Journal of Artificial Intelligence and Capsule Networks (ISSN: 2582-2012)
www.irojournals.com/aicn/

Polynomial Regression Model to Predict Future Covid Cases

Dr. S. Suriya[1], R. Sanjay Krishna[2]

[1]Associate Professor, [2]PG Scholar

[1,2]Department of Computer Science and Engineering, PSG College of Technology, Coimbatore, India.

Email: [1]suriyas84@gmail.com, [1]ss.cse@psgtech.ac.in, [2]22mz05@psgtech.ac.in

Abstract

Accurate case predictions are essential for efficient public health management and resource allocation since the COVID-19 pandemic has had a substantial impact on economies and global health. Using polynomial regression, a machine learning technique that fits a polynomial function to the data, this research seeks to create a predictive model for future COVID-19 cases. The model takes into consideration the elements such as population density, healthcare facilities, and governmental initiatives using historical COVID-19 case data from India. In order to forecast the number of upcoming COVID-19 instances, the polynomial regression model is employed. The model's effectiveness is assessed using a number of measures, including mean squared error and R-squared. The outcomes demonstrate that the polynomial regression model can precisely forecast the trend of COVID-19 instances over time. This approach can be useful for forecasting the spread of the virus and informing public health policies. The limitations and future directions of the model are also discussed. Furthermore, the model's adaptability to changing trends and its ability to capture non-linear relationships between variables, make it a promising tool for forecasting future pandemics and other public health crises.

Keywords: COVID-19, Regression, Machine Learning, Polynomial regression.

1. Introduction

Millions of people have been affected by the COVID-19 pandemic, which is brought on by the new coronavirus SARS-CoV-2. It has also caused enormous disruptions to economies and healthcare systems. Accurate and timely predictions of future COVID-19 cases are

DOI: https://doi.org/10.36548/jaicn.2023.2.004
Received: 12.04.2023, received in revised form: 12.05.2023, accepted: 27.05.2023, published: 06.06.2023

essential for governments, healthcare professionals, and policymakers to make informed decisions regarding public health measures, resource allocation, and pandemic response strategies. Predictive modelling techniques have been widely employed to forecast the spread of infectious diseases, providing valuable insights for effective planning and management.

Among various predictive modelling techniques, polynomial regression has shown promise in recording non-linear relationship among variables, making it a suitable candidate for predicting future COVID-19 cases. By fitting a polynomial function to the dataset, polynomial regression, a sort of regression analysis, models the connection between a dependent variable and one or more independent variables. This approach allows for greater flexibility in modelling the data, as it can account for non-linear trends and fluctuations that may not be captured by simpler linear or exponential models.

This research aims to develop a robust and accurate polynomial regression model for predicting future COVID-19 cases using historical case data from various countries and regions. The model takes into account factors such as population density, healthcare infrastructure, and government interventions, which are known to influence the spread of the virus. By comparing the performance of the polynomial regression model to other predictive models, its effectiveness in forecasting future COVID-19 cases and its potential applicability to other public health crises are demonstrated.

India, being the second-most populous country in the world, has faced unique challenges in managing the COVID-19 pandemic. The country's diverse population, varying levels of healthcare infrastructure, and regional differences in government interventions have resulted in a complex pattern of COVID-19 spread. For the purpose of developing and validating the polynomial regression model for projecting future COVID-19 instances, this work particularly concentrates on the Indian dataset. The study intends to capture the regional differences and trends in the transmission of the virus by the analysis of historical case data from various states and union territories in India, which would improve the accuracy and reliability of the model.

The polynomial regression model includes elements like population density, urbanization, healthcare facilities, and governmental initiatives like lockdowns and vaccination programs to account for the numerous factors impacting the development of COVID-19 in India. These factors are crucial in understanding the dynamics of the pandemic in India and

will help improve the model's predictive capabilities. Additionally, the impact of socio-economic factors, such as income levels and education, which may indirectly affect the spread of the virus by influencing people's adherence to public health guidelines and their access to healthcare services, are also considered.

By applying the polynomial regression model to the Indian dataset, valuable insights are provided into the future trajectory of COVID-19 cases in the country. These predictions can help policyholders and healthcare professionals regarding resource allocation, public health measures, and vaccination strategies. Furthermore, the model's adaptability to changing trends and its ability to capture non-linear relationships between variables make it a promising tool for forecasting future pandemics and other public health crises in India and beyond. By continuously updating the model with new data and refining its parameters, it can be ensured that the predictions remain accurate and relevant, ultimately contributing to more effective public health management and pandemic response strategies.

2. Literature Review

Machine learning has grown in importance in the recent years with applications in many industries, including marketing, finance, and healthcare. In the work "Machine learning algorithms-a review," Mahesh B offered a thorough analysis of several machine learning algorithm types. The major types of machine learning algorithms, unsupervised learning, supervised learning, and reinforcement learning were introduced and various algorithms under each category, along with their advantages and disadvantages were discussed. The author also highlighted some real-world applications of these algorithms [1].

In "Identifying child abuse through text mining and machine learning," Alharbi et al., proposed a novel approach to automatically detect potential cases of child abuse by analysing written documents. The authors argue that traditional methods for identifying child abuse rely on human experts, which can be time-consuming and error-prone. The work demonstrated the potential of combining text mining and machine learning techniques to automatically identify potential cases of child abuse, achieving a high level of accuracy [2].

"Reinforcement learning-based mobile robot navigation" proposed a model that combines reinforcement learning algorithms with mobile robot platform to improve mobile robot navigation. The paper demonstrated the potential of using reinforcement learning

techniques for mobile robot navigation, achieving a high level of accuracy in different environments [3].

“Reinforcement learning for multi-product multi-node inventory management in supply chains" proposed a new approach to logistics management in supply chain using reinforcement learning techniques. The authors demonstrated the potential of using reinforcement algorithm to optimize logistics management decisions in complex supply chain environments [4]. The paper "Evaluation of machine learning technique for face detection and recognition" evaluated the performance of machine learning algorithms for face recognition tasks. The authors demonstrated the potential of using SVM and RF algorithms for these tasks [5].

“Supervised machine learning techniques to the prediction of tunnel boring machine penetration rate" used supervised machine learning technique to find TBM penetration rate using a set of input variables. The authors demonstrated the potential of using supervised learning to find TBM penetration rates [6]. The paper “Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford Exercise Testing (FIT) project" used supervised machine learning technique for predicting diabetes mellitus with a set of input variables. The authors demonstrated the potential of using supervised learning and SMOTE technique to find the imbalance of class in the dataset and predict diabetes mellitus [7].

The authors of "Implementation of machine learning model to predict heart failure disease” used a dataset of patient records from the Cleveland Clinic Foundation. The dataset contains several clinical and demographic variables. The model achieved an accuracy of 85.5%, demonstrating the potential of using machine learning algorithms to find heart disease [8]. The authors of "Music recommendation system using machine learning" used a dataset of user listening histories from the Million Song Dataset to train and test the music recommendation system. The dataset contains information on user listening histories, such as the songs they listened to and the time they spent listening to each song. The system achieved a precision of 0.25 and a recall of 0.15, demonstrating the potential of using machine learning algorithms to recommend music to users [9].

The authors of "Sales prediction for big mart outlets” used a dataset of sales records from Big Mart outlets. The dataset contains information on outlet locations, product types, and sales volumes. The model achieved an RMSE of 1158.8, demonstrating the potential of using

machine learning algorithms to predict sales for retail outlets [10]. Das et al. objected to find community death risk in COVID-19 patients and to develop an online prognostic tool based on their findings. The study included data from 3,897 COVID-19 patients in India and used various clinical and demographic variables. The authors found that their models were able to accurately predict. mortality risk for COVID 19 patients, with an area under curve (AUC) of 0.93 [11].

The paper titled "Machine learning prediction for mortality of patients diagnosed with COVID-19: a nationwide Korean cohort study" aimed to explore the use of machine learning techniques for finding the mortality rate of COVID 19 patients in Korea. The authors used a dataset of 10,062 COVID-19 patients who were diagnosed between January 20 and April 14, 2020, in Korea. The study provided insights into the factors that contribute to COVID-19 mortality and highlights the potential of machine learning in improving patient outcomes [12]. The paper titled "COVID Mortality Prediction with Machine Learning Methods: A Systematic Review and Critical Appraisal" is a systematic review that aims to evaluate the effectiveness of machine learning models in predicting COVID 19 death rates. The authors searched various databases and identified 43 studies that met their inclusion criteria. The authors found that the most commonly used machine learning algorithms in these studies were logistic regression, random forest, and support vector machine [13].

The paper "Polynomial Regression Model for COVID-19 Pandemic Prediction" presented a polynomial regression model to predict the number of COVID-19 cases. The authors used data from the WHO and Johns Hopkins University to develop the model. The study found that a 5^{th} degree polynomial regression model provided the best fit for the data, with a high coefficient of determination (R-squared value). The model was able to find the count of cases with reasonable accuracy, demonstrating the potential of polynomial regression for pandemic prediction [14].

The paper "COVID-19 Prediction for Egypt using Polynomial Regression Model" focused on predicting the count of COVID-19 cases in Egypt using a polynomial regression model. The authors used data from the Egyptian Health Ministry and the WHO. The results showed that a 4^{th} degree polynomial regression model provided the best fit for the data, with a high R-squared value. The framework was able to find the number of cases with good accuracy, suggesting that polynomial regression could be a useful method for finding COVID 19 cases in Egypt [15].

3. Proposed Methodology

Polynomial regression is defined as a type of analysis of regression in which the relation between the dependent variable y and the independent variable x is modelled as the degree of n^{th} polynomial.

The formula for polynomial regression is:

$$y = a0 + a1*x + a2*x^2 + ... + an*x^n$$

where x is independent variable, y is dependent variable, a0, a1, a2, ..., an are coefficients, and n is degree of polynomial. The degree of the polynomial determines the shape of best fit curve of the data.

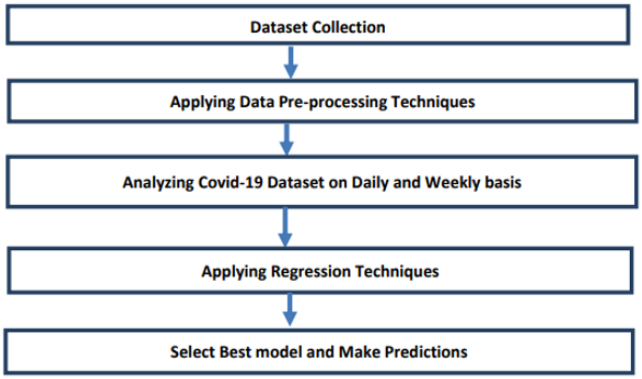

Figure 1. Methodology

Steps Involved in Polynomial Regression

Step-1: The first step in polynomial regression is to gather data. The data must be collected on the dependent variable and the independent variable. The data should be in numerical format and should be organized in a table or spreadsheet.

Step-2: The next step is to visualize the data using a scatter plot. This will help to identify any patterns or relationships between the variables. If the data appears to have a curved pattern, a polynomial regression may be a good fit.

Step-3: The polynomial degree determines the curvature of the line of best fit. A high degree polynomial will fit in the data more closely, but overfitting may occur and lead to bad predictions for new values. The polynomial degree may be chosen based on the visual inspection of the data or by using a method such as cross-validation.

Step-4: Once the degree of the polynomial is chosen, the polynomial function can be fit to the data using a regression model. This involves calculating the coefficients of the polynomial function that best fits the data. The regression model will give the formula for the polynomial function.

Step-5: To evaluate the model, the R-squared value, which is a measure of how well the polynomial function fit in the data, is calculated. The polynomial function can be plotted on the scatter plot to visualize how well it fits the data.

Step-6: Once the polynomial function is fitted to the data, it can be used to make predictions for new data points which is shown in figure 1.

Manual Tracing with Respect to a Small Sample

Considering the corona cases of India from March 7 to March 16, 2023 as shown in Table 1, they are manually traced using polynomial regression.

Table 1. Corona Cases of India from March 7, 2023, and March 16, 2023.

Date (X-axis)	**Count of Cases (Y-axis)**
March 7	326
March 8	379
March 9	440
March 10	456
March 11	524
March 12	444
March 13	402
March 12	618
March 13	402
March 14	618
March 15	754
March 16	796

To calculate the coefficients of the quadratic polynomial manually, the following steps are used:

1. Convert the date into a numeric value, as shown in Table2, by subtracting March 7, 2023, from each date. For example, March 7 is day 0, March 8 is day 1, March 9 is day 2, and so on.

Table 2. Numeric value

Date	**Numeric Value (x)**
March 7	0
March 8	1
March 9	2
March 10	3
March 11	4
March 12	5
March 13	6
March 12	5
March 13	6
March 14	7
March 15	8
March 16	9

2. Create a table of values for x, x^2, and y, where y is the count of cases as shown in table 3.

Table 3. Values for x, x^2, and y

x	**x^2**	**y**
0	0	326
1	1	379
2	4	440
3	9	456
4	16	524
5	25	444
6	36	402
5	25	618
6	36	402
7	49	618
8	64	754
9	81	796

3. Calculate sum of x, x^2, y, xy, x^2y.

Sum (x) =45

Sum (x^2) =285

Sum (y) =5139

Sum (xy) = 36259

Sum (x^2y) = 295043

4. Use the following formulas to calculate the coefficients of the quadratic polynomial:

ao =(sum(y)sum(x^2)-sum(x)sum(xy)/(nsum(x^2)- sum(x)^2)

a1 = (nsum(xy) - sum(x)sum(y)) / (nsum(x^2) - sum(x)^2)

a2 = (nsum(x^2y) - sum(x*y) sum(x)) / (nsum(x^2) - sum(x)^2)

where n is the count of data, which is 10 in this case.

a0 = -113.8727273

a1 = 97.38484848

a2 = -0.1857142857

y = a0 + a1x + a2x^2

To calculate the predicted number of cases for March 17th, the formula used is:

y = -113.8727273 + 97.38484848x - 0.1857142857x^2

where x is the number of days since March 7th, and y is the predicted number of cases.

Since March 17th is 10 days after March 7th, x = 10 is set.

y = -113.8727273 + 97.38484848*10 - 0.1857142857*10^2

By simplifying this expression,

y = -113.8727273 + 973.8484848 - 18.57142857

y = 841.4043292

Therefore, the predicted number of cases for March 17th is 841.4043292. Officially it was found to be 843.

4. Dataset Description

The research data was obtained from https://ourworldindata.org/coronavirus. The attributes present in the dataset are:

1. State/Union Territory Glucose
2. ConfirmedIndianNational
3. ConfirmedForeignNational
4. Cured
5. Deaths
6. Total Confirmed

The description of the dataset are as follows.

- State/Union Territory: State/Union territory of India helps to find which part of India is more affected by Covid.
- Confirmed Indian National: Total confirmed cases within the nation (Numerical)
- Confirmed Foreign National: Total confirmed cases in the international airport (Numerical)
- Cured: Total cured cases (Numerical)
- Deaths: Total deaths due to Covid (Numerical)
- Total Confirmed: Total confirmed cases within the nation as well as airport (Numerical)

Table 4. Dataset

State/Unic	Confirmed	Confirmed	Cured	Deaths	Confirmed
Kerala	1	0	0	0	1
Kerala	1	0	0	0	1
Kerala	2	0	0	0	2
Kerala	3	0	0	0	3
Kerala	3	0	0	0	3
Kerala	3	0	0	0	3
Kerala	3	0	0	0	3
Kerala	3	0	0	0	3
Kerala	3	0	0	0	3
Kerala	3	0	0	0	3
Kerala	3	0	0	0	3
Kerala	3	0	0	0	3
Kerala	3	0	0	0	3
Kerala	3	0	0	0	3
Kerala	3	0	0	0	3
Kerala	3	0	0	0	3
Kerala	3	0	0	0	3
Kerala	3	0	0	0	3
Kerala	3	0	0	0	3
Kerala	3	0	0	0	3
Kerala	3	0	0	0	3
Kerala	3	0	0	0	3
Kerala	3	0	0	0	3
Kerala	3	0	0	0	3

5. Experimental Results and Discussion

A) Python

Python, a versatile programming language, offers a wide range of libraries and tools for implementing machine learning algorithms, including polynomial regression. In this study, libraries such as Pandas, NumPy, and scikitlearn are employed to pre-process data, develop the polynomial regression model, and assess its performance. The results achieved for Covid dataset is shown below in Figure 2.

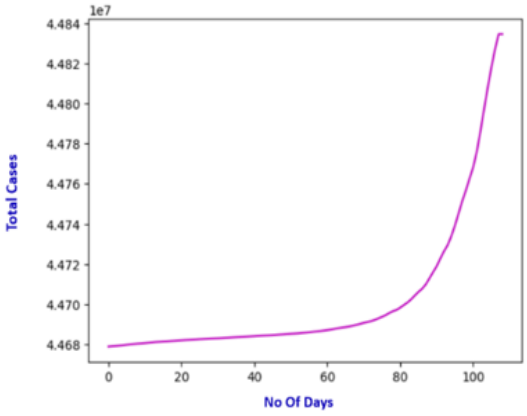

Figure 2. Prediction of Actual Cases

```
from sklearn.preprocessing import PolynomialFeatures
from sklearn import linear_model
polyFeat = PolynomialFeatures(degree=3)
x= polyFeat.fit_transform(x)
```

```
model=linear_model.LinearRegression()
model.fit(x,y)
accuracy=model.score(x,y)
print(f'Accuracy:{round(accuracy*100,3)}%')
```

```
Accuracy:96.644%
```

Figure 3. Model Fitting and Accuracy

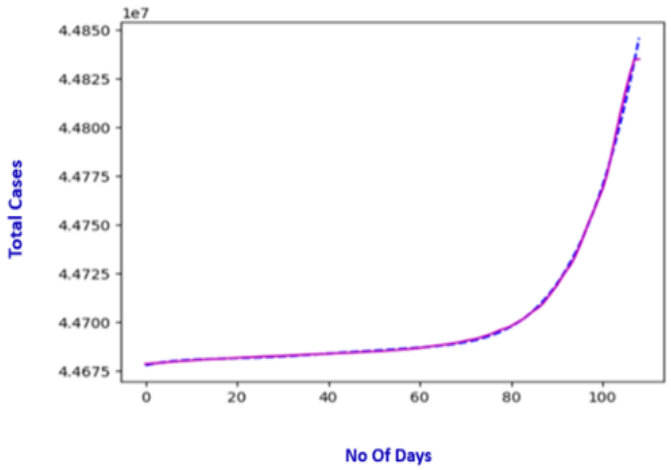

Figure 4. Best fit line for predicted cases along with actual cases

```
days=5
print(f'prediction - Cases after {days} days:',end='')
print(round(int(model.predict(polyFeat.fit_transform([[111+days]])))/1000000,2),'Million')
```

```
prediction - Cases after 5 days:44.89 Million
```

```
days=30
print(f'prediction - Cases after {days} days:',end='')
print(round(int(model.predict(polyFeat.fit_transform([[111+days]])))/1000000,2),'Million')
```

```
prediction - Cases after 30 days:45.19 Million
```

Figure 5. Prediction of cases after 5 days and 30 days

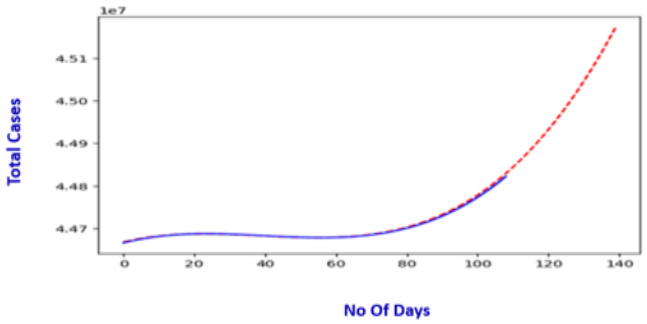

Figure 6. Prediction after 30 days

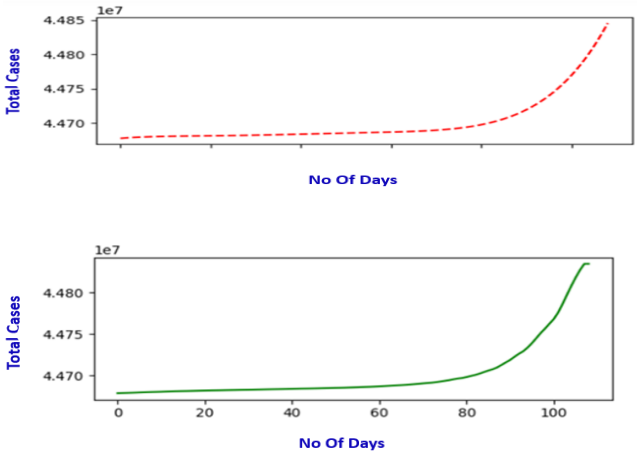

Figure 7. Plot Visualization using Subplots

B. Discussion

The polynomial regression model employed in this research demonstrated success in predicting cases of COVID-19 for the next 5 and 30 days, achieving a high accuracy of 96.644% as shown in Figures 3, 4, 5, 6 and 7. The model's predictions indicate that the observed increase in cases is not representative of a 4th wave but is instead a temporary spike. This conclusion is corroborated by the visualization of actual and predicted cases, which exhibit a similar trend. Nonetheless, it is crucial to acknowledge the model's limitations and the data used. Polynomial regression models can be sensitive to outliers and may not generalize well to novel data.

6. Conclusion and Future Works

The current polynomial regression model shows promising results in predicting COVID-19 cases. The polynomial regression model employed in this study has demonstrated promising results in predicting COVID-19 cases and identifying the current increase as a spike rather than a new wave. However, it is essential to continuously update the model with new data and consider other factors that may influence the virus's spread. For future work, it is recommended to explore other machine learning algorithms, incorporate additional features (e.g., vaccination rates, government policies), and continuously update the model with new data to maintain its accuracy and relevance. Regular evaluation and comparison of different models will help in refining the prediction capabilities and contribute to better decision-making in managing the pandemic.

References

[1] Mahesh, B., 2020. "Machine learning algorithms-a review". International Journal of Science and Research (IJSR), 9, pp.381-386.

[2] Amrit, C., Paauw, T., Aly, R. and Lavric, M., 2017. "Identifying child abuse through text mining and machine learning". Expert systems with applications, 88, pp.402-418.

[3] ALTUNTAŞ, N., Imal, E., Emanet, N. and Öztürk, C.N., 2016. "Reinforcement learning-based mobile robot navigation". Turkish Journal of Electrical Engineering and Computer Sciences, 24(3), pp.1747-1767.

[4] Sultana, N.N., Meisheri, H., Baniwal, V., Nath, S., Ravindran, B. and Khadilkar, H., 2020. "Reinforcement learning for multi-product multi-node inventory management in supply chains". arXiv preprint arXiv:2006.04037.

[5] Amaro, E.G., Nuño-Maganda, M.A. and Morales-Sandoval, M., 2012, February. "Evaluation of machine learning techniques for face detection and recognition". In CONIELECOMP 2012, 22nd International Conference on Electrical Communications and Computers (pp. 213-218). IEEE.

[6] Xu, H., Zhou, J., G. Asteris, P., Jahed Armaghani, D. and Tahir, M.M., 2019. "Supervised machine learning techniques to the prediction of tunnel boring machine penetration rate". Applied sciences, 9(18), p.3715.

[7] Alghamdi, M., Al-Mallah, M., Keteyian, S., Brawner, C., Ehrman, J. and Sakr, S., 2017. "Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford ExercIse Testing (FIT) project". PloS one, 12(7), p.e0179805.

[8] Alotaibi, F.S., 2019. "Implementation of machine learning model to predict heart failure disease". International Journal of Advanced Computer Science and Applications, 10(6).

[9] Verma, V., Marathe, N., Sanghavi, P. and Nitnaware, P., 2021. "Music recommendation system using machine learning". International Journal of Scientific Research in Computer Science, Engineering and Information Technology.

[10] Jyothi, J., Madhuri, U.C., Srija, B.O.S., Sravani, J. and Mounica, P., "Sales prediction for Big Mart outlets".

[11] Das AK, Mishra S, Saraswathy Gopalan S. "Predicting CoVID-19 community mortality risk using machine learning and development of an online prognostic tool". PeerJ. 2020 Sep 28;8:e10083. doi: 10.7717/peerj.10083. PMID: 33062451; PMCID: PMC7528809.

[12] An C, Lim H, Kim DW, Chang JH, Choi YJ, Kim SW. "Machine learning prediction for mortality of patients diagnosed with COVID-19: a nationwide Korean cohort study". Sci Rep. 2020 Oct 30;10(1):18716. doi: 10.1038/s41598-020-75767-2. PMID: 33127965; PMCID: PMC7599238.

[13] Bottino F, Tagliente E, Pasquini L, Napoli AD, Lucignani M, Figà-Talamanca L, Napolitano A. "COVID Mortality Prediction with Machine Learning Methods: A Systematic Review and Critical Appraisal". J Pers Med. 2021 Sep 7;11(9):893. doi: 10.3390/jpm11090893. PMID: 34575670; PMCID: PMC8467935.

[14] Depuru, S. S. S. R., Wang, L., & Wang, J. (2020). "Polynomial Regression Model for COVID-19 Pandemic Prediction" 8(5), p.681.

[15] Elaziz, M. A., Ewees, A. A., & Atangana, A. (2020). "COVID-19 Prediction for Egypt using Polynomial Regression Model". Chaos, Solitons & Fractals, 140, p.110122.