset. We compared data analysis for all states wise data as well. Data is almost the same in all cases.

A. Polynomial Regression

Polynomial Regression is one of the forms of regression analysis methods. In polynomial regression, the relationship between the independent variables x and dependent variables y is modeled as an nth degree polynomial in x.

We have used polynomial function available in the sci-kit learn library. Sci-kit learn is a very useful python library that contains many machine learning algorithms. The dataset contains a dependent variable as Daily Confirm cases and an independent variable as Daily Test.

$$y = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \alpha_3 x^3 + \cdots + \alpha_n x^n \quad (1)$$

First, we have imported the sci-kit learn library and other dependencies. At the beginning of the implementation, we started with parameters normalize = True, fitintercept = True, and order=F (for fast computation). We implemented polynomial regression on datasets of covid-19 daily cases of India. Both datasets have different natures in increase and decrease daily cases of covid-19. Both the dataset had a good training efficiency at the different degrees of polynomial. For dataset India, we plot two different graphs with degree = 3 and 10.

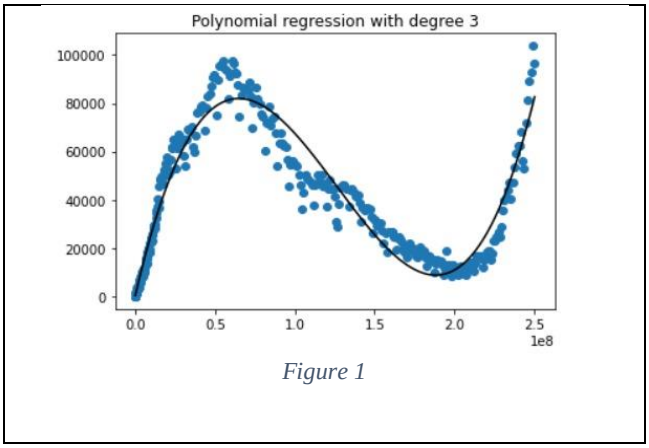

Figure 1

As we can see the results of polynomial regression in figure (c) and figure (d). From that, we can conclude that if we want to achieve high accuracy then we have to increase the degree of a polynomial.

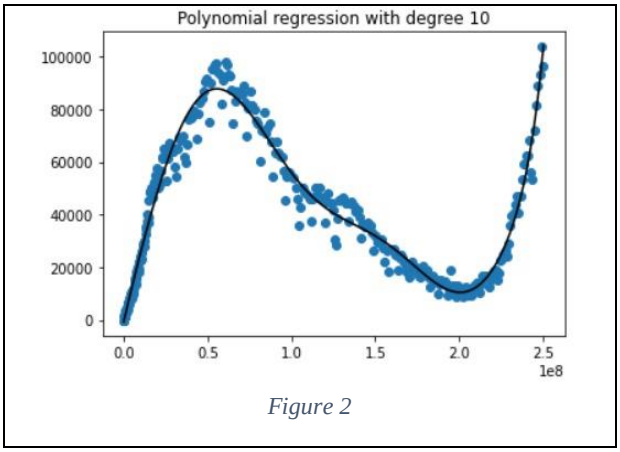

Figure 2

We have good efficiency for the higher degree of the polynomial but higher degree polynomial should not be considered as a better model. And on the other side, we have taken the independent variable as a Daily Test. Yet, Ideally, we realize that 'Daily Test' isn't the real explanation of the reliance of the spread of the Coronavirus cases. We have reffered litureature and got to know that we can achive a higer accuracy and good prediction for the problems like covid-19 cases where we do not have proper independent feature. Instead of polynomial regression we can implement regression alorithms of time series analysis like ARMA and ARIMA for the better short term prediction. A.Khakharia et al.[1] has a implement many algorithms to the dataset of covid-19 and got very good results for the dataset for Inida.

B. Time Series Analysis

There are certain steps to perform time series analysis on the data. We have performed following steps to analys time series.

First we checked wheter the data is stationary or not. For that we performed two test test- Rolling test and Dickey Fuller test. From that we got the p-value = 0.889. Then we plot the Estimated trend. After that we find the difference between moving average and acutal value for making time series stationary. In that we convert the data into the log

scale and drop the null value (i.e. NaN). After that we test the data with Augmented Dickey-Fuller (ADCF) test by performing rolling statistics and Dickey-Fuller test. After this test we got the stationarity in data with p-value = 9.9e-5. After that we shifted the value in time series so that we can use those values in forcasting and drop the null values. Then we plot the decompose of three main component of time series- Trend, Seasonal and Residual.

1. Moving Average

Moving Average is a statistical term which uses the series mean and previous errors to regress the data and forecast future values.[3]

Moving average of order q can be written as,

$$\mathbf{Y_t} = \boldsymbol{\alpha} + \boldsymbol{\epsilon_t} + \emptyset_1\boldsymbol{\varepsilon_{t-1}} + \emptyset_2\boldsymbol{\varepsilon_{t-2}} + .. + \emptyset_q\boldsymbol{\varepsilon_{t-q}} \quad (2)$$

$where\ \emptyset_k = parametere\ of\ the\ model$
$\varepsilon_k = error\ term$
$q = order\ of\ the\ moving\ average\ (lag)$
$\alpha = constant\ term$

In our model we first performed moving average on 'Daily confirm cases' column in the excel. Fisrt we understood the concept of Moving Average and then perform Moving Average in Excal. We got the efficient output in the Excel that is shown in Figure 5.

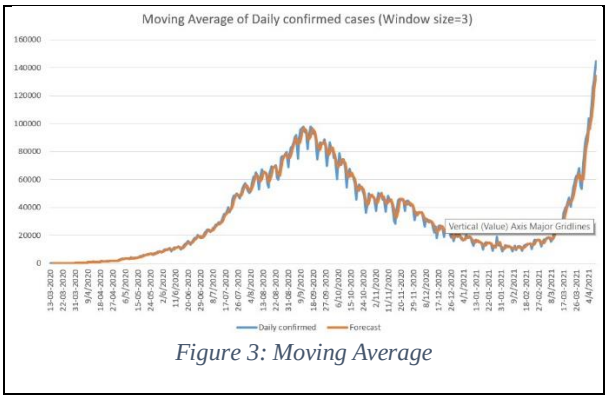

Figure 3: Moving Average

2. Auto Regression

In Autoregession (AR) we forcast the modle by taking linear combination of previous values of dependent variable. The term autoregression indicates that regression of the dependent variable with itself.[3]

Auto regressive model of order q can be written as,

$$\boldsymbol{Y_t} = \boldsymbol{\beta} + \emptyset_1\boldsymbol{y_{t-1}} + \emptyset_2\boldsymbol{y_{t-2}} + .. + \emptyset_q\boldsymbol{y_{t-q}} + \boldsymbol{\varepsilon_t} \quad (3)$$

$where\ \emptyset_k = parametere\ of\ the\ model$
$\varepsilon_t = noise\ term$
$q = order\ of\ the\ auto\ regression\ (lag)$
$\beta = constant\ term$

We have implement autoregrssion for the 'Daily Confirm cases' for the prediction purpose. First we imported the dataset and then we trained the model as per the training data and test data. We used the python library 'statsmodels.api'. Then we passed parameter of auto regression q and change the parameter values based on what the model fits well. We got the results shown in Figure. We got the efficient output in the both case- from Excel and from python library.

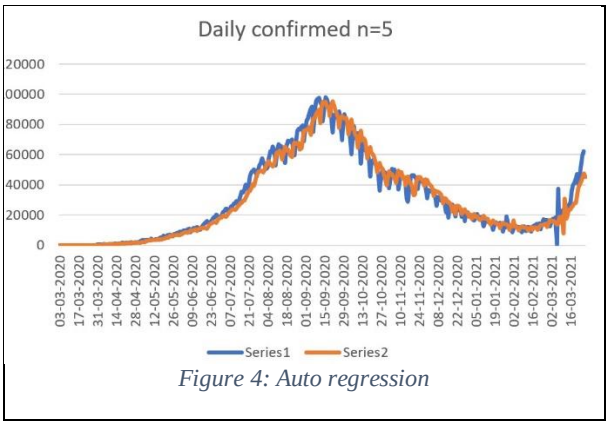

Figure 4: Auto regression

3. ARMA

ARMA stands for Auto Regressive Moving Average. From the name of the ARMA we got the two term Auto Regression and Moving Avereage. ARMA is a combination of Autoregression and Moving average of the data. ARMA provides weakly stochastic time series analysis in terms of two polynomial – Auto regression and Moving Average. To perform ARMA on the dataset, data must be stationary. For that p-value of data must be less then 0.05.[3]

Equation of the ARMA can be written as,

$$Y_t = \sum_{i=0}^{p}(\emptyset_i y_{t-i}) + \sum_{j=0}^{q}(\theta_j \varepsilon_{t-j}) + \varepsilon_t \quad (4)$$

We have implement ARMA model to forecast Daily Confirm cases of covid-19. For that, we have loaded the dataset and use in-buit python library 'statsmodels.api' to forecast the data using ARMA algorithm. And then put the parameter values. We changed parameter value to get the best fitted model. We got the best prediction for the value of p = 2, q=3. Summary of the ARMA model is shown in Figure 7.

ARMA Model Results

Dep. Variable:	y	No. Observations:	389
Model:	ARMA(5, 3)	Log Likelihood	207.160
Method:	css-mle	S.D. of innovations	0.140
Date:	Tue, 06 Apr 2021	AIC	-394.320
Time:	08:37:04	BIC	-354.684
Sample:	0	HQIC	-378.607

	coef	std err	z	P>|z|	[0.025	0.975]
const	7.1543	4.179	1.712	0.088	-1.036	15.345
ar.L1.y	0.4159	0.161	2.585	0.010	0.101	0.731
ar.L2.y	-0.2897	0.100	-2.897	0.004	-0.486	-0.094
ar.L3.y	0.7850	0.091	8.633	0.000	0.607	0.963
ar.L4.y	0.2683	0.155	1.731	0.084	-0.035	0.572
ar.L5.y	-0.1810	0.054	-3.366	0.001	-0.286	-0.076
ma.L1.y	0.4766	0.160	2.983	0.003	0.163	0.790
ma.L2.y	0.9930	0.057	17.301	0.000	0.880	1.105
ma.L3.y	0.1481	0.154	0.964	0.336	-0.153	0.449

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	-0.2115	-0.9841j	1.0066	-0.2837
AR.2	-0.2115	+0.9841j	1.0066	0.2837
AR.3	1.0006	-0.0000j	1.0006	-0.0000
AR.4	-1.9256	-0.0000j	1.9256	-0.5000
AR.5	2.8302	-0.0000j	2.8302	-0.0000
MA.1	-0.1695	-1.0158j	1.0299	-0.2763
MA.2	-0.1695	+1.0158j	1.0299	0.2763
MA.3	-6.3673	-0.0000j	6.3673	-0.5000

Figure 5: Summary of the ARMA model on 'Daily Confirm cases'

ARIMA Model Results

Dep. Variable:	D.y	No. Observations:	392
Model:	ARIMA(3, 1, 3)	Log Likelihood	223.383
Method:	css-mle	S.D. of innovations	0.136
Date:	Sun, 11 Apr 2021	AIC	-430.765
Time:	08:56:26	BIC	-398.995
Sample:	1	HQIC	-418.174

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.0553	0.041	1.333	0.183	-0.026	0.137
ar.L1.D.y	0.5799	0.015	37.685	0.000	0.550	0.610
ar.L2.D.y	-0.5643	0.018	-30.558	0.000	-0.600	-0.528
ar.L3.D.y	0.9713	0.016	60.702	0.000	0.940	1.003
ma.L1.D.y	-0.6571	0.028	-23.068	0.000	-0.713	-0.601
ma.L2.D.y	0.6603	0.023	28.338	0.000	0.615	0.706
ma.L3.D.y	-0.9083	0.028	-32.998	0.000	-0.962	-0.854

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	1.0055	-0.0000j	1.0055	-0.0000
AR.2	-0.2123	-0.9894j	1.0119	-0.2836
AR.3	-0.2123	+0.9894j	1.0119	0.2836
MA.1	1.0440	-0.0000j	1.0440	-0.0000
MA.2	-0.1586	-1.0146j	1.0269	-0.2747
MA.3	-0.1586	+1.0146j	1.0269	0.2747

Figure 6: Summary of ARIMA for Daily Confirm cases

4. ARIMA

ARIMA stands for Auto Regressive Integrated Moving Average. ARMA model is stationary model. If your model is not stationary then you can make the model stationary by taking series of combination. ARIMA algorithm first makes your model stationary if it was not and then regress the model and make prediction. If your model is stationary then ARIMA is same as ARMA.[1]

Equation of ARIMA can be written as,

$$l_t = \beta_0 + \sum_{i=0}^{p} (\beta_i l_{t-i}) + \sum_{j=0}^{q} (\emptyset_j \varepsilon_{t-j}) + \varepsilon_t \quad (5)$$

Where $l_t = last\ term$

We have implement ARIMA algorithm to forecast Daily Confirm cases of covid-19. For that first, we have loaded the dataset and use in-buit python library 'statsmodels' to forecast the data using ARIMA algorithm. And we implemented time series analysis steps to find the parameter, autoregressive lag (p), moving average (q) and order of diffrentiaton (d).

Then we found the value of p,q and d using AutoARIMA function for best fitting of model.

We got the best prediction for the value of p = 3, q=3 and d = 1. Summary of ARIMA model for the 'Daily confirm case' is shown in Figure 8.

IV. Result and Conclusion

We have used ARMA and ARIMA model to forecast the data of Daily Confirm cases of covid-19 in India.

Figure 9 shows the result of forecst using ARMA for the 'Daily confirm cases' for India.

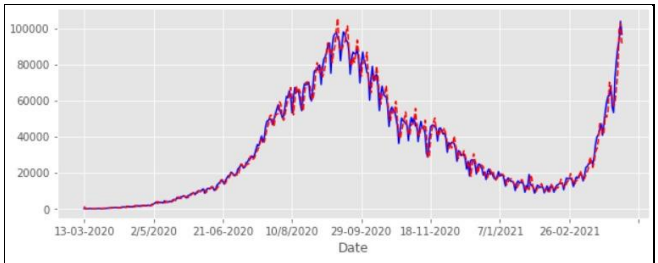

Figure 7: ARMA Forecast

Figure 10 shows the result of forecst using ARIMA for the 'Daily confirm cases' fot India.

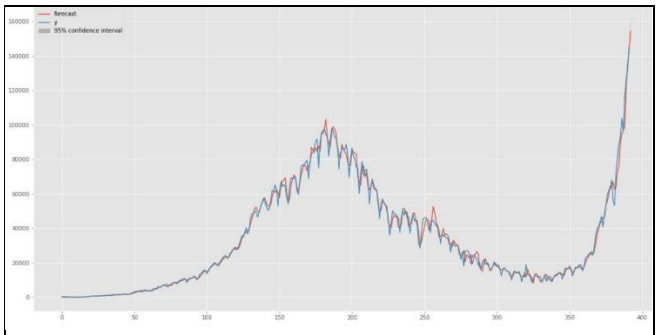

Figure 8: ARIMA Forecast

Figure 11 shows the acutal prediction of the 'Daily confirm cases' for the next 7 days using ARMA model. For this model we got the accuracy of 97%

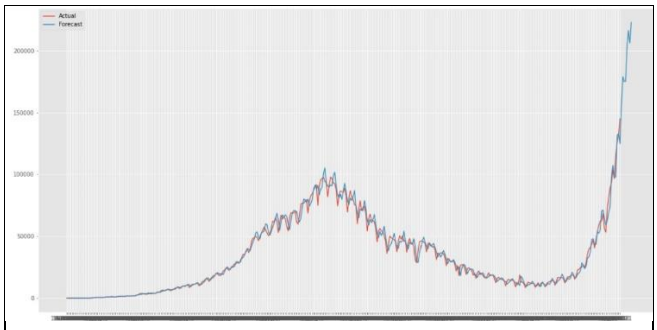

Figure 9: Prediction of 'Daily confirm cases' using ARMA

Figure 12 shows the acutal prediction of the 'Daily confirm cases' for the next 7 days using ARIMA model.Accurecy of the model is 98%

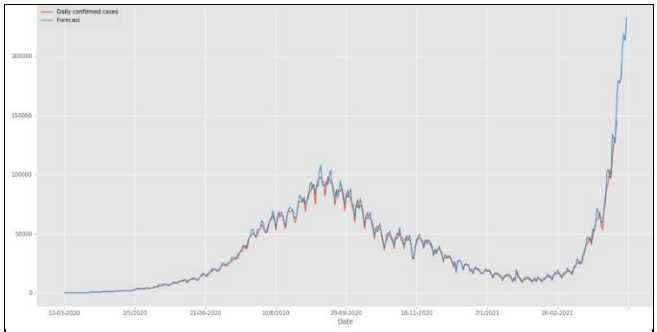

Figure 10: Prediction of 'Daily confirm cases using ARIMA

REFERENCES

Article reference:

[1] A. Khakharia et al., "Outbreak Prediction of COVID-19 for Dense and Populated Countries Using Machine Learning," Ann. Data Sci., vol. 8, no. 1, pp. 1–19, 2021, doi: 10.1007/s40745-020-00314-9.

[2] Y. Zoabi, S. Deri-Rozov, and N. Shomron, "Machine learning-based prediction of COVID-19 diagnosis based on symptoms," npj Digit. Med., vol. 4, no. 1, pp. 1–5, 2021, doi: 10.1038/s41746-020-00372-6.

[3] Hyndman. R. and Athanasopoulos. G.. 2021. *Forecastina: Principles and Practice*. 2nd ed. [ebook] Available at: https://otexts.com/fpp2/index.html

[4] Time series analysis. (2020, June 17). Retrieved April 11, 2021, from https://www.statisticssolutions.com/time-series-analysis/

Dataset reference:

[4] https://www.covid19india.org/

[5] https://github.com/nshomron/covidpred