RESEARCH ARTICLE OPEN ACCESS

Prediction of COVID-19 Cases with Time Series Analysis and Machine Learning

Sirisha Alamanda[1], Suresh Pabboju[2], Rishitha Reddy[3], Sheetal Naini[4]
Department of Information Technology, Chaitanya Bharathi Institute of Technology, Hyderabad
Email: asirisha_it@cbit.ac.in
Department of Information Technology, Chaitanya Bharathi Institute of Technology, Hyderabad
Email: psuresh_it@cbit.ac.in
Department of Information Technology, Chaitanya Bharathi Institute of Technology, Hyderabad
Email: ugs18012_it.rishitha@cbit.ac.in
Department of Information Technology, Chaitanya Bharathi Institute of Technology, Hyderabad
Email: ugs18017_it.sheetal@cbit.ac.in

Abstract:

People in Millions have been infected and lakhs of people have lost their lives due to the Coronavirus. It is of utmost importance to predict the future infected cases and the rate of virus spread for advance preparation in the healthcare domain to avoid deaths. For the research community forecasting the spread of COVID-19 accurately, is an analytical and challenging problem in real-world. In this work time series analysis is done on the dataset taken from the 'www.worldometers.info/coronavirus' website. A polynomial regression technique and neural net model are used to do the statistical prediction of active cases and deaths for various countries, and the best models with least root mean square error is selected for the prediction. Being able to accurately predict the time of outbreak would significantly minimize the impact of the virus and help the government to be prepared to control the spread of virus.

***Keywords* — Time Series Analysis, Polynomial Regression, Machine Learning, Time Series Forecasting, Pandemic, COVID-19.**

I. INTRODUCTION

The COVID-19 pandemic had created a huge havoc in the entire world. As many people have affected in the world and so many have lost their life (3,220,000 deaths as on 1st May, 2021), it has compelled people to stay at their homes and has caused a huge damage to the world's economy. The first reported COVID-19 case in India was on 30th January. India has observed a sudden hike in the number of cases and since then, the numbers are increasing day by day. As of 1st May 2021, India has quite 300,000 active cases with around 220,000 deaths and is currently world's most infected country.

Time series data analysis and forecasting have become significantly important these days. Time series forecasting is a technique to predict future events by analysing the trends of the past based on the assumption that future trends will be similar to historical trends. Pandemics like COVID 19 will occur more frequently and with increased impact in the recent past, people, the health sector and the government have to be prepared to face and handle such pandemics. Time series prediction helps us to note the trends and analyse the speed at which the virus is spreading. This will help us in preparing ourselves and taking all the required measures to be ready and alert.

In this work, machine learning models are employed to study and understand the future trend of the COVID-19. For this, work polynomial regression model and neural net model have been used to a examine the COVID-19 pandemic accordingly. This work predicts the count of cases and count of deaths around the world and in India particularly for the upcoming days. The forecasting results can assist governments to plan policies to control the spread of the virus. This work could help in administrations and health care systems to plan in advance and make arrangements for reducing the stress due to future pandemic situation.

II. RELATED WORK

From the time of declaration of COVID pandemic, many researchers have worked [8], [9] on several issues of COVID-19. In [7], an autoregressive model is proposed to determine the worldwide COVID-19 “confirmed” and “recovered” cases. Their work supported a variation of Gaussian distribution (GD) known as two-piece scale mixture GD and performance is compared with standard Gaussian autoregressive model. The results in their work have showed that the proposed work has performed well in forecasting confirmed and recovered COVID-19 cases within the world. In [5], authors have worked on Kaggle dataset with multilayer perceptron and vector autoregressive models for forestalling the long term COVID-19 effects in India.

[1] presented a deep learning approach to forecast COVID-19 cases using LSTM. The COVID-19 data utilized in their work is collected from Canadian Health authority and Johns Hopkins University. Given number of confirmed cases until March 31, 2020, the results of their work supported the prediction of possible ending point of this outbreak to be around June 2020. [2] had adopted the concept of Support Vector Machine (SVM) on time series data for the prediction covid-19 pandemic. Application of machine learning Time series analysis for prediction COVID-19 pandemic [3] has used several forecasting techniques like moving average, exponential smoothing, naive method, Holt linear trend method, Holt Winter method and ARIMA for comparison. The results of their work described the naive method as the best. [4] has used the prediction models based on genetic programming(GP) for confirmed cases and death cases across three most affected states i.e Delhi, Gujarat and Maharashtra. The results of their work have showed that the GEP-based models are highly reliable for statistic prediction of COVID-19 cases in India. In [5] and a series of machine learning techniques were studied including linear regression, multi-linear regression, polynomial regression and Lasso regression models. In addition, in order to improve the accuracy, they have focused on cumulative features. The results in their work showed that multi-linear regression and multi-polynomial regression are suitable to predict the COVID-19 pandemic. [6] has used several models, first the Holt model which employs double exponential smoothing; The second was the Autoregressive Integrated Moving Average (ARIMA) model, third was the (Trigonometric Exponential smoothing state space model with Box-Cox transformation, ARMA errors, Trend and Seasonal component) TBATS model. The last model used was the cubic smoothing spline model it's been shown that this model may be a special case of an ARIMA (0, 2, 2) model. Its advantage over the complete ARIMA model is that it provides a smooth historical trend also as a linear forecast function. In this current work, we have used machine learning models including polynomial regression model and neural net model to study and understand the future trend of the COVID-19

III. METHODOLOGY

Firstly, we grab the data from the website using data grabbers and do the necessary processing. Training and testing of the data is done using several machine learning models and an optimal model is selected. We can now view the predicted number of cases or deaths for any number of days as desired. Time series graph is also displayed which gives us a clearer idea of the worsening situation. The work flow of the proposed work is shown in Fig. 1.

Using the proposed work, we can predict the statistics for active cases and deaths of various countries as per our wish and the world aggregate as well. We can also alter the number of days till which we would like to know the prediction. With this model, one can get the anticipated cases/deaths count as well as the time series graph for the specified country. The time series graphs for deaths in India are displayed as shown in the Fig 4. Along with the predicted cases/deaths count, we have included the feature that displays the root mean square error as well.

We have employed two algorithms in our work,

i) Polynomial Regression

ii) Neural Network model

We have the liberty to execute the code with either one of the models and compare the obtained root mean square error. According to our study, Polynomial Regression model is more accurate for this proposed work.

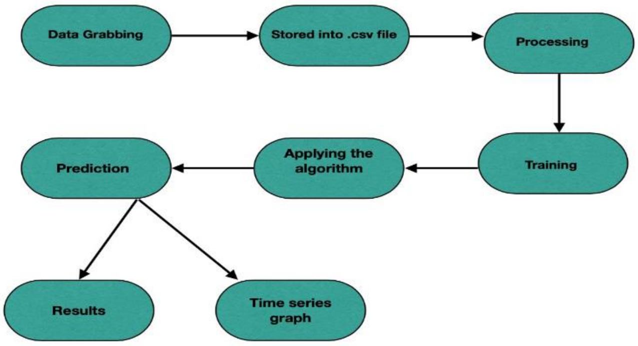
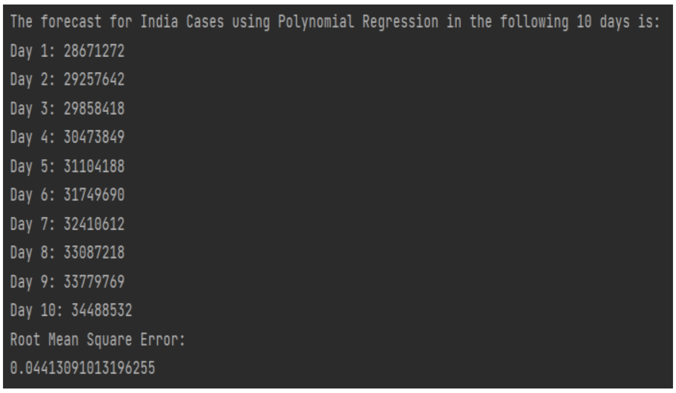


Fig. 1. Workflow of the proposed work

IV. RESULTS AND DISCUSSION

In this work, the machine learning models i.e. Polynomial Regression and Neural Net models are trained on the dataset taken from the website ‘www.worldometers.info/coronavirus’. This dataset contains information about the observation data, provenance/ state, country/region, active cases, deaths and latest updates. The results obtained for statistical prediction of active cases and deaths for various countries from the polynomial regression technique and neural net model are compared and the best model with least root mean square error is selected for the prediction.

The results obtained for statistical prediction of active cases using polynomial regression for the next 10 days in India are shown in Fig.2. and the comparison of actual and the prediction are shown in Fig.3.



```
The forecast for India Cases using Polynomial Regression in the following 10 days is:
Day 1: 28671272
Day 2: 29257642
Day 3: 29858418
Day 4: 30473849
Day 5: 31104188
Day 6: 31749690
Day 7: 32410612
Day 8: 33087218
Day 9: 33779769
Day 10: 34488532
Root Mean Square Error:
0.04413091013196255
```

Fig. 2. Forecast for Active Cases in India for the next 10 days

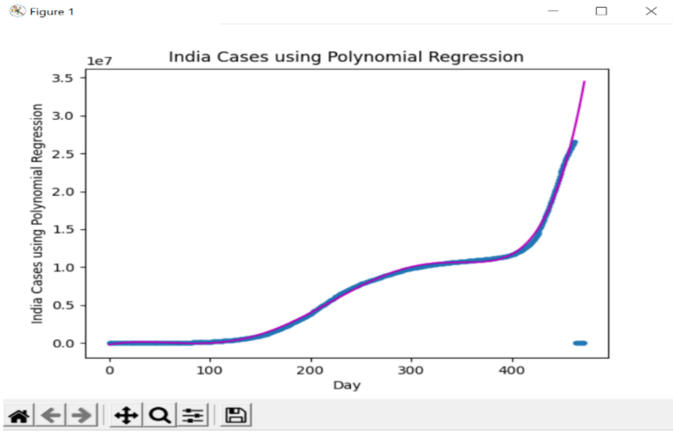
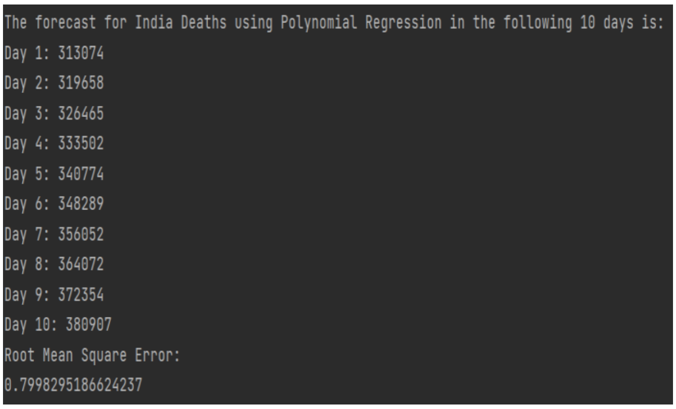


Fig. 3. Graph for Active Cases in India for the next 10 days

The results obtained for statistical prediction of no. of death cases using polynomial regression for the next 10 days in India are shown in Fig.4. and the comparison of actual and the prediction are shown in Fig.5.



```
The forecast for India Deaths using Polynomial Regression in the following 10 days is:
Day 1: 313074
Day 2: 319658
Day 3: 326465
Day 4: 333502
Day 5: 340774
Day 6: 348289
Day 7: 356052
Day 8: 364072
Day 9: 372354
Day 10: 380907
Root Mean Square Error:
0.7998295186624237
```

Fig. 4. Forecast of Deaths in India for the next 10 days

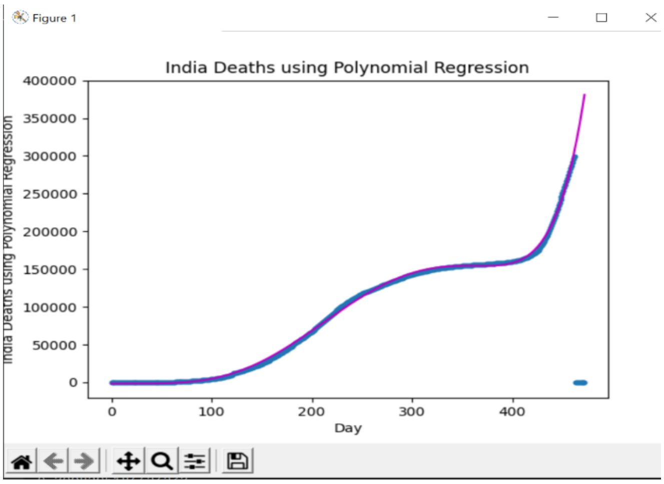
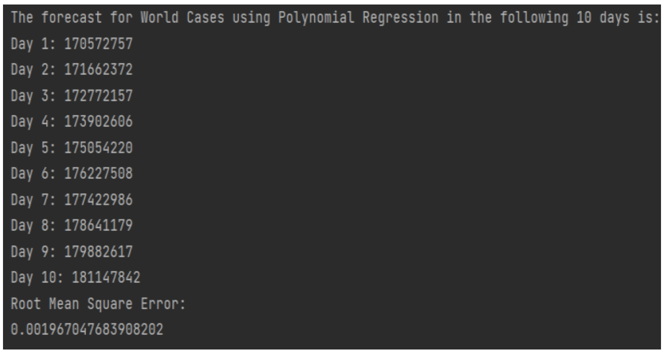


Fig. 5. Graph for Deaths in India for the next 10 days

The results obtained for statistical prediction of active cases using polynomial regression for the next 10 days in world are shown in Fig.6 and the comparison of actual and the prediction are shown in Fig.7.



```
The forecast for World Cases using Polynomial Regression in the following 10 days is:
Day 1: 170572757
Day 2: 171662372
Day 3: 172772157
Day 4: 173902606
Day 5: 175054220
Day 6: 176227508
Day 7: 177422986
Day 8: 178641179
Day 9: 179882617
Day 10: 181147842
Root Mean Square Error:
0.001967047683908202
```

Fig. 6. Forecast of Active Cases Worldwide for the following 10 days

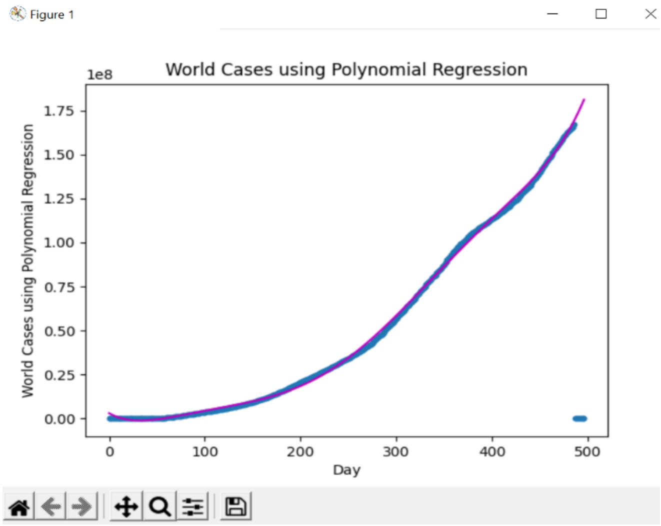
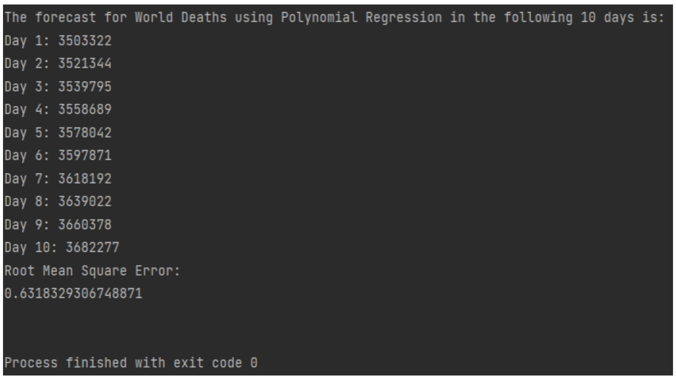


Fig. 7. Graph for Active Cases Worldwide for the following 10 days

The results obtained for statistical prediction of death cases using polynomial regression for the next 10 days in world are shown in Fig.8. and the comparison graph of actual and the prediction are shown in Fig.9.



```
The forecast for World Deaths using Polynomial Regression in the following 10 days is:
Day 1: 3503322
Day 2: 3521344
Day 3: 3539795
Day 4: 3558689
Day 5: 3578042
Day 6: 3597871
Day 7: 3618192
Day 8: 3639022
Day 9: 3660378
Day 10: 3682277
Root Mean Square Error:
0.6318329306748871


Process finished with exit code 0
```

Fig. 8. Forescast of Deaths Worldwide for the following 10 days

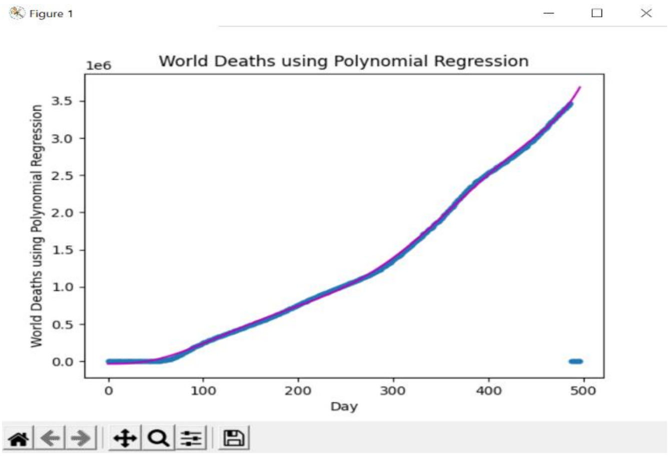
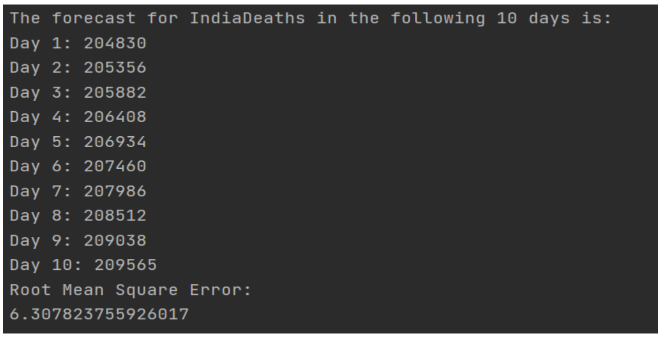


Fig. 9. Graph for Deaths Worldwide for the following 10 days

The results obtained for statistical prediction of death cases using Neural Net model for the next 10 days in India are shown in Fig.10. and the comparison of actual and the prediction are shown in Fig.11.



```
The forecast for IndiaDeaths in the following 10 days is:
Day 1: 204830
Day 2: 205356
Day 3: 205882
Day 4: 206408
Day 5: 206934
Day 6: 207460
Day 7: 207986
Day 8: 208512
Day 9: 209038
Day 10: 209565
Root Mean Square Error:
6.307823755926017
```

Fig. 10. Forecast of India deaths using Neural Net model

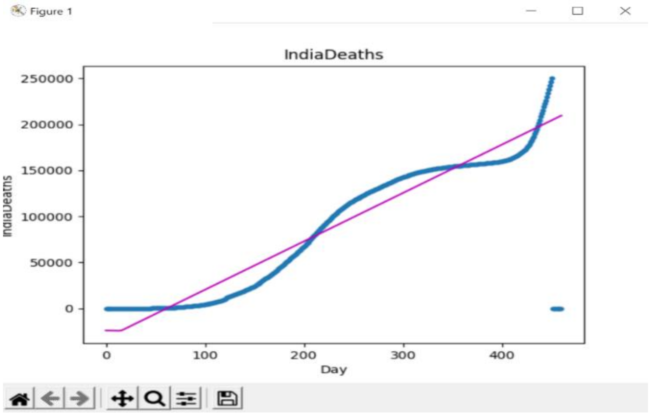
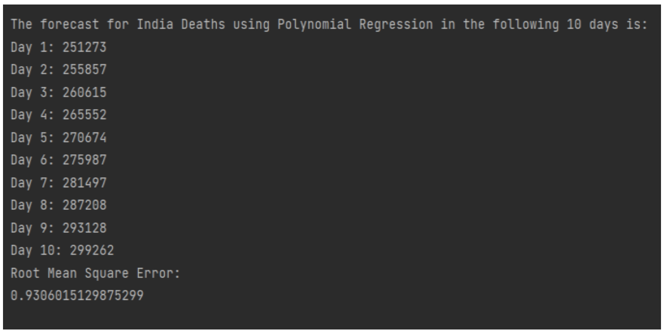


Fig. 11. Graph for India Deaths using Neural Net model

When we compare the results using Neural Net model and that of those using Polynomial Regression model, we can see that the root mean square error value while using Neural Net model is more. Hence, we can say that Polynomial Regression model is the best suited and the more accurate algorithm for this work.



```
The forecast for India Deaths using Polynomial Regression in the following 10 days is:
Day 1: 251273
Day 2: 255857
Day 3: 260615
Day 4: 265552
Day 5: 270674
Day 6: 275987
Day 7: 281497
Day 8: 287208
Day 9: 293128
Day 10: 299262
Root Mean Square Error:
0.9306015129875299
```

Fig. 12. Forecast of India Deaths using Polynomial Regression Model

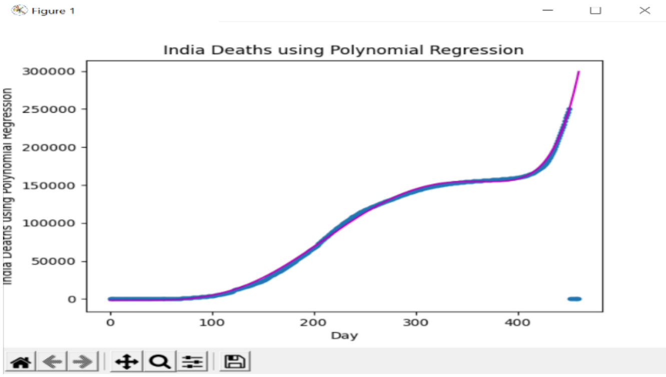


Fig. 13. Graph for India Deaths using Polynomial Regression model

From the Fig. 10 and Fig.12, it can be observed that, when the prediction is done using Neural Net model the root mean square error recorded is 6.30783, but when we do the predictions using Polynomial Regression model, the mean square error is only 0.93060. Hence, we can say that Polynomial Regression model is the best model for COVID-19 prediction.

V. CONCLUSIONS

In order to seek out a better way to predict the COVID-19 pandemic situation, we adopted machine learning methods to process the information and predict the future development of the COVID-19 for the world and India. So as to enhance the accuracy of the prediction results, we have considered cumulative daily data in India and round the world. Then, we employed machine learning algorithms polynomial regression and neural net model, to predict the number of active cases and number of deaths for the upcoming days, From the results its observed that polynomial regression model is best suitable for the prediction of COVID-19. Through the forecast, people can make rational and feasible plans so as to encourage the restoration of the traditional life and serious-affected economies. By knowing the anticipated data, we will caution people around the world and see that the spread of such viruses is prevented at the beginning stages itself. Forecasting the cases and deaths can help us save many lives across the planet. Considering the strong variability of the COVID-19, the spread of the virus has shown different characteristics in different periods. Therefore, future research should take more data related to the COVID-19 into consideration.

REFERENCES

1. *Vinay Kumar Reddy Chimmula, Lei Zhang, "Time series forecasting of COVID-19 transmission in Canada using LSTM networks".Chaos, "Solitons & Fractals", vol. 135, 2020,109864, ISSN 0960-0779.*

2. *Vijander Singh, Ramesh Chandra Poonia, Sandeep Kumar, Pranav Dass, Pankaj Agarwal, Vaibhav Bhatnagar & Linesh Raja, "Prediction of COVID-19 corona virus pandemic based on time series data using support vector machine", Journal of Discrete Mathematical Sciences and Cryptography, 23:8, 1583-1597.*

3. *Chaurasia, V., Pal, S., "Application of machine learning time series analysis for prediction COVID-19 pandemic". Res. Biomed. Eng., 2020.*

4. *Rohit Salgotra, Mostafa Gandomi, Amir H Gandomi, "Time Series Analysis and Forecast of the COVID-19 Pandemic in India using Genetic Programming", Chaos, Solitons & Fractals, vol.138, 2020, 109945, ISSN 0960-0779.*

5. *Zhenyu Li ,Shentong Yang, Junhong Wu ,"The Prediction of the Spread of COVID-19 using Regression Models", International Conference on Public Health and Data Science (ICPHDS). 2020.*

6. *Emrah Gecili ,Assem Ziady, Rhonda D. Szczesniak , "Forecasting COVID-19 confirmed cases, deaths and recoveries: Revisiting established time series modeling through novel applications for the USA and Italy", 2021.*

7. *Singh R.K., Rani M., "Prediction of the COVID-19 Pandemic for the Top 15 Affected Countries: Advanced Autoregressive Integrated Moving Average (ARIMA) Model", JMIR Public Health Surveill, 05 13; 6(2):e19115, 2020.*

8. *Jason Brownlee, "Deep Learning for Time Series Forecasting Predict the Future with MLPs, CNNs and LSTMs in Python".*

9. *Zhao B, Wang Z, Yu Z, Tian C, Cao J., "Time series analysis of the novel coronavirus (COVID-19)", J Immuno Allerg.1(3):1-13, 2020..*