International Journal of Future Generation Communication and Networking
Vol. 13, No. 4, (2020), pp. 2500–24506

Regression Model Based Prediction on COVID-19 Confirmed Cases, Recoveries and Deaths in India

Narayana Darapaneni*, Anwesh Reddy Paduri*, Sourabh Verma[1], Ankit Namdeo[2], Siddhartha Khushu[3], S. Sainath[4], Tarun Kr. Mittal[5]

Great Learning, Gurugram

[*]darapaneni@gmail.com, [*]anwesh@greatlearning.in, [1]chatsourabh@gmail.com, [2]ankitnamdeo285@gmail.com, [3]snathsai@gmail.com, tarunkar@gmail.com, siddhartha_k07@yahoo.com,

Abstract

To leverage machine learning techniques to effectively develop a regression model-based prediction of COVID-19 confirmed cases, recoveries and deaths in India. This would help to be better prepared for upcoming cases thereby mitigating the impact risk. As part of this study, we have taken input data on COVID-19 in India (for data cleaning) from the following source: https://www.mohfw.gov.in website and application programming interface (API) provided by https://www.COVID19india.org/ [2]. The time period of data is from 30th Jan 2020 to 30th Jul 2020. The data includes cumulative confirmed cases, deaths and recoveries in India. Considering the nature and scope, supervised learning methods and algorithms were used namely- Polynomial regression and Support Vector Regression (SVR). The regression accuracies of the Polynomial Regression model used for L_Poly_reg_C, L_Poly_reg_R, and L_Poly_reg_D are 98.97%, 98.83% and 97.21% respectively for test data whereas the regression accuracies of the SVR model used for SVM_reg_C, SVM_reg_R and SVM_reg_D are 94.79%, 97.01% and 94.64% respectively for test data. Models trained on only the COVID-19 dataset perform poorly on test data. Also, the Root Mean Square Error (RMSE) values of the Polynomial Regression model used for L_Poly_reg_C, L_Poly_reg_R, and L_Poly_reg_D are 39170, 28263 and 1178 respectively whereas the RMSE values of the SVR model used for SVM__reg_C, SVM_reg_R, and SVM_reg_D are 82790, 42354 and 1634 respectively. Both Polynomial regression & SVR can be used for predicting the COVID-19 Confirmed, Recovery & Death cases for India. But based on the Training / Test Model Accuracies and RMSE values deduced, it seems that Polynomial regression model is slightly better in performance than SVR model for predicting the above data.

Keywords*: COVID-19, polynomial regression, support vector regression, machine learning, prediction, supervised learning, RMSE[1]*

1. Introduction

Year 2020 has unfortunately been a year of COVID-19 so far. The pandemic of this scale and severity was unforeseen and perhaps the worst since Spanish flu in 1918. The catastrophic impact it had made to mankind, economy and overall well-being in such a short time is unimaginable. It has affected nearly 10 million people worldwide, and in India as on 29th June, 2020, there are 528,859 confirmed cases with 16,095 deaths.

COVID-19 is caused by a severe acute respiratory syndrome (SARS) corona virus 2 (SARS-CoV-2) with its origin in China in later part of 2019. Unlike bacteria and fungi, which produce distinct unpredictable

# All the authors have equally contributed.
*Corresponding author, Guide.

ISSN: 2233-7857 IJFGCN

metabolic signatures associated with inherent differences in both primary and secondary metabolic processes, viruses are wholly reliant on the metabolic structure of infected cells for duplication and transmission [7].

As per the World Health Organization (WHO) reports, the pandemic situation is categorized into four stages. The first stage begins with the infected cases reported for the people who travelled in already affected regions, whereas in the second stage, infected cases are reported locally among family, friends and others who came into contact with the person arriving from the affected regions. At this point the affected people are traceable. Later, the third stage makes the situation even worse as the transmission source becomes untraceable and spreads across the individuals who neither have any travel history nor met the infected person. This situation demands immediate lockdown across the nation to reduce the social contacts among individuals and control the rate of transmission. The worst of all, stage four being when the transmission becomes pandemic and uncontrollable [8].

Considering that it is highly infectious and world is yet to find a vaccine for it, it's more important to take a preventive and precautionary approach and try measures to curb its spread. Accordingly, the purpose of this study is to leverage machine learning techniques to effectively predict the confirmed cases, recoveries and fatalities. This would help to be better prepared for upcoming cases and mitigate risk of spread including scaling up the health infrastructure wherever needed. As the resources are limited and the government doesn't have previous experience to handle pandemic of this magnitude, it will help to prioritize and direct the resources where and when the probability of occurrence is higher.

We truly believe that partnering of statistical and machine learning science with medicinal science is the need of the hour. While many studies have/are being done in this area, there are still many inconsistencies of actual vs. predicted COVID-19 cases across the world.

There are mainly two key areas we think where focus needs to be there to tackle COVID-19. First area is around finding the medicinal cure for this pandemic. This includes deep research around COVID-19 virus classification and origin, how it transmits, various risk factors, its pathogenesis and immune response, clinical manifestations, diagnosis, study of lab and radiology findings and various treatment measures. These measures mainly include medicinal interventions like inventing a vaccine, antibiotics and ventilator support for severe cases.

Above is indeed a primary measure to cure and ensure recoveries and reduce fatalities. This also includes creating precautionary guidelines for prevention- hand sanitization, masks, immunity boosters, social distancing, etc. The active scrutiny of close contacts of confirmed COVID-19 cases and the execution of control measures, including home quarantine for those evaluated at moderate/high risk of exposure, reduce the risk of human-to-human transmission originating from imported cases and subsequently delay spread of the virus in the general population [6].

However, it's equally important to complement these with more mitigated measures which may not necessary be entirely biologically oriented. Being a pandemic, it is crucial to predict its outbreak in each of the geography (clusters, local districts, zones of few kilometers etc.). It is necessary that data are accurately circulated to citizens and decision makers to make informed choices [9].

In other words, same reactive and preventive control measures for same duration across PAN India are not advisable. This is attributed to the fact that India is having one of the highest population densities but also contrasting demographics including climate, ethnicity, religious practices, cultural beliefs, hygiene awareness, income disparities, local state enforcements and health infrastructure availability. Hence, it is imperative to approach the preventive measures with sound facts. This needs to be backed by robust live COVID-19 data by leveraging scientific machine learning models and get closer to correctly predicting the outbreak. Well, when we say predicting the outbreak, it actually sums up multiple indicators in this context like:

- Where i.e. which are the most prone areas to witness COVID-19 outbreak in coming period (days, weeks, months)?
- When i.e. likelihood of outbreak in those areas (relative probability)?
- How i.e. velocity of pandemic spread in those areas?
- Who i.e. identification of high risk factors for individual- age, co-morbidities, etc.?

ISSN: 2233-7857 IJFGCN

- What i.e. the probable prediction and break-up of infected, cured and fatalities?

Addressing all these questions requires a comprehensive framework encompassing multiple data sources, machine learning techniques (both regression & classification) and various resources. Considering the objective of our project, we would like to restrict our scope to regression model with focus on predicting the COVID-19 cases, recoveries and fatalities in India.

We selected this problem statement as we believe that this is the most burning and relevant question at the moment around COVID-19. Secondly, many studies have been done around this topic meaning we can develop good understanding and appreciate the problem post literature reviews. Thirdly, we believe that being a recent and live global issue, there is still much scope to value-add by developing better COVID-19 prediction model. Last but not the least, we can further fine tune the algorithms on the basis of latest data pulled at regular intervals, to improve model accuracy.

2. Materials and Methods

According to the World Health Organization (WHO) COVID-19 situation report [1] as on 29th June, 2020, a total of 9,843,073 confirmed cases and 495,760 deaths have been reported across the globe. A total of 528,859 confirmed cases and 16,095 deaths have been reported across India till 29th June, 2020. Worldwide increase of this pandemic has been very rapid.

The literature search was conducted on 29th June, 2020 by using the following database: https://www.kaggle.com/imdevskp/COVID19-corona-virus-india-dataset [10]. It included studies published in 2019 and 2020. In this study, outbreak of this disease has been analyzed for India till 29th June, 2020 and predictions have been made for the number of cases for the next 2 weeks. Regression model have been used for predictions based on the data collected from Ministry of Health and Family Welfare (MoHFW) Government of India repository in the time period of 22nd January, 2020 to 29th June, 2020. MoHFW updates state level data on confirmed cases, deaths and recoveries. The objectives of above study were as follows:

- Finding the rate of spread of the disease in India.
- Prediction of COVID-19 outbreak using Regression models.

We also reviewed another literature by Ramjeet Singh Yadav on the topic: Data analysis of COVID-2019 epidemic using machine learning methods: a case study of India https://link.springer.com/content/pdf/10.1007/s41870-020-00484-y.pdf.

The literature search was conducted from 1st March to 11th April, 2020. In this study, six regression analysis based models were utilized namely: quadratic, third degree, fourth degree, fifth degree, sixth degree, and exponential polynomial respectively. Root mean square error (RMSE) of these six regression analysis models was also determined. It was observed that the best fit line was obtained using sixth degree polynomial where it RMSE was very less compared to other models over 7 days of forecast period [5]. The objectives of above study were as follows:

- Finding the rate of spread of the disease in the next 7 days with the help of regression analysis models.
- To develop a machine learning-based regression analysis models for exposed COVID-2019.

As explained in the introduction, the objective of this study is to develop a regression model based prediction on COVID-19 cases, recoveries and fatalities in India. Following models have been used for the same:

2.1 Regression Model

Regression models are statistical sets of processes which are used to estimate or predict the target or dependent variable on the basis of independent variables. The regression model has many variants like linear regression, ridge regression, stepwise regression, polynomial regression etc. This study has used polynomial regression and support vector regression (SVR) for prediction of COVID-19 cases.

The polynomial regression used in the study includes transformation of data into polynomials and applying linear regression to fit the parameter. Choosing the value of a degree is a tricky task. If the degree of

ISSN: 2233-7857 IJFGCN

polynomial is less, it will not be able to fit the model properly and if the value of degree of polynomial is greater than actual, it will over-fit the training data.
One of the major advantages of SVR is that its computational complexity does not depend on the dimensionality of the input space. Furthermore, it has outstanding generalization potential, with high prediction accuracy. [4]

2.2 Performance metrics

A performance metric can be defined as a logical and mathematical construct designed to measure how close are the actual results from what has been expected or predicted. The most commonly mentioned metrics in research studies are Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), etc.

2.3 Measuring the Quality of Fit

We need to quantify the extent to which the predicted response value for a given observation is close to the true response value for that observation. In the regression setting, the most commonly-used measure is the mean squared error (MSE), given by:

$$\text{MSE} = 1/n \sum_{i=1}^{n} (y_i - f(x_i))^2 \quad (1)$$

Where, f(xi) is the prediction that f gives for the ith observation. The MSE will be small if the predicted responses are very close to the true responses, and will be large if for some of the observations, the predicted and true responses differ substantially. The MSE in (1) is computed using the training data that was used to fit the model, and so should more accurately be referred to as the training MSE. But in general, it is of lesser significance to how well the method works on the training data. Rather, we are more interested in the accuracy of the predictions that we obtain, when we apply our method to previously unseen test data. [3]
We want to choose the method that gives the lowest test MSE, as opposed to the lowest training MSE. Root mean square error (RMSE) is simply the square root of MSE and one of the important metrics to evaluate model performance. As a part of this study, a standard dataset from Kaggle website has been used:
https://www.kaggle.com/sudalairajkumar/COVID19-in-india?select=COVID_19_india.csv [11]
Accessed for latest update for data as on July 30, 2020 and this Kaggle website data has been sourced from the official website of MoHFW (Ministry of Health & Family Welfare), Government of India.
For evaluation of the two regression models, a data-frame was used for fitting Polynomial & SVR algorithms to the dataset. For analysis using regression model for COVID-19 confirmed cases prediction, we split the dataset from data-frame into two different sets, training set (X_train, yC_train) and the testing Set (X_test, yC_test) in a specific ratio. After doing the above, we fitted Polynomial Regression as well as SVR algorithms to the datasets created in order to predict (for a 20-day rolling period window in future) the number of COVID-19 Confirmed cases for India. All the above steps were executed in a similar manner using 2 more datasets relevant for predicting the COVID-19 recovery and death cases in India.

3. Results and Discussion

The regression models using Polynomial Regression and SVR algorithms were created for predicting the total no. of COVID-19 confirmed cases, recoveries and deaths respectively. The name of the models created and their performance metrics (viz. Accuracy Score & RMSE values) after fitment of training and test dataset and running them in order to finally predict the total no. of COVID-19 Confirmed cases, recoveries and deaths respectively are given in the Table 1 below.
From the below performance metrics values, we can see that training accuracy scores are higher than testing accuracy scores in all the below regression models used, but the drop in testing accuracy scores in case of SVR model is of a higher order than Polynomial Regression models which also means that using the SVR algorithm is making the models to be more over fit than Polynomial Regression algorithm. Also, the corresponding RMSE values in case of Polynomial Regression models are appreciably lower than the SVR models used for prediction of total no. of COVID-19 confirmed cases, recoveries and deaths, which means the Polynomial Regression algorithm is better in predicting these numbers.

ISSN: 2233-7857 IJFGCN

This is also getting broadly proven as the predicted total no. of COVID-19 confirmed and death cases (for a 20-day rolling period window in future) using Polynomial Regression are comparatively closer (than SVR) to actual total number of COVID-19 confirmed and death cases published for the period from 30-Jan-2020 till 01-Aug-2020 (Source: Times of India website [12]) as shown in Table 2 below:

Table 1: Accuracy scores of models

S.N	ID-19 Predict	el used	el Name	racy (in %) ning		E
1	irmed cases	nomial Regres	ly_reg_C			70
2			_reg_C			00
3	very cases	nomial Regres	ly_reg_R			53
4			_reg_R			54
5	n cases	nomial Regres	ly_reg_D			
6			_reg_D			

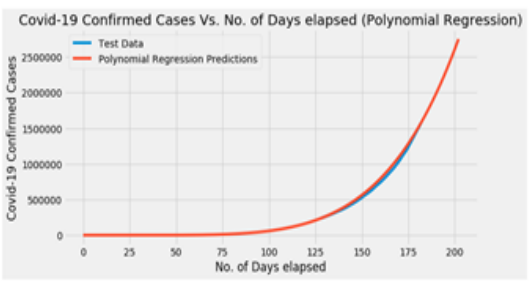
Table 2: Predicted vs. Actual Cases

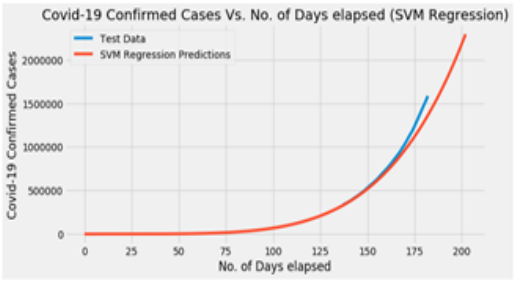
VID-19 Confirmed Cases-1			VID-19 Recovery Cases-1-			VID-19 Death Cases-1-		
Predicted		al	Predicted		al	Predicted		al
nomial Regr			nomial Reg			nomial Reg		
3,250	7,851	,723	536	005	,629	7	0	4

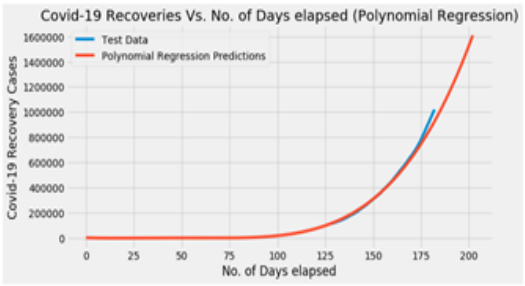
Below, we are displaying the graphical representation of the comparison of predictions of models for test data vs. model prediction for confirmed cases, recoveries and deaths, respectively. In the graphs shown below, X-axis is depicting the number of days elapsed (since the first reported COVID-19 case on 30-Jan-2020) and Y-axis is depicting the number of confirmed cases, recoveries and deaths for Polynomial and SVR models respectively.

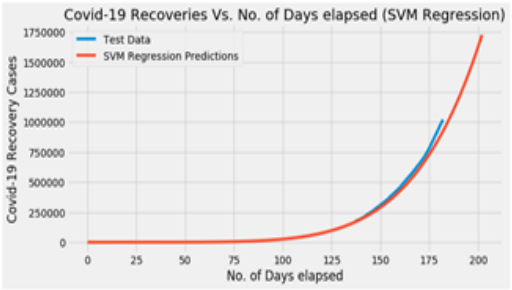
Prediction Comparisons

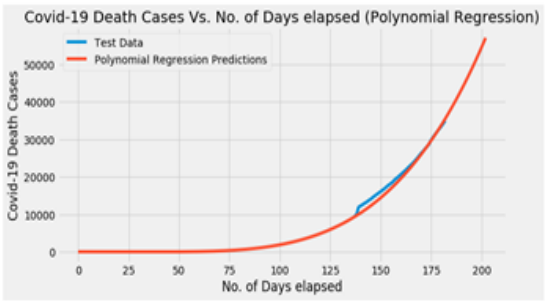
ISSN: 2233-7857 IJFGCN

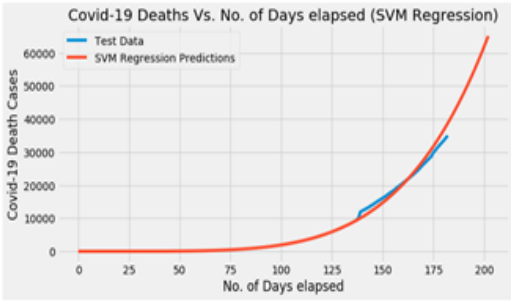












4. Conclusion

We can say that both the Polynomial regression and SVR can be used for predicting the COVID-19 Confirmed, Recovered & Death cases for India, but based on the Training / Test Model Accuracies and the RMSE (Root Mean Square Error) parameter values deduced between the above two models, it seems that Polynomial Regression model is slightly better in performance than Support Vector Regression (SVR) model for predicting the above data.

Since the SVR model is evolved from the perception algorithm, hence the drop in testing accuracy scores is also explainable. Beyond this the model accuracies for SVR algorithms are also very much dependent upon the hyper-parameter tuning (for e.g.: setting the values of SVM Parameters like kernel, epsilon, gamma, degree and cost function).

In case of the Polynomial Regression model, better results were obtained by using the degree four & five in the algorithm. We also found that the model accuracies for both Polynomial regression and SVR) also depend upon the split ratio of Training & Testing data sets.

The above regression models are very clearly predicting that the no. of COVID-19 Confirmed cases, recoveries and deaths are going to increase at a steeper rate with time. This study can also be useful in helping the Government of India to plan out the availability of important resources to better handle it in tackling the spread of COVID-19 w.r.t. prevention & treatment facilitation including the following:

- Ensuring regular production and supply of face masks, hand gloves and sanitizers at affordable prices to urban and rural population across the country, including educating them through social media platform or with help from social workers on how to use these meticulously for maintaining good hygienic practices.

ISSN: 2233-7857 IJFGCN

- Providing the necessary equipment and infrastructure support for enabling smooth implementation of social distancing and proper air ventilation measures with regards to marketplaces, people transportation, production of goods and services, increasing production of immunity kits, etc.
- Enhancement in nationwide Hospital infrastructure support including adequate supply of PPE kits for health workers, patient ambulances, patient beds, isolation rooms, ICU rooms, ventilators, oxygen cylinders and reliable testing services etc.
- Building IT and telecom infrastructure capability for conducting virtual education classes for school children (coming from various sections of the society) through television channel and web broadcast so that their studies do not suffer.
- Creating more web-based digital training platforms for re-skilling of large pool of talented human resources who have been either rendered jobless or are on bench due to economic impact of COVID-19.

5. References

[1] Who.int. [Online]. Available: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200628-covid-19-sitrep-160.pdf?sfvrsn=2fe1c658_2 (**2020**).
[2] "Coronavirus in India: Latest map and case count," COVID19india.org. [Online]. Available: https://www.COVID19india.org (**2020**).
[3] "Springer Texts in Statistics," Springer.com. [Online]. Available: https://www.springer.com/series/417 (**2020**).
[4] M. Awad and R. Khanna, "Support Vector Regression," in Efficient Learning Machines, Berkeley, CA: Apress, (**2015**), pp. 67–80.
[5] R. S. Yadav, "Data analysis of COVID-2019 epidemic using machine learning methods: a case study of India," Int. J. Inf. Technol., (**2020**).
[6] S. Bernard Stoecklin et al., "First cases of coronavirus disease 2019 (COVID-19) in France: surveillance, investigations and control measures, " Euro Surveill., vol. 25, no. 6, (**2020**).
[7] Giorgia Purcaro, Christiaan A Rees, Wendy F Wieland-Alter, Mark J Schneider, Xi Wang, Pierre-Hugues Stefanuto, Peter F Wright, Richard I Enelow and Jane E Hill, "Volatile fingerprinting of human respiratory viruses from cell culture," J. Breath Res., vol. 12, no. 2, p. 026015, (**2018**).
[8] N. S. Punn, S. K. Sonbhadra, and S. Agarwal, "COVID-19 epidemic analysis using machine learning and deep learning algorithms," bioRxiv, (**2020**).
[9] M. Kwok and T. M. L. Tran, "For the future and possible ensuing waves of COVID-19: A perspective to consider when disseminating data," J. Popul. Ther. Clin. Pharmacol., vol. 27, no. S Pt 1, pp. e53–e57, (**2020**).
[10] D. Kp, "COVID-19 Corona Virus India Dataset." https://www.kaggle.com/imdevskp/COVID19-corona-virus-india-dataset, (**2020**).
[11] SRK, "COVID-19 in India. https://www.kaggle.com/sudalairajkumar/COVID19-in-india?select=COVID_19_india.csv", (**2020**).
[12] "Coronavirus cases in India and World Live: Covid-19 India tracker live, State wise corona cases in India and World," Indiatimes.com, https://timesofindia.indiatimes.com/coronavirus, (**2020**).

ISSN: 2233-7857 IJFGCN