Grenze International Journal of Engineering and Technology, Jan Issue

Analysis and Prediction of COVID-19 Spread in India

Santosini Bhutia[1], Bichitrananda Patra [2] and Mitrabinda Ray[3]
[1-3]Siksha O Anusandhan (Deemed to be) University, Bhubaneswar, India
Email: santosinibhutia@gmail.com, bichitranandapatra@soa.ac.in, mitrabindaray@soa.ac.in

***Abstract*—"Novel Coronavirus", also known as COVID-19 has spread globally which is sparked in Wuhan, the capital city of Hubei in China. It is a big problem for the world's health organization as the number of affected cases as well as deaths increases in each country. Several Machine Learning models are designed to predict and forecast the COVID-19 outbreak. In this paper, we are designed to understand the current scenario of coronavirus, and predicting and forecasting the total number of confirmed cases of novel COVID-19 in India using linear regression, polynomial regression, and support vector machine regression models. The decision-making process of an organization depends on the data analysis and it is presenting with interactive visual representation. It is not possible to predict without having necessary information from the past and the present. Trends over time tell us where we were and where we can go with our model. The model is predicting the number of confirmed cases from 30th May 2021 to 15th June 2021on the data available from 22 January 2020 to 29 May 2021. Forecasting the future of confirmed cases can help health organizations to plan medical facilities. Moreover, the time series forecasting technique is being engaged to forecast the overall total of confirmed cases in the upcoming.**

***Index Terms*— Coronavirus, COVID-19, Machine Learning, Linear Regression, Polynomial Regression, Support Vector Machine (SVM) Regression component.**

I. INTRODUCTION

The coronavirus is a respiratory disease caused by a new virus discovered at the end of 2019 in Wuhan in China. Thereafter it has been spreading all around the globe making it disastrous to humankind [1]. Then by World Health Organization (WHO), it is declared a pandemic on 11th March 2020 [2]. Due to the widespread of the disease and unavailability of drugs, it is a serious issue for every country. It is an infectious disease. The severe infection can cause pneumonia, severe acute respiratory syndrome, and multiple organ failure and also lead to death [3]. In a Chinese study, two strains such as L and S are found where the S type is older and the L type is very common at the early stage of the outbreak. It is a normal behavior of the virus to mutate. Hence, several viruses are developed and named according to the regions where they have developed. It is not possible to say how long the pandemic will continue due to the involvement of many factors such as the researcher's work to understand the nature of the virus, search for treatment and vaccine, and finally the public's action to slow down the spread. As coronavirus is new, people are not immune to it, and when the virus infects a person it triggers the body to make antibodies. During the end of 2020and at the beginning of 2021, the antibodies of some COVID-19 patients are examined and various clinical trials have experimented. After the experimental research, companies like Pfizer and BioNTech claimed the vaccine developed by the company is 95% efficient against the coronavirus. However, in a recent study, it is published due to the COVID-19 the global economy is expected to

Grenze ID: 01.GIJET.8.1.15_1

slow down by 1% as lockdown restricted many economic activities. Hence the unlocking process has started with proper measures.

Machine Learning (ML) is incredibly important in healthcare and particularly to handle this pandemic. Machine learning can help to fight against coronavirus in many ways such as: (1) Tracing the outbreak- ML models are used to trace the outbreak of coronavirus which helps in isolation as required for the current situation. (2) Diagnosis of the patient- If the diagnosis of the patient is done immediately who is infected by the virus can be isolated to stop the spreading of coronavirus. (3) Disinfecting the area- where the infection rate is high as identified by the ML model needs to be disinfected. (4) Improving in the health sector- ML can help in the treatment of the disease in health organizations. The numbers of COVID-19 confirm cases are around 28 million while 3lakh above death have occurred till the end of May 2021 in India. This is a challenging task for the medical organization to handle this situation. Early detection and generating forecast are very necessary to prevent and manage COVID-19. It is also required for the Government to provide and evaluate the efficiency of COVID-19 guidelines for the citizen to control the spread of the disease [4, 5]. ML plays an important role in this current scenario of the pandemic which includes helping in diagnosis, predicting, treatment, generating dashboard, tracking COVID-19 confirm cases, and many more control measurements [6, 7]. It is a vital role in the health sector for decision making.

To control the spread of the virus some safety measures are taken into consideration by the countries to avoid contact and maintain social distancing. The first case in India is reported in the last week of January 2020 and is the second-largest populated country in the world. Hence it stands in the top four affected countries in the world. To control this spread of the virus lockdown is implemented in two phases in the country and it showed a significant result. However, still deceased is the foremost concerning issue. For this reason, it is very important to predict the trend of the pandemic to help the government and medical professionals in decision-making. In this study, we have analyzed the COVID-19 dataset for Confirmed, Recovered, and Death cases in India using regression models and also forecasted the trend using time-series data. In section.2 of this paper, some literature published in the area of analysis and prediction of COVID-19 is presented. The detailed methodology starting from data collection and preprocessing, data visualization and all the regression models used are described in section.3. Thereafter results and discussion are presented in the section.4 followed by conclusion in section.5.

II. Literature Review

This section's related work focuses on the prediction of the transmission of the COVID-19 virus and the current scenario of the spread. The existing research and publication are addressed here. A novel support vector regression model is proposed instead of a simple regression line to obtain better classification accuracy and the result is compared with simple linear regression and polynomial regression model. The model predicts the spread of coronavirus, analyzes the growth rate, transmission rate, number of recoveries, and the correlation between coronavirus and weather [8]. Various classical and ensemble ML techniques are used for classification after the feature engineering process to a better understanding of the viral spread of COVID-19 and the performance is measured in terms of accuracy, precision, recall, and F1 score [9]. Mathematical models are proposed depending on susceptible, infected, and recovered (SIR) cases and susceptible, exposed, infected, quarantined, and recovered (SEIQR) cases with some parameter settings and optimization algorithms. The optimized SIR and SEIQR models were also compared with ML models and the results demonstrate optimized SIR and SEIQR model performs better [10].

III. Methodology

In this section the flowchart of the detailed methodology of the process: analysis, prediction, and forecasting of confirmed cases of COVID-19 are described as presented in Fig-1. The First COVID-19 dataset of India is collected. Thereafter data preprocessing is done followed by data visualization. Finally regression models: Linear Regression, Polynomial Regression, Support Vector Machine (SVM) Regression are implemented followed by the forecasting approach.

A. Data Collection and Preprocessing

The dataset used in this study is collected from Johns Hopkins University Centre for Systems Science and Engineering (JHU CSSE). The dataset is downloaded in CSV file format. The dataset is in time-series format with date, month, and year. The data is taken from 22 January 2020 to 29 May 2021 which consists of Observation Date, Province/State, Country/Region, Last Update, Confirmed, Deaths, and Recovered as input variables. In this study, the dataset is analyzed in Jupyter Notebook with Python 3 software by importing the

corresponding libraries. Here Pandas is a flexible data analysis and manipulation tool. For the analysis of the data preprocessing is done by dropping some columns and converting the Observation Date to date-time format. Further analysis is done by extracting the dataset for the country India which is the most populated country in the world. After that analysis is done state-wide.

B. Data Visualization

The graphical representation of the growth of COVID-19 disease across the country is shown in Figure-2. Figures 3-5 show the total number of confirmed cases, recovered, and deaths respectively from the different states and union territories in India. The objective of the study is to understand and visualize the outbreak of COVID-19 disease. Figure-6 gives the detailed information of confirmed cases, cured, deaths, active cases, death rate, and cured rate among different states and union territories in India which is sorted according to the confirmed cases. The more the number of cases the higher the intensity of the color.

C. Linear Regression

To build a Linear Regression model [11], linear regression analysis is done which predicts for the next few days numerical values of COVID-19 cases in India. It is a supervised machine learning algorithm. The performance of the algorithm is regression analysis. The model is fitted using the mean-square approach. This algorithm model a target prediction value for the next few days depending upon the number of confirmed cases as dependent variable (y) and date as an independent variable (x). So the linear regression model outlines a linear relationship between date (x) and confirmed cases (y) as follows:

$$\boldsymbol{y = a + bx} + \mathcal{E} \qquad (1)$$

Where b is the slope of the line and □ is an error.

D. Polynomial Regression

Polynomial regression [12] is also known as the multiple linear regression model in machine learning which models the relationship between the independent variable (x) and dependent variable (y) as nth degree polynomial. This regression model predicts the number of future cases based on confirmed cases as the dependent variable and date as the independent variable with the 8th degree of the polynomial. This algorithm is designed with some modifications to improve the accuracy of the model. The polynomial regression model is represented as:

$$y = \boldsymbol{\beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \cdots + \beta_n x^n} + \mathcal{E} \qquad (2)$$

Where n is the degree of the polynomial, β is the set of coefficients and □ is the error.

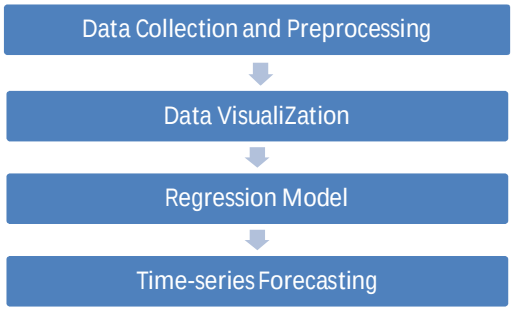

Figure 1. Flowchart of the methodology

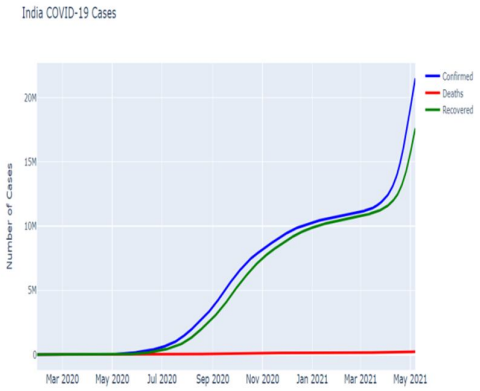

Figure 2. Spread of COVID-19 across the country

E. Support Vector Machine Regression

Support Vector Machine (SVM) regression [13] model is most popular in a supervised machine learning method used for regression analysis. SVM used to solve the regression problem is also known as support vector

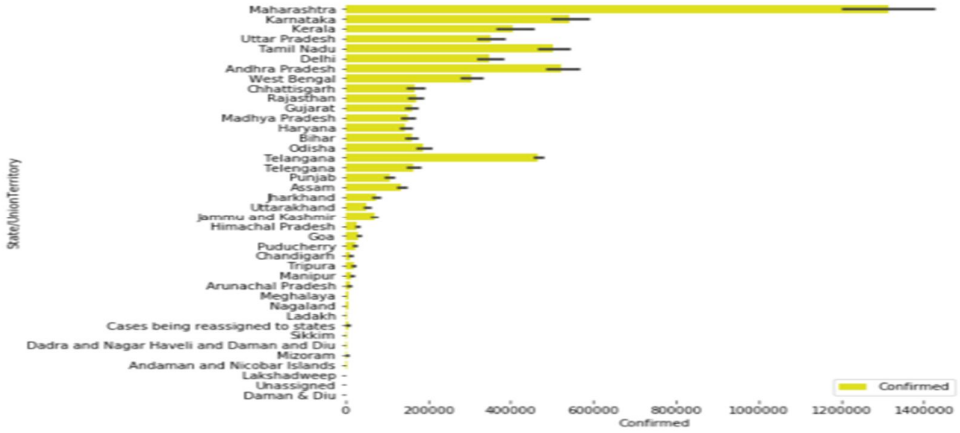

Figure 3. Confirmed cases of COVID-19

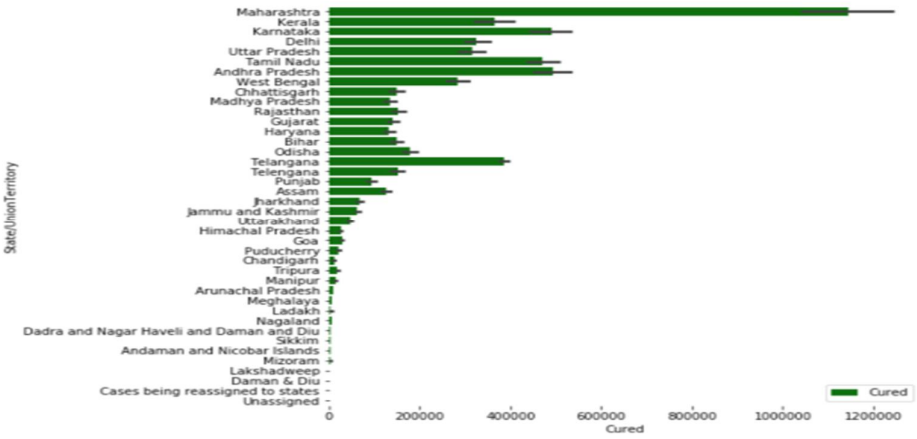

Figure 4. Recovered cases of COVID-19

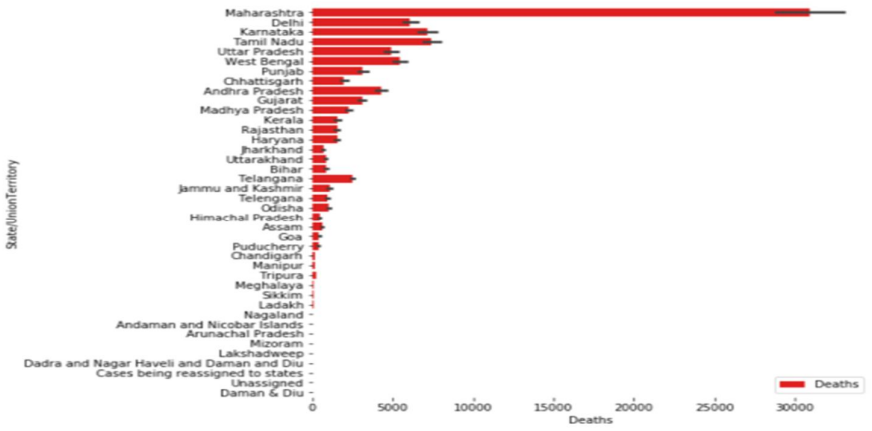

Figure 5. Deaths of COVID-19

regression (SVR) [14]. It can handle non-linear data using the kernel technique. The Kernel is a function that can map a low dimensional plane to a high dimensional space. Mainly three types of kernel are used in SVM namely: linear, polynomial, and Radial Basis Function (RBF). Here polynomial kernel is used which draws curved lines in the input space. The parameters used in this model are: C=1, degree=6, kernel='poly',

	State/Union Territory	Confirmed	Deaths	Cured	Active	Death Rate (per 100)	Cure Rate (per 100)
22	Maharashtra	4942736	73515	4227940	641281	1.49	85.54
17	Karnataka	1790104	17212	1255797	517095	0.96	70.15
18	Kerala	1786396	5628	1389515	391253	0.32	77.78
37	Uttar Pradesh	1425916	14501	1151571	259844	1.02	80.76
32	Tamil Nadu	1297500	14974	1151058	131468	1.15	88.71
10	Delhi	1273035	18398	1164008	90629	1.45	91.44
1	Andhra Pradesh	1228186	8446	1037411	182329	0.69	84.47
39	West Bengal	935066	11964	800328	122774	1.28	85.59
7	Chhattisgarh	816489	9950	675294	131245	1.22	82.71
30	Rajasthan	702568	5182	499376	198010	0.74	71.08
12	Gujarat	645972	8035	490412	147525	1.24	75.92
21	Madhya Pradesh	637406	6160	542632	88614	0.97	85.13
13	Haryana	573815	5137	452836	115842	0.9	78.92
4	Bihar	553803	3077	435574	115152	0.56	78.65
27	Odisha	500162	2121	423257	74784	0.42	84.62
33	Telangana	481640	2625	405164	73851	0.55	84.12
34	Telengana	443360	2312	362160	78888	0.52	81.69
29	Punjab	416350	9979	339803	66568	2.4	81.61
3	Assam	277687	1531	242980	33176	0.55	87.5
16	Jharkhand	270089	3479	205977	60633	1.29	76.26
38	Uttarakhand	220351	3293	154147	62911	1.49	69.96
15	Jammu and Kashmir	201511	2562	157283	41666	1.27	78.05

Figure 6. State/Union Territory-wide analysis of COVID-19 in India

epsilon=0.01, and thereafter Root Mean Square Error (RMSE) is calculated for the SVR model. The lower the value of RMSE the better is the model.

F. Steps for Forecasting Model

- Start
- Open Jupyter Notebook with Python 3
- Load the dataset with CSV file format
- The data is taken from 22 January 2020 to 29 May 2021 which consists of Observation Date, Province/State, Country/Region, Last Update, Confirmed, Deaths, and Recovered
- In the Analytic section, the three Forecast models for 17 days

```
for i in range(1,18):
    new_date.append(datewise_india.index[-1]+timedelta(days=i))
    new_prediction_lr.append(lin_reg.predict(np.array(datewise_india["Days Since"].max()+i).reshape(-1,1))[0][0])
    new_prediction_ poly.append (poly.predict(np.array(datewise_india["Days Since"].max()+i).reshape(-1,1))[0][0])
    new_prediction_svm.append(svm.predict(np.array(datewise_india["Days Since"].max()+i).reshape(-1,1))[0])
```

- The forecast data is obtained from 30th May 2021 to 15th June 2021.
- The result is displayed in the following format

```
pd.set_option('display.float_format', lambda x: '%.1f' % x)
model_predictions=pd.DataFrame(zip(new_date,new_prediction_lr,new_prediction_poly,new_prediction_svm),
                               columns=["Dates","Linear Regression Prediction","Polynonmial Regression Prediction","SVM Regression Prediction"])
model_predictions
```

- End

IV. Result and Discussion

Due to the pandemic, global health and the economy have been greatly affected. In this paper, Linear regression, Polynomial Regression, and Support Vector Machine Regression are used to forecast the confirmed cases in India in the COVID-19 dataset. The graphical representations of the models are shown in figures 7-9 respectively. The blue line represents the train data for confirmed cases and the black dot line represents the best fit line for the corresponding model. The formulation of root means square error (RMSE) values evaluated after the prediction of confirmed cases, recovered, and deaths of the models and is presented in Table-I. RMSE is the prediction error that tells us how concentrated the data is around the best fit line. It is commonly is for

forecasting and regression analysis to verify the experimental result. The lower the value of RMSE the better is the model. In this study, it is found that the SVM Regression model gives a lower RMSE value for the confirmed cases. Hence, the better the model is as compared to the other two models. But for the number of recovered cases and deaths polynomial regression model gives little better results compared to the SVM regression model.

$$\boldsymbol{RMSE} = \sqrt{\overline{(\boldsymbol{F}-\boldsymbol{O})^2}} \qquad (3)$$

Where F = forecasts or expected values

O = observed values

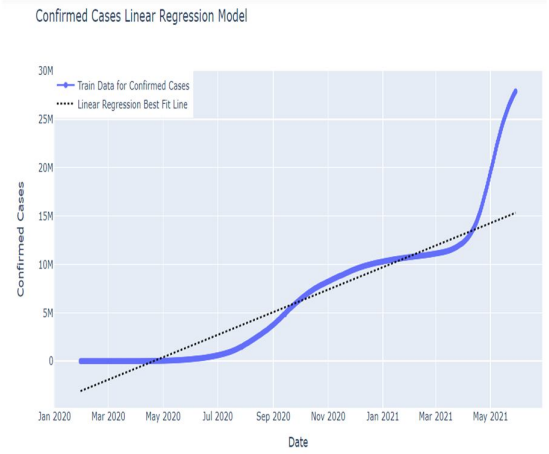

Figure 7. Linear Regression Model for Confirmed cases

Figure 8. Polynomial Regression Model for Confirmed cases

Figure 9. SVM Regression Model for Confirmed cases

The spread of COVID-19 analysis is done from 22 January 2020 to 29 May 2021 in India. The spread is growing more rapidly. Till now it is spread in 229 countries in the world by confirming more than 160 million cases. It is important to analyze and predict the disease to control the spread of COVID-19. In this study, Linear Regression, Polynomial Regression, and SVM Regression are used to build the models which forecast for the next 17 days that is from 30th May 2021 to 15th June 2021 confirmed cases in India which are presented in Table-II. When we compare the predictive data with the actual data the %of the accuracy of the SVM regression is more than the other two linear regression and polynomial regression which are shown in Table-III.

TABLE I. RMSE VALUES OF THE MODELS

	Linear Regression	Polynomial Regression	SVM Regression
Confirmed	10246280	5127334	1603095
Recovered	8140171	777071	858053
Death	79865	17593	20301

TABLE II. PREDICTED VALUES OF THE MODELS

Sl No.	Date	Actual Data	Linear Regression	Polynomial Regression	SVM Regression
1	30-05-2021	28047000	15375267	40035027	27103831
2	31-05-2021	28173883	15413275	41177458	27416098
3	01-06-2021	28307035	15451283	42357143	27731587
4	02-06-2021	28441079	15489290	43575080	28050325
5	03-06-2021	28573503	15527298	44832289	28372339
6	04-06-2021	28693957	15565306	46129813	28697656
7	05-06-2021	28808445	15603313	47468717	29026303
8	06-06-2021	28909654	15641321	48850089	29358306
9	07-06-2021	28995458	15679329	50275040	29693694
10	08-06-2021	29088245	15717336	51744706	30032493
11	09-06-2021	29182128	15755344	53260244	30374732
12	10-06-2021	29273977	15793352	54822838	30720439
13	11-06-2021	29358551	15831359	56433696	31069641
14	12-06-2021	29439076	15869367	58094052	31422367
15	13-06-2021	29510077	15907375	59805163	31778645
16	14-06-2021	29570085	15945382	61568316	32138503
17	15-06-2021	29632302	15983390	63384821	32501971

TABLE III. % OF THE ACCURACY OF THE MODELS

Date	Linear Regression	Polynomial Regression	SVM Regression
30-05-2021	54.18	57.25	96.63
31-05-2021	54.70	53.84	97.31
01-06-2021	54.58	50.36	97.96
02-06-2021	54.46	46.78	98.62
03-06-2021	54.34	43.09	99.29
04-06-2021	54.24	39.23	99.98
05-06-2021	54.16	35.22	99.24
06-06-2021	54.10	31.02	98.44
07-06-2021	54.07	26.61	97.59
08-06-2021	54.03	22.11	96.75
09-06-2021	53.98	17.49	95.91
10-06-2021	53.95	15.36	95.07
11-06-2021	53.92	12.72	94.17
12-06-2021	53.90	10.66	92.31
13-06-2021	53.90	9.68	92.31
14-06-2021	53.92	8.68	91.31
15-06-2021	53.93	9.68	90.31

V. CONCLUSIONS

In this study, Linear Regression, Polynomial Regression, SVM regression models have been experimentally tested for the forecasting of confirmed cases of COVID-19 in India. The comparison among the models is done and it is observed that the SVM Regression model gives more predictive values performance than the other two

models for time series data forecasting. However, the novel SVM Regression model gives significant results which much more accurate with the help of kernel function compared to the other two Regression models. From the analysis, it is observed that much more COVID-19 restriction guidelines are required to control the spread of the disease.

REFERENCES

[1] Hui, David S., et al. "The continuing 2019-nCoV epidemic threat of novel coronaviruses to global health—The latest 2019 novel coronavirus outbreak in Wuhan, China." *International journal of infectious diseases* 91 (2020): 264-266.

[2] World Health Organization. "Coronavirus disease 2019 (COVID-19): situation report, 82." (2020).

[3] Centers for Disease Control and Prevention. "Symptoms of coronavirus." (2020).

[4] Menni, Cristina, et al. "Loss of smell and taste in combination with other symptoms is a strong predictor of COVID-19 infection." *MedRxiv* (2020).

[5] Remuzzi, Andrea, and Giuseppe Remuzzi. "COVID-19 and Italy: what next?." *The lancet* 395.10231 (2020): 1225-1228.

[6] Prakash, KollaBhanu, and M. A. DoraiRangaswamy. "Content extraction studies using neural network and attribute generation." *Indian Journal of Science and Technology* 9.22 (2016): 1-10

[7] Ogundokun, Roseline O., et al. "Predictive modelling of COVID-19 confirmed cases in Nigeria." *Infectious Disease Modelling* 5 (2020): 543-548.

[8] Yadav, Milind, MurukessanPerumal, and M. Srinivas. "Analysis on novel coronavirus (COVID-19) using machine learning methods." *Chaos, Solitons& Fractals* 139 (2020): 110050.

[9] Khanday, AkibMohiUd Din, et al. "Machine learning based approaches for detecting COVID-19 using clinical text data." *International Journal of Information Technology* 12.3 (2020): 731-739.

[10] Rahimi, Iman, et al. "Analysis and Prediction of COVID-19 Using SIR, SEIQR and Machine Learning Models: Australia, Italy and UK Cases." *Information* 12.3 (2021): 109.

[11] Gambhir, Ekta, et al. "Regression analysis of COVID-19 using machine learning algorithms." *2020 International conference on smart electronics and communication (ICOSEC)*. IEEE, 2020.

[12] Parbat, Debanjan, and MonishaChakraborty. "A python based support vector regression model for prediction of COVID19 cases in India." *Chaos, Solitons& Fractals* 138 (2020): 109942.

[13] Sánchez A, V. David. "Advanced support vector machines and kernel methods." *Neurocomputing* 55.1-2 (2003): 5-20.

[14] Wu, Xiao, et al. "Air pollution and COVID-19 mortality in the United States: Strengths and limitations of an ecological regression analysis." *Science advances* 6.45 (2020): eabd4049.

[15] Kumar, SwarnAvinash, et al. "Deep Analysis of COVID-19 Pandemic using Machine Learning Techniques." Global Journal on Innovation, Opportunities and Challenges in Applied Artificial Intelligence and Machine Learning [ISSN: 2581-5156 (online)] 4.2 (2020).