Prediction of COVID-2019 cases through a Machine Learning Approach for India and United States of America

Mayank Pandey[1,*], **Shailendra Rai**[1,2] **and Shweta Singh**[3]

[1]*K. Banerjee Centre of Atmospheric and Ocean Studies, University of Allahabad, Prayagraj-211002, UP, INDIA*

[2]*M. N. Saha Centre of Space Studies, University of Allahabad, Prayagraj-211002, UP, INDIA*

[3]*Indian Institute of Technology (IIT), ISM, Dhanbad*

Abstract

To efficiently manage the world catastrophic of COVID-19 that involves serious casualties, recently several prediction models are efficiently being explored to accurately estimate the nature and pattern of the disease. The choice of an accurate prediction model not only leads to appropriate resource allocation and efficient management but also helps in controlling this severe pandemic to great extent. This paper presents the performance analysis of three predictive models to estimate the COVID-19 infected cases and death cases for India and USA. India and the USA reported the highest number of cases in the world. The behaviour of reported cases for COVID-19 was nonlinear so we validated our results using three predictive models namely support vector regression (SVR) model of machine learning (ML), linear regression(LR) model and polynomial regression(PR) model in order to depict the suitability of these models. The COVID-19 dataset has been drawn from the period of 30/01/2020 to 31/08/2021. The results demonstrate the degree of accuracy involved by each model for prediction. The results depict a higher level of accuracy attained using the SVR model as compared to other models. We obtained improvement in results from the PR model by increasing the degree of the model but after reaching a certain 4th degree further increment does not affect system performance. It was found that accuracy from the SVR model is the highest as compared to the rest of the models.

Keywords: COVID-19, INDIA, USA, Machine learning, Linear Regression, Polynomial Regression, Support Vector Machine.

1.Introduction

The world earmarked year 2020 as a catastrophic year for the entire world due to the spread of coronavirus disease of 2019 (COVID-19) pandemic.World Health Organization (WHO) declared the outbreak of COVID-19 as a global pandemic on March 11, 2020. It was believed that this pandemic started from Wuhan, China in December 2019 [1]. The first case of mortality was observed in January 2020 and the total mortality cases have gone up to 218 million till 31 August, 2021 all over the world and it is still continuing. The outbreak of COVID-19 throughout the world has posed severe restrictions on the government of each country to regulate the uprising of the COVID-19 cases of their respective country. Many countries like South Korea, New Zealand, Germany, France, Vietnam could successfully find their ways to control the outbreak of this disease however countries like India and USA had a serious devastating effect seen, that poses a great challenge to suppress this pandemic [2]. As on 31 August 2021, out of a total 65,907,563

**Corresponding Author(Email:mayankau@allduniv.ac.in)*

reported cases in India and USA, contribution for COVID-19 positive cases was 2.6% and 6.6% respectively [3]. This catastrophe has not only impacted human health and life but also resulted into great financial dearth on the economy of many countries. The researchers have started to forecast the trend line of cases based on some regression analysis model which can aid in appropriate allocation of resources [4]. In this regard, several researchers have proposed an efficient machine learning (ML) tool to appropriately predict short term and long term case analysis. This ML tool can be an effective prediction model to estimate the parameter which finds its application in several domains. ML is an efficient tool for feature prediction thus mitigating the amount of involved risk and helps in proper planning and management of resources by intercepting the disease in prior [5]. The efficiency of ML methods in handling a vast variety of datasets by characterizing the identification of patterns of an undetermined nature makes it a suitable choice for classification and prediction of complex datasets [6,7]. The main motive of such a prediction model using ML and Artificial Intelligence (AI) is to strategically plan the inventory so that optimal allocation of resources can take place [8-10]. However, there are only a few studies which reported the prediction model for COVID-19 which has given below.

A. Bansal et. al. has elaborated on the significance of ML as an efficient tool to predict COVID-19 cases[11]. This study has illustrated the detailed approach of each step involved in the COVID-19 pandemic. There are few studies which established their hypothesis of the number of reported cases for COVID-19 based on region and temperature [12]. Population based analysis for the United States of America (USA) was presented [13] that demonstrates the inverse proportionality behaviour between temperature of region and number of cases. Study for forecasting the number of death cases due to COVID-19 for China has been studied based on a patient information based algorithm (PIBA) [14].

However, this was limited as the hypothesis built did not hold true. Further, time series model was introduced which explored the auto regressive integrated moving average (ARIMA) model for predicting the number of COVID-19 positive cases in the most impacted portion of Europe including Spain, Italy and France [15]. This model helps in understanding the basic trends by suggesting the hypothetic epidemic's inflection point and final size [16,17]. A combination of ML and deep learning (DL) algorithms is presented for estimating the number of positive cases for COVID-19 [18]. High degree of effectiveness involved in the supervised learning based model of ML, regression analysis for prediction has been a major research interest. This is a data driven approach that is trained using previous learning and accordingly adopts a strategy that builds best to the given dataset. Piecewise linear regression has been investigated by Yang et al. [19]. Logistic regression has been explored and found the effectiveness of the predictive model based on logistic regression by comparison with other ML approaches [20]. The prediction of the number of cases due to COVID-19 for India was investigated through partial derivative regression and nonlinear machine learning (PDR-NML) model [21]. Tuli et al. [22] has used ML and cloud computing to find the increasing rate of the COVID-19 pandemic. The study has shown that using recursive weighting for fitting Generalized Inverse Weibull (GIW) distribution, a good approach can be obtained to develop a forecasting structure [22]. Pinter et al. [23] has worked on COVID-19 pandemic using hybrid ML methods of adaptive network based fuzzy inference system (ANFIS) and multi-layered perceptron-imperialist competitive algorithm (MLP-ICA). They have predicted the number of confirmed and death cases of COVID-19 pandemic [23]. Their model predicts with accuracy if there is no significant barrier occurs and provides initial tools for future research work using machine learning techniques [23]. There are studies [24] which briefly described the method to find accurate results of COVID-19 using ML techniques. They have analyzed the data using convolution neural networks with different validation techniques and prove better accuracy of the result. The work described by Singh et al. [25] on COVID-19 pandemic using auto regressive integrated moving average (ARIMA) and support vector machine (SVM) to predict current status of this disease. They have evaluated on the dataset derived from five countries for COVID-19 cases which are the most affected countries and also validated these models of the forecasting results [25]. The results showed the high accuracy of the SVM model over the ARIMA model and also suggested a rise of COVID-19 confirmed cases [25].

The COVID-19 pandemic has affected the population of the entire world while the sudden surge in cases for India and USA that has opened new challenges for the research community. Although ML methods have been formerly used to predict cases related to other pandemics like cholera, Ebola and few others but there recent research interest have been towards utilizing these efficient predictive tool dedicated to COVID-19. Multiple regression analysis tools have been used to predict the mortality cases in India caused due to SARS-CoV-2 for a time span of 56 weeks [26]. The hybrid approaches of auto regressive integrated moving average have been used to forecast the COVID-19 cases [27]. In this study, we have evaluated the performance parameter of three different regression models namely support vector regression (SVR), linear regression (LR) model and polynomial Regression (PR) model for prediction of COVID-19 pandemic for India and USA. The main aim of the presented analysis includes interpreting a time series COVID-19 pandemic data set for the prediction of the confirmed and death cases. The result outcome of present work leads to effectively predict and track the spread of the virus thus it can be a magnificent weapon for early alerts against battling COVID-19.

2.Mathematical Framework for Prediction of Covid-19 Cases

In this section three models of ML has been studied which are used to classify the output based on given input using different regression models. The results demonstrate the accuracy to predict the exact classification by analyzing the data based on respective algorithm. The data set has been taken from *ourworldindata.org* the format of the data set is comma separated value (CSV) which is included many columns and combined into a single . CSV file. The dataset has covered data from January 2020 to August 2021 for India and USA region. The dataset obtained has been utilized to predict the cumulative confirmed and death cases.

2.1. Linear regression

It represents the basic regression analysis termed as linear regression (LR) model. LR is the interpretation of relationship among the dependent and independent variables by assuming the relation between X and Y to be approximately linear [30]. Linear regression takes the following form:

$$y = p + qX \tag{1}$$

In the equation (1), y is the dependent variable, X is the independent variable. The symbol p denotes the intercept of the regression line and q is the slope of it. LR model has been fitted and tested into the dataset to predict the total number of positive and death cases for India and USA. The objective is to find the best fitting line to model our data. The algorithm tries to minimize the error with respect to slope and intercept.

2.2. Polynomial regression:

When the linear regression(LR) model is unable to capture the patterns in the data we increase the complexity of the model. To overcome under-fitting, we need to increase the complexity of the model [30]. To generate a higher order equation we can add powers of the original features as new features.

$$y = p - qX$$

The linear model, can be transformed to

$$y = p + qX + rX^2 + sX^3 + \ldots\ldots\ldots \quad (2)$$

This is still considered to be a linear model as the coefficients/weights associated with the features are still linear. To change the genuine properties into their higher order terms we will use the polynomial features class and after that we train the model using Linear Regression.

2.3. Support vector regression:

A support vector regression (SVR) is a binary perceptron divider method. Every raw information acts as points in space and classes are divided by a straight line "margin". The SVR maximizes the margin by placing the largest possible distance between the margin and the occurrence on both sides. A new data point would be classified according to which side of the margin it is going. Sequential minimal optimization (SMO) method is used when we train the model, which mainly uses for splitting the binary classification into many sub-problems and finds the maximum distance between the point and the instances in each class. Support vector are established from core method require only a user-specified kernel function over pairs of data points into kernel space on which learning algorithms operate linearly [31].

2.4 SVM model description

On the contrary to the regression analysis ML SVM algorithm is capable of handling complex datasets [30]. The key feature of the supervised learning in SVM algorithm leads to trained dataset that helps in better classification of complex data points. Thus, SVM is an efficient tool for classification of dataset which segregate the data points into two classes by a hyper plane/line. Based on the co-ordinates of dataset with reference to hyper plane the decision is made for classifying into classes. It assumes an n-dimensional space for the data points. SVM can be used for linear and nonlinear datasets. Specifically nonlinear classification is used for dividing the complex datasets into different class based on hyper plane positioning.

$$y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + \ldots\ldots\ldots + b_n x_n$$

$$y = b_0 + \sum_{i=1}^{n} b_i x_i \quad (3)$$

$$y = b_0 + b^T X$$

$$y - b + B^T X \quad (4)$$

where,

b_i =coefficient vectors (b_0,b_1,b_2,b_3,b_4,.....b_n)
b = biased term (b_0)
X = variable.

SVM- maximal margin classifiers (MMC) provides an imaginary classifier for the datasets. The term margin specifies the distance between the hyper plane and the datasets. The line possessing the highest margin and can distinctly classify the two classes is termed as optimal line also called the maximal-margin hyper plane. The term margin can also be defined as the right angle distance from the hyperplane to the nearest data point. Only these support points are the defining line and to the making of the classifier. These points are called the support vectors. The equivalent of a two-layer perceptron neural network is an SVM model with a radial bias kernel feature. SVM are used for training methods for radial basis function, polynomial and perceptron classifiers that use a kernel function and solve a nonlinear programming problem with linear constraints rather than a non-convex, unconstrained problem. In this paper, we have used SVM radial bias function. The equation for estimating new input using the dot product between the input and support vector is given as

$$y = C + \sum_{i=1}^{n} (t_i * (x * x_i)) \tag{5}$$

The dot product between the vectors is called kernel using equation (5). Here *x* denotes the new input vector with all support vectors (x$_i$). The coefficients *C* and t$_i$ must be estimated from the training dataset by supervised learning algorithm.

$$k(x, x_i) = \sum (x * x_i) \tag{6}$$

We have assumed a radial basis function (rbf) kernel to classify non linear data set due to its efficiency of efficiently evaluating the hyperplane for classification which can effectively classify the datasets.

$$k(x, x_i) = e^{-\gamma * \sum(x - x_i|^2)} \tag{7}$$

Where γ denotes the learning algorithm parameter which varies in the range of 0 <γ < 1. We have assumed value for gamma to be 0.1. The non-linear radial kernel can effectively create complex regions within the space.

Algorithm

Algorithm SVM using rbf:

```
Initialize
INPUT:Read the entire data set ${(x_i, y_i)}$ = where i =1, 2, 3, 4…….n
OUTPUT: Find the y_i on the each x_i of the testing dataset.
Reshape the dataset into a 1D dimension array of both x_i and y_i dataset.
Split the dataset into training and testing.
Training dataset = {(x_k, y_k)} where k ϵ {x_i, y_i)} and $k \not\in$ test data set
Testing dataset = $\{(x_t, y_t)\}$ where $t \in \{x_i, y_i\}$ and $t \not\in$ training dataset.
Scale the training and testing dataset
$\{(x_k, y_k)\}$ and $\{(x_t, y_t)\}$ :
$\frac{\{(x_i, y_i)\} - \{(x_i, y_i)\}.\mathbf{mean()}}{\{(x_i, y_i)\}.\mathbf{std}}$
```

Training using SVR rbf model

```
Hyperplane: classified the whole dataset
If $y_i(B^T * X_i + b) \geq 1$
    $X_i$ is correctly classified class 1
else
    $X_i$ is  classified class 0
end if
```

Radial kernel function

```
Initialize parameter $\gamma = 0.1$
for i=1 to n $x, x_i \in R^n$
    Calculate $k(x, x_i) - e^{-\gamma * \sum(|x - x_i|^2)}$
end for
for i =1 to n , $y_i \in R^n$
    $y = C + \sum_{i=1}^{n} (t_i * (x * x_i))$
end for
```

2.4 Measure the performance of Model

$$RMSE = \sqrt{\sum_{i-1}^{n} \frac{(Predicted_t - Actual_t)^2}{n}} \tag{8}$$

$$MSE = \sum_{i-1}^{n} \frac{(Predicted_t - Actual_t)^2}{n} \tag{9}$$

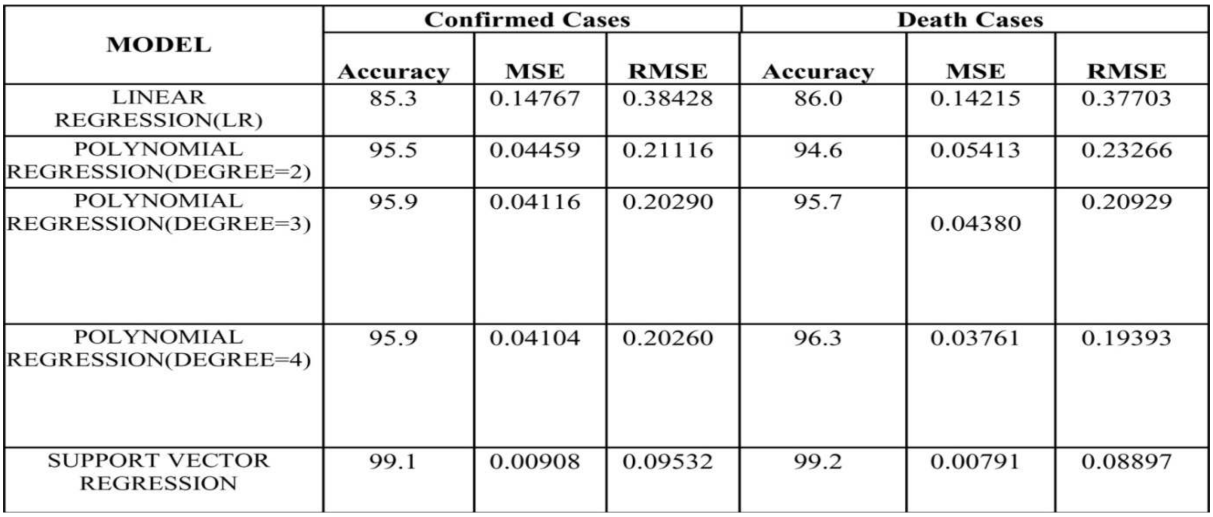
Where n is a number of dataset.

3. Results and Discussion

This section focuses on the results of prediction related to COVID-19 confirmed and death cases. The analysis is based on univariate cumulative forecasting of confirmed and mortality cases of COVID-19 from India and USA. To evaluate the accuracy of our test results, we have used the COVID-19 training dataset from 30 January 2020 to 31 August 2021. The Centre of Oxford Martin School at Oxford University has made the data publicly available at (https://ourworldindata.org/covid-cases).

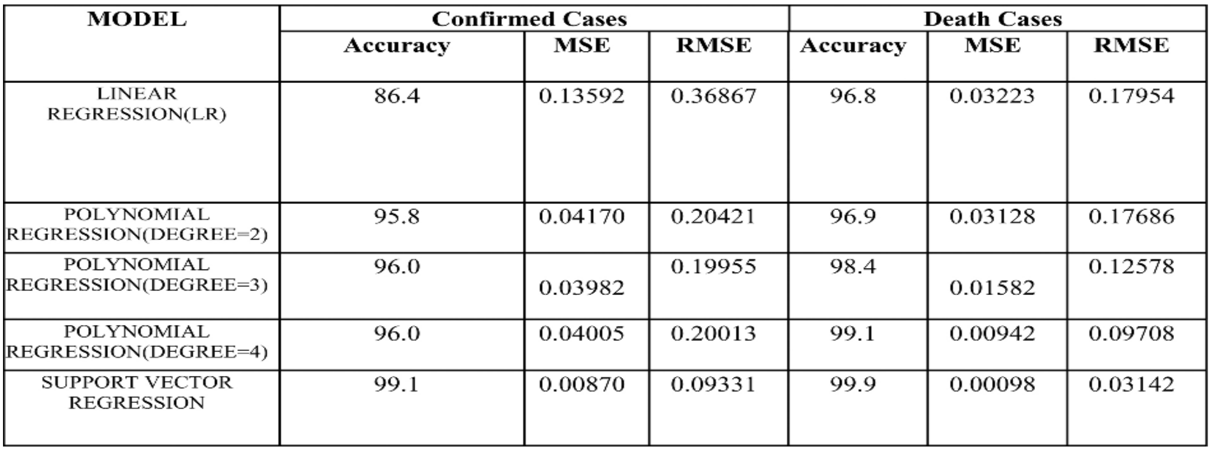
As shown in Figs. 1 and 2, the COVID-19 time-series dataset in the considered countries i.e. India and USA has been shown from January 2020 to August 2021. We first apply standard scalar within the appropriate ML model to normalise the dataset and then use this dataset in the ML models. We have used a reverse operation after the forecasting method to correlate the predicted data with the original COVID-19 time-series data for confirmed and death cases. We divided the normalised data into two sub-datasets namely training and test. We partitioned the dataset into training and test after shuffling the entire dataset. The objective is to evaluate the performance of the considered models to the one of standard ML models (LR, PR and SVR) due to limited datasets. The shuffling test dataset includes total confirmed and death COVID-19 cases present in India and USA from 30 January 2020 to 31 August 2021 using the forecasting outcomes from trained ML models. We compute the evaluation metrics (Accuracy, RMSE, and MSE) from the test datasets for the three ML models (i.e. LR, PR, SVR) after forecasting the COVID-19 time series using each trained model to improve the performance of each model quantitatively.

Table 1. Performance Parameter for India



MODEL	Confirmed Cases			Death Cases		
	Accuracy	MSE	RMSE	Accuracy	MSE	RMSE
LINEAR REGRESSION(LR)	85.3	0.14767	0.38428	86.0	0.14215	0.37703
POLYNOMIAL REGRESSION(DEGREE=2)	95.5	0.04459	0.21116	94.6	0.05413	0.23266
POLYNOMIAL REGRESSION(DEGREE=3)	95.9	0.04116	0.20290	95.7	0.04380	0.20929
POLYNOMIAL REGRESSION(DEGREE=4)	95.9	0.04104	0.20260	96.3	0.03761	0.19393
SUPPORT VECTOR REGRESSION	99.1	0.00908	0.09532	99.2	0.00791	0.08897

Table 2. Performance Parameter for USA



MODEL	Confirmed Cases			Death Cases		
	Accuracy	MSE	RMSE	Accuracy	MSE	RMSE
LINEAR REGRESSION(LR)	86.4	0.13592	0.36867	96.8	0.03223	0.17954
POLYNOMIAL REGRESSION(DEGREE=2)	95.8	0.04170	0.20421	96.9	0.03128	0.17686
POLYNOMIAL REGRESSION(DEGREE=3)	96.0	0.03982	0.19955	98.4	0.01582	0.12578
POLYNOMIAL REGRESSION(DEGREE=4)	96.0	0.04005	0.20013	99.1	0.00942	0.09708
SUPPORT VECTOR REGRESSION	99.1	0.00870	0.09331	99.9	0.00098	0.03142

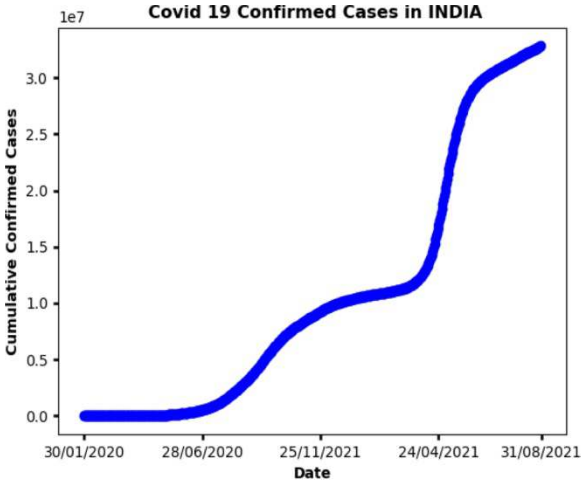


Fig1(a): Cumulative Confirmed cases of COVID-19 in India during 30/01/2020 to 31/08/2021.

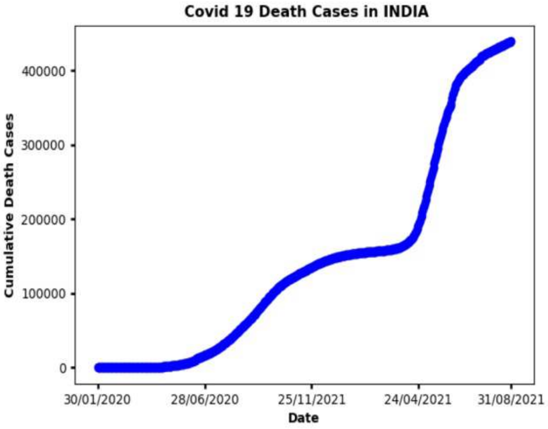


Fig1(b): Cumulative Death cases of COVID-19 in India during 30/01/2020 to 31/08/2021.

Tables 1 and 2 shows the forecasting results that have been obtained. The quality of the forecasts from the three models is promising, as shown in Tables 1 and 2 in which it was observed that ML models outperformed as compared to other classical model in prediction. Figures 3 and 4 shows the graph between the actual and the predicted data using the LR Model for India and USA respectively. In Figs.3 and 4, we have seen that the test dataset from 30/01/2020 to 28/03/2020, the LR model predicts a negative value, which is not possible in the current scenario. The minimum value is either 0 or greater than 0, which has not obtained in the LR model. We can also see that after March 28, 2020, the predicted value does not reach saturation as it does in the actual dataset, and it simply shows linear trends. The accuracy, MSE and RMSE of the LR model is obtained to be 85.3%, 0.14767 and 0.38428 respectively for total confirmed cases in India. Additionally, the accuracy, MSE and RMSE of LR model for total death cases of India is found to be 86.0%, 0.14215 and 0.37703 respectively. The total confirmed cases of the USA are investigated for the LR model at 86.4% accuracy; MSE and RMSE are 0.13592 and 0.36867 respectively. The quality of forecast for total death cases in USA for LR model is 96.8% (Accuracy), 0.03223(MSE) and 0.17954 (RMSE). We have presented the predicted and actual datasets using the PR model for degree 2 for India and USA in figs. 5 and 6 respectively. In figures 5 and 6, we have observed that the negative trends of the LR model has removed in PR model degree 2, but when saturation has come in test dataset after 25/11/2020, then PR model has not recognised the exact pattern and predicted as like LR model. The PR model for degree 2, the accuracy is found to be 95.5% and 94.6% of total confirmed and death cases in India and it is found to be 95.8% and 96.9% for USA. The accuracy of the model is increased in PR 2 model as compared to LR model. In Figs. 7, 8, 9 and 10 we have observed that quality of prediction for PR model degree 3 and 4 does not improve much in total daily cases of India and USA but we can see that in total death cases of USA the improvement in prediction is observed as compared to LR and PR model of degree 2. The accuracy increased slightly in case of PR model of degree 3 for the test dataset for India. The accuracy of PR models of degree 3 was found to be 95.9% and 95.7% for total confirmed and death cases respectively in India. The RMSE value of the total confirmed and death cases is 0.20290 and 0.20929 respectively for India. In USA, Fig. 8 presents the results of PR model of degree 3 that estimated the accuracy of 96.0% and 98.4% for cumulative confirmed and death cases having RMSE value of 0.19955 and 0.12578 respectively.

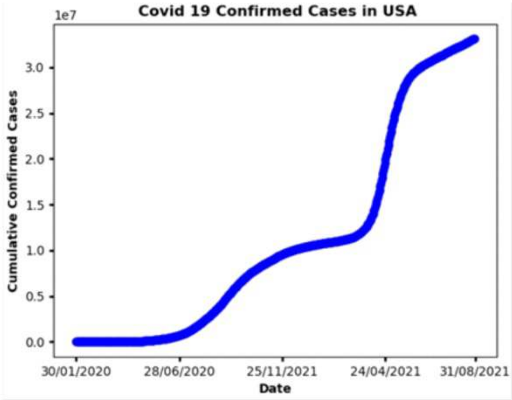


Figure 2(a): *Cumulative Confirmed cases of COVID-19 in the USA during 30/01/2020 to 31/08/2021.*

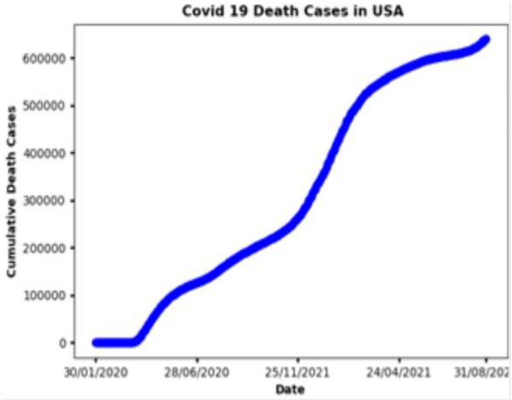


Figure2(b): *Cumulative Death cases of COVID-19 in the USA during 30/01/2020 to 31/08/2021.*

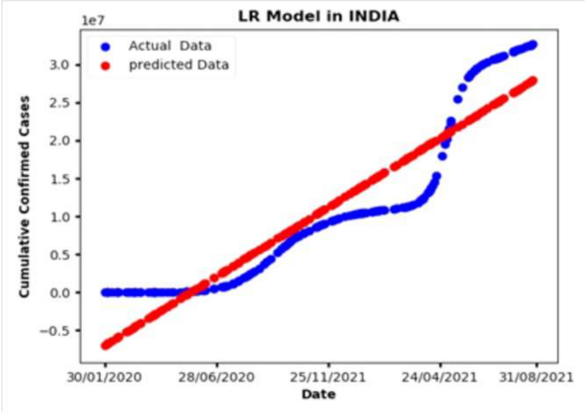


Figure3(a): *Observed and predicted values of cumulative confirmed cases of India using Linear Regression (LR).*

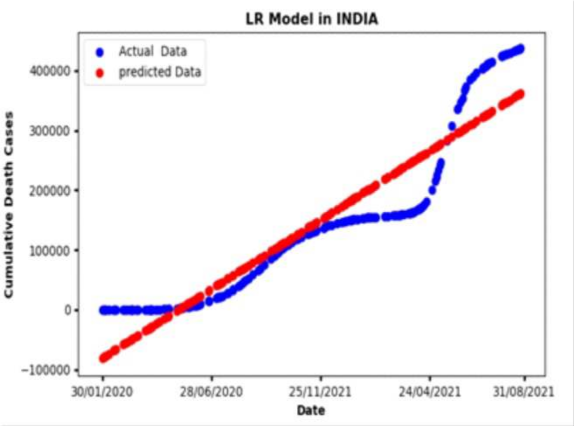


Figure3(b): *Observed and predicted values of cumulative death cases of India using Linear*

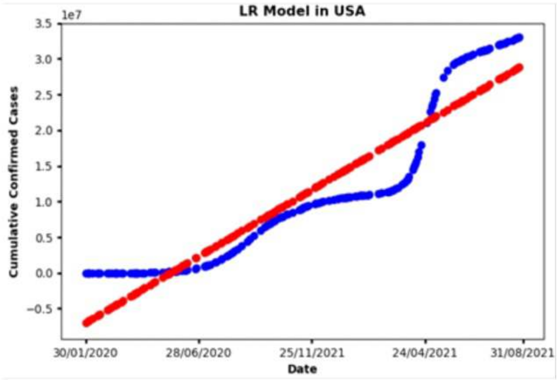


Figure4(a): *Observed and predicted values of cumulative confirmed cases of USA using Regression (LR).*

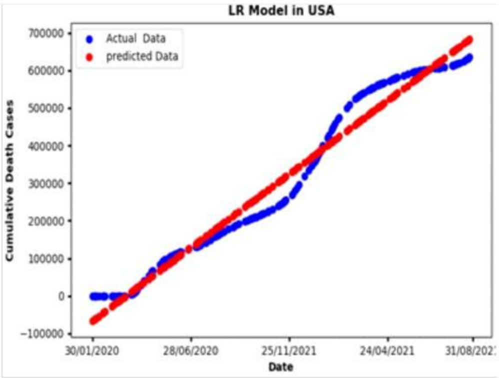


Figure4(b): *Observed and predicted values of cumulative death cases of the USA using Linear Regression (LR).*

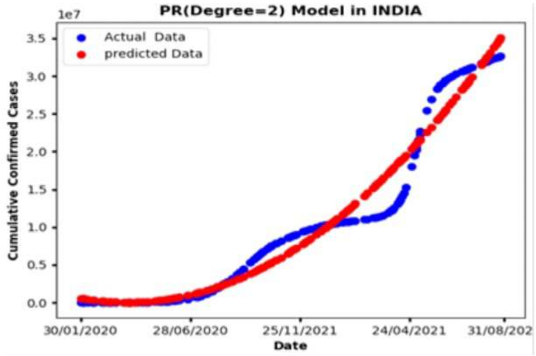


__Figure5(a):__ Observed and predicted values of cumulative confirmed cases of India using Polynomial Regression (Degree=2).

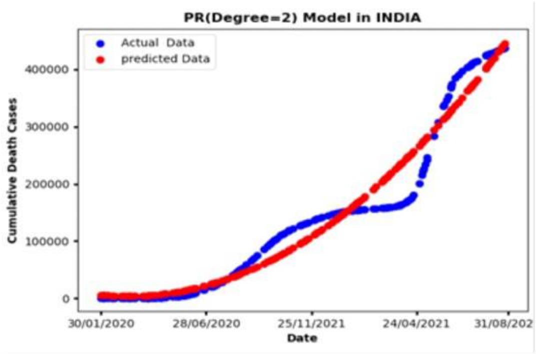


__Figure5(b):__ Observed and predicted values of cumulative death cases of India using Polynomial Regression (Degree=2).

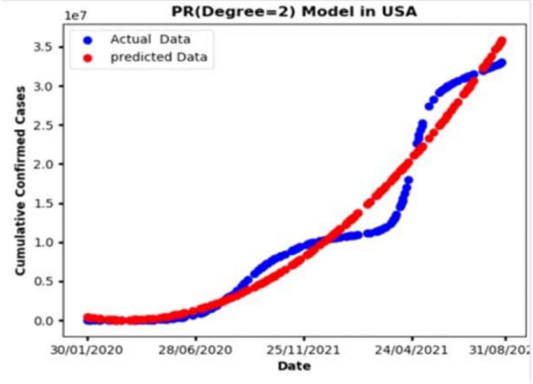


__Figure6(a):__ Observed and predicted values of cumulative confirmed cases of the USA using Polynomial Regression (Degree=2)

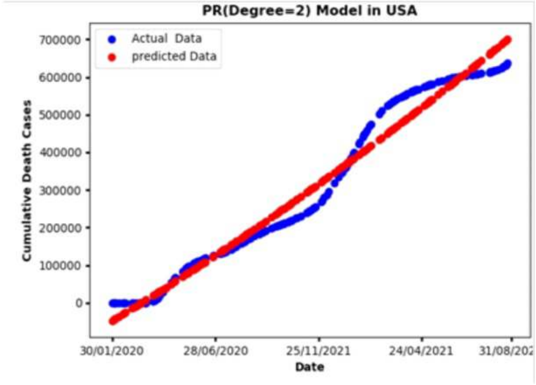


__Figure6(b):__ Observed and predicted values of cumulative death cases of the USA using Polynomial Regression (Degree=2)

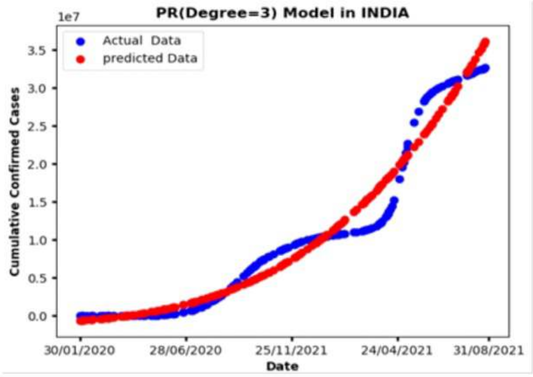


__Figure7(a):__ Observed and predicted values of cumulative confirmed cases of India using Polynomial Regression (Degree=3)

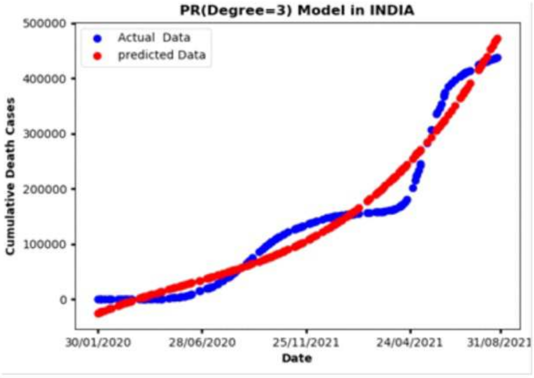


__Figure7(b):__ Observed and predicted values of cumulative death cases of India using Polynomial Regression (Degree=3).

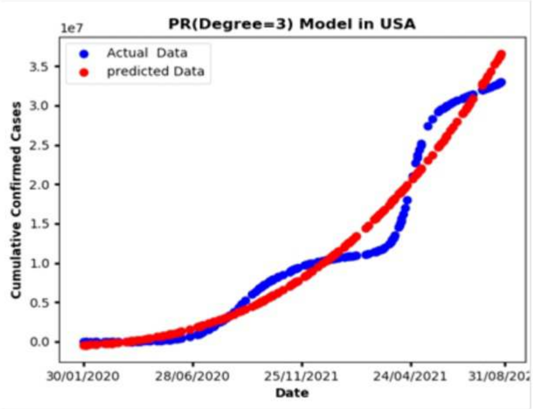


Figure 8(a): *Observed and predicted values of cumulative confirmed cases of the USA using Polynomial Regression (Degree=3).*

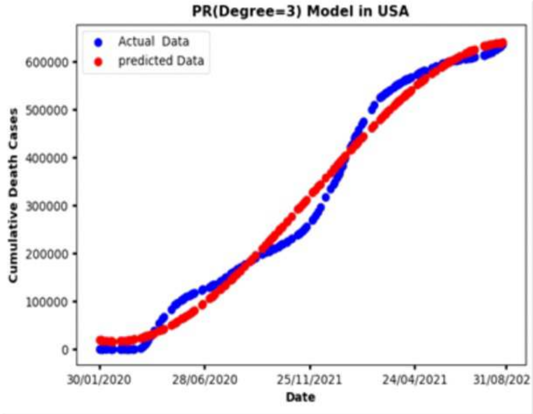


Figure8(b): *Observed and predicted values of cumulative death cases of the USA using Polynomial Regression (Degree=3).*

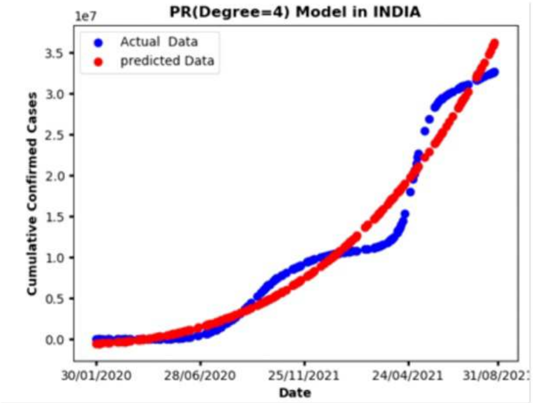


Figure9(a): *Observed and predicted values of cumulative confirmed cases of India using Polynomial Regression (Degree=4)*

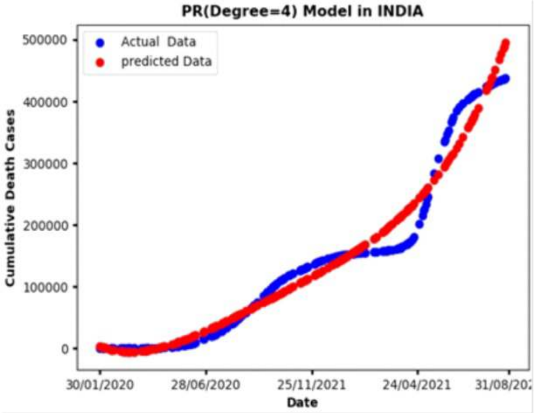


Figure9(b): *Observed and predicted values of cumulative death cases of India using Polynomial Regression (Degree=4).*

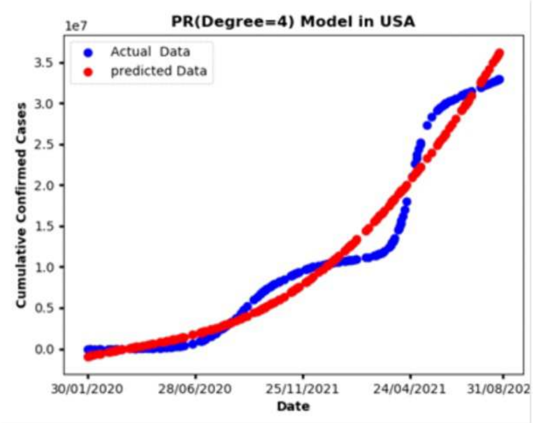


Figure10(a): *Observed and predicted values of cumulative confirmed cases of the USA using Polynomial Regression (Degree=4).*

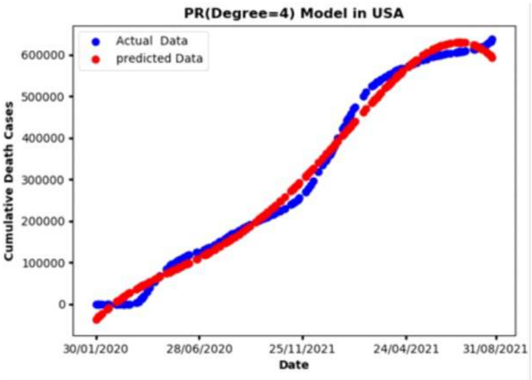


Figure10(b): *Observed and predicted values of cumulative death cases of the USA using Polynomial Regression (Degree=4).*

In Fig.9, it is seen that the PR model with degree 4 reached at saturation in terms of quality of prediction and no remarkable improvement is seen. The accuracy of 95.9% and 96.3% for total confirmed and death cases for India which is very close to PR model of degree 3. The similar results were obtained for USA also. It is also found that when we further increase the degree of the PR model then performance of the model does not improve so we have taken only PR model till degree 4.

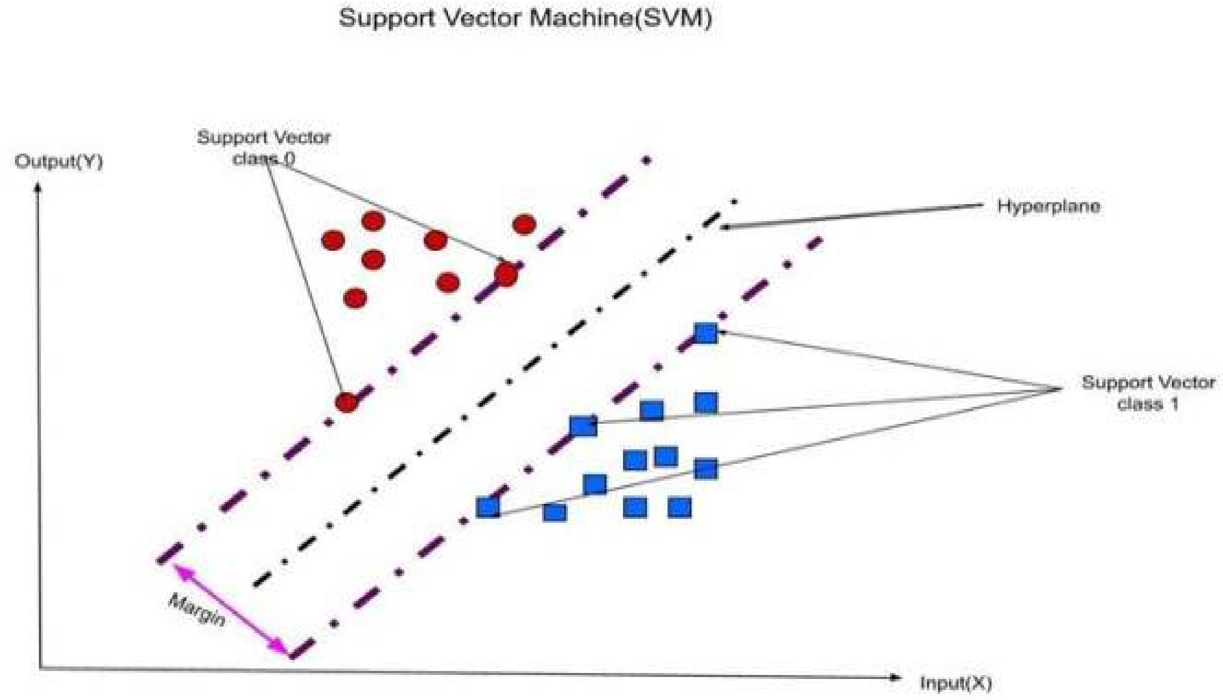


Figure11: SVM model Classification using hyperplane.

In Fig. 11, the SVM model classified the data points in two classes. The separation of the dataset is basically dependent on its classification into either class 0 or class 1. In Fig. 12, we have used the rbf that are the equivalent to a two-layer perceptron neural network model for SVM model. SVM are used to train rbf that use a kernel function. Figs.13 and 14 shows the forecasting of actual and predicted dataset using the SVR Model for India and USA respectively. In Fig. 13, we have observed that SVR model has provided better performance and predicted that saturation point which the LR and PR model has not recognised the correct pattern of the actual dataset. We see that the RMSE and MSE values for SVR for India and the USA cases are quite low, denoting high degree of correlation between test dataset. It is noteworthy that SVR are the most efficient and accurate forecasting methods for COVID-19 dataset. It could be attributed to its capability to model time-dependent data and extraction of high features. The COVID-19 dataset is evaluated by Ardabili et al. to establish a relationship between each dependent and independent characteristic of the dataset in order to construct several ML models [28]. They used SVM, Naive Bayes, and a decision tree model. In the COVID-19 dataset, 80% of the training dataset is applied to train the models, while the remaining 20% is used to test the models. According to Ardabili et al., the decision tree model has an accuracy of 94.99 %, whereas the SVM and the Naive Bayes Model have 92.40% and 94.36% accuracy respectively. In comparison, the presented SVM model has been used in predicting the test dataset for total confirmed and death cases of India and USA with high accuracy as earlier reported. The test results in the SVR model for India are obtained to be 99.1% accuracy of total confirmed cases and 99.2% accuracy for total death cases. In USA total confirmed and death cases result is found to be 99.1% and 99.9% accuracy respectively. Amir Ahmad et al. have investigated ML models for estimating the quantity of confirmed COVID-19 cases[29]. They found different complex problems for accurate prediction using machine learning approaches in their work. The first challenge they find is that training appropriate machine learning models with minimal datasets is a difficult task. However, we could able to predict COVID-19 total confirmed and death cases in USA and India accurately using SVR method with high degree of accuracy.

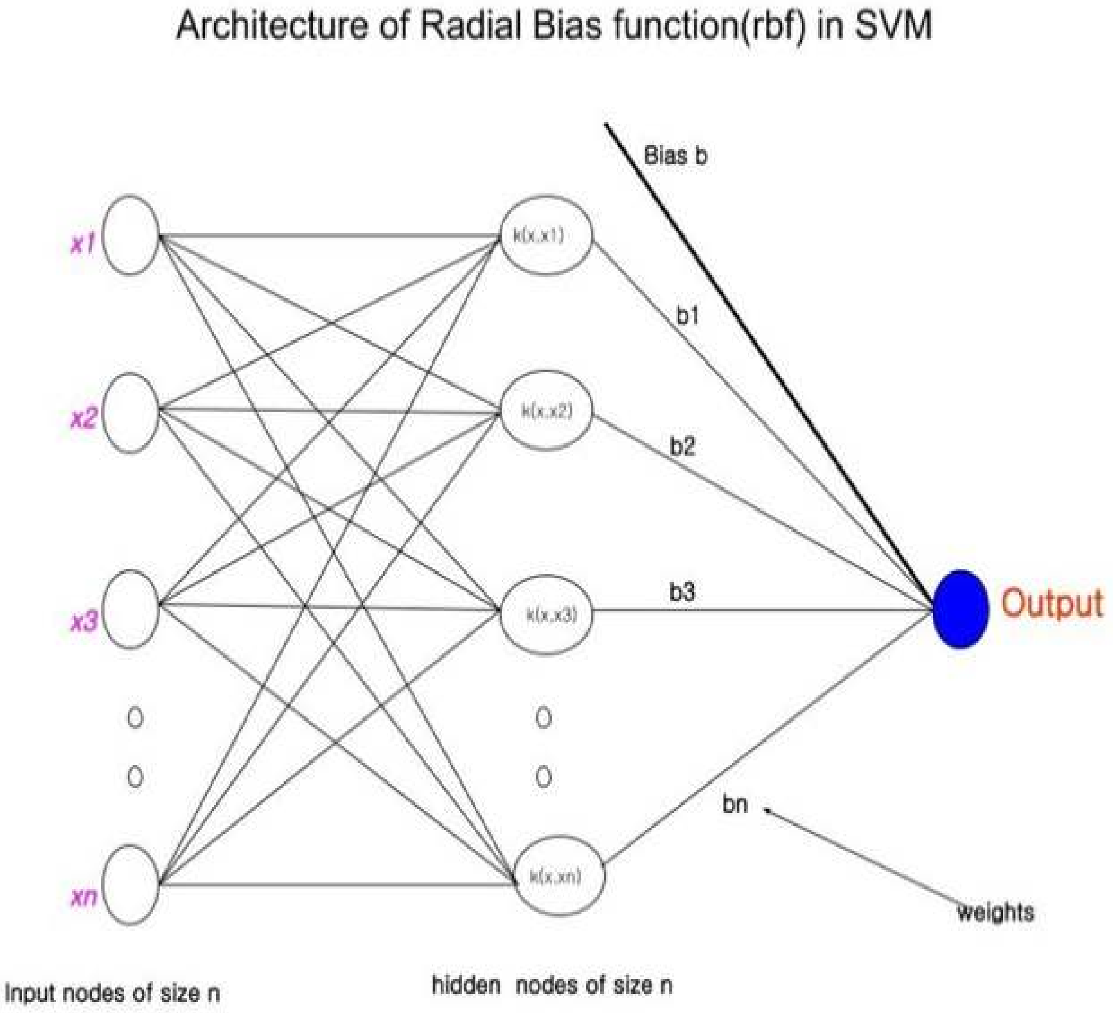


Fig. 12: Architecture of SVM (rbf) kernel function

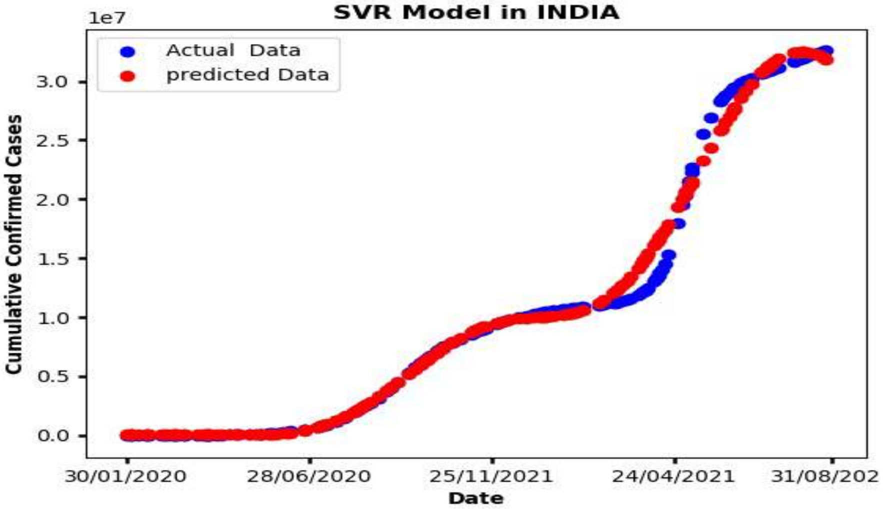
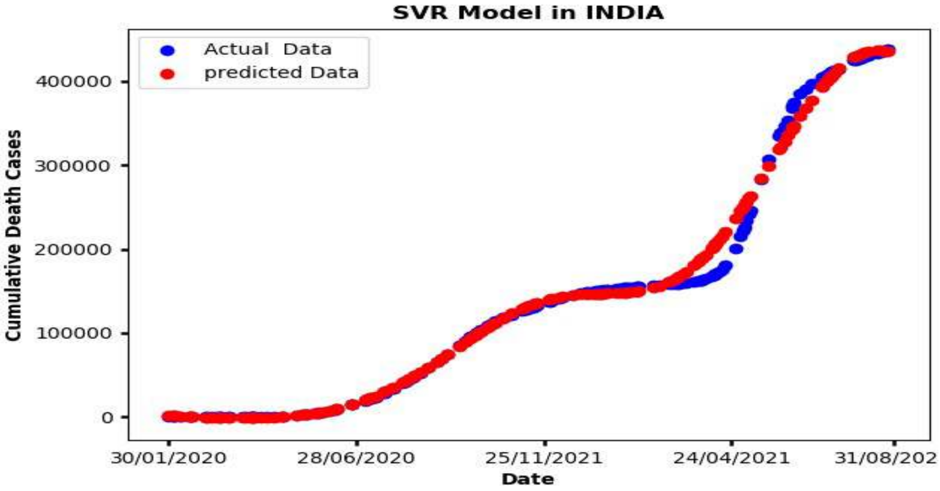


Fig. 13(a): Observed and predicted value ofcumulative confirmed cases of India using (SVR).



\

Fig. 13(b): Observed and predicted values ofcumulative death cases of India (SVR).

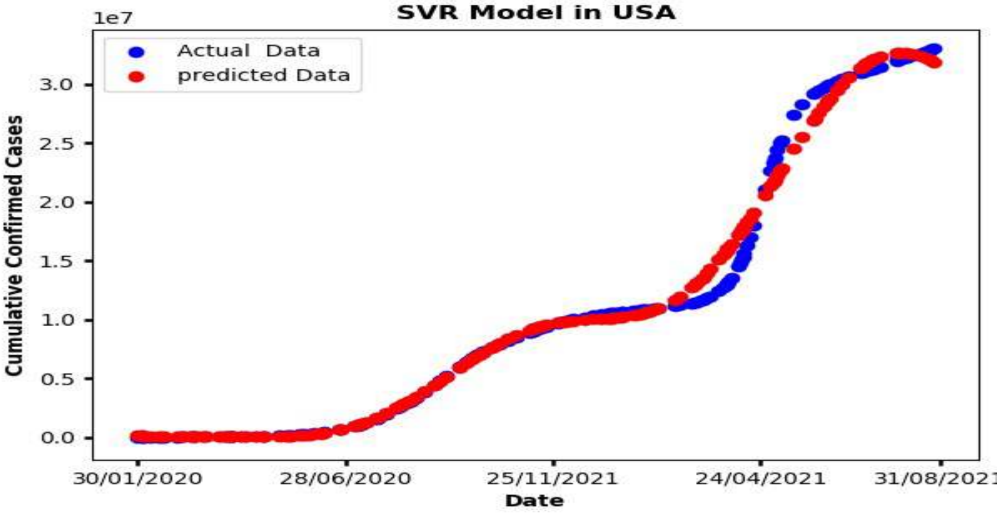


Fig. 14(a): Observed and predicted values ofcumulative confirmed cases of the USA using (SVR).

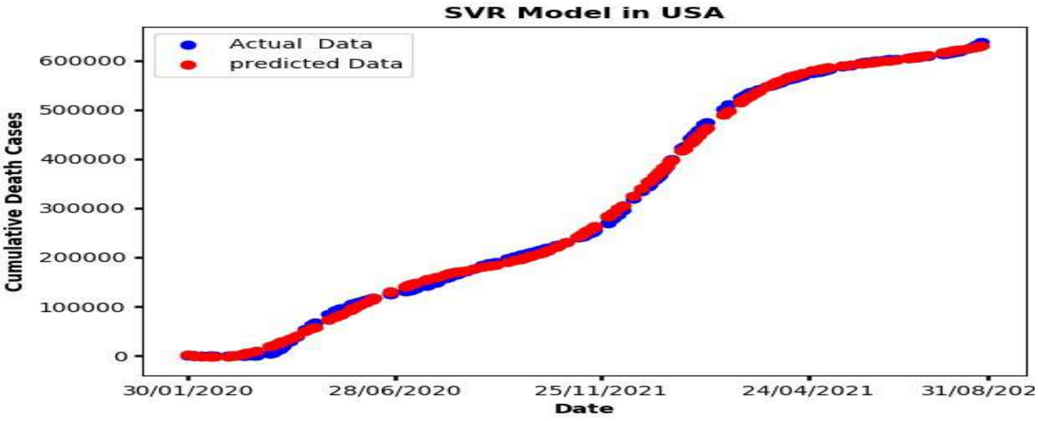


Fig. 14(b): Observed and predicted values ofcumulative death cases of the USA using (SVR).

SVR models provided good estimates with an accuracy of greater than 99% in both cumulative confirmed and death cases of India and USA (shown in the table 1 and 2). For all the cases, the results indicate that in the initial stage of COVID-19, the reported cumulative confirmed and deaths cases is satisfying with SVR models based on the ML. Therefore, the estimate that the result comes from the study on COVID-19 total confirmed and deaths cases of test dataset derived from SVM models could give better understandings of the pandemic in India as well as USA.

4. Conclusion

This paper presents a novel modified prediction and analysis SVR tool for evaluating the COVID-19 affected cases and death cases in India and USA. The results are compared against popularly explored LR and PR models to provide a better insight of the high degree of accuracy attained by SVR method. The LR is the basic regression model with reasonable accuracy hence this model is taken as a baseline for evaluation of performance of SVR and PR model. The improvement in results obtained from the PR model by increasing the degree of the model but after reaching certain threshold degree further increment does not affect model performance. The high degree of accuracy offered by SVR and its versatile prediction analysis that works well for time series data of COVID-19 scenario makes it a suitable choice for extending for prediction. This study evaluates that SVR models provide prediction upto a reasonable level of accuracy. In SVR, we obtained accurate results due to its rbf methodology. The high level of correlation involved between predicted and actual dataset reveals that SVR has been provided 99% accuracy for COVID-19 dataset analysis. The results illustrates that India should adopt more precautionary and regulating steps to control the increasing number of cases of COVID-19.

Acknowledgement

MP Acknowledge the funding from University Grants Commission (UGC), MHRD, New Delhi for fellowship to conduct PhD study.

References

[1] Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, Zhang L, Fan G, Xu J, Gu X, Cheng Z(2020) Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. The Lancet 395(10223):497-506.

[2] Senapati A, Nag A, Mondal A, Maji S A novel framework for COVID-19 case prediction through piecewise regression in India. International Journal of Information Technology13(1):41-48(2021).

[3] Mortality analyses-johns hopkins coronavirus resource center. https://coronavirus.jhu.edu/

[4] Fanelli D, Piazza F Analysis and forecast of COVID-19 spreading in China, Italy and France. Chaos, Solitons & Fractals134:109761(2020).

[5] Dahiwade D, Patle G, Meshram E Designing disease prediction model using machine learning approach.In 3rd International Conference on Computing Methodologies and Communication (ICCMC),pp.1211-1215 (2019).

[6] Uddin S, Khan A, Hossain ME, Moni MA Comparing different supervised machine learning algorithms for disease prediction. BMC Medical Informatics and Decision Making19(1):1-6(2019).

[7] Sharmila SL, Dharuman C, Venkatesan P Disease classification using machine learning algorithms-a comparative study.International Journal of Pure and Applied Mathematics114(6):1-0(2017).

[8] Davenport T, Kalakota R The potential for artificial intelligence in healthcare. Future healthcare journal6(2):94(2019).

[9] Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, Wang Y, Dong Q, Shen H, Wang Y Artificial intelligence in healthcare: past, present and future. Stroke and vascular neurology1;2(4) (2017).

[10] Grampurohit S, Sagarnal C Disease prediction using machine learning algorithms. In2020 International Conference for Emerging Technology (INCET) ,pp. 1-7 (2020).

[11] Bansal A, Padappayil RP, Garg C, Singal A, Gupta M, Klein A Utility of artificial intelligence amidst the COVID 19 pandemic: a review.Journal of Medical Systems44(9):1-6(2020).

[12] Meraj G, Farooq M, Singh SK, Romshoo SA, Nathawat MS, Kanga S Coronavirus pandemic versus temperature in the context of Indian subcontinent: a preliminary statistical analysis. Environment, Development and Sustainability23(4):6524-34. https://doi.org/10.1007/s10668-020-00854-3 (2021)

[13] Gupta S, Raghuwanshi GS, Chanda A Effect of weather on COVID-19 spread in the US: A prediction model for India in 2020. Science of the total environment1;728:138860 (2020).

[14] Wang L, Li J, Guo S, Xie N, Yao L, Cao Y, Day SW, Howard SC, Graff JC, Gu T, Ji J Real-time estimation and prediction of mortality caused by COVID-19 with patient information based algorithm. Science of the total environment727:138394 (2020).

[15] Ceylan Z Estimation of COVID-19 prevalence in Italy, Spain, and France.Science of The Total Environment729:138817(2020).

[16] Benvenuto D, Giovanetti M, Vassallo L, Angeletti S, CiccozziM Application of the ARIMA model on the COVID-2019 epidemic dataset. Data in brief 1;29:105340 (2020).

[17] Solanki A, Singh T COVID-19 Epidemic Analysis and Prediction Using Machine Learning Algorithms. Emerging Technologies for Battling Covid-19: Applications and Innovations.57-78 (2021).

[18] Punn NS, Sonbhadra SK, Agarwal S COVID-19 epidemic analysis using machine learning and deep learning algorithms. https://doi.org/10.1101/2020.04.08.20057679 (2020).

[19] Yang X, Yang H, Zhang F, Zhang L, Fan X, Ye Q, Fu L Piecewise linear regression based on plane clustering. IEEE Access 7:29845-55(2019).

[20] Itoo F, Singh S Comparison and analysis of logistic regression, Naïve Bayes and KNN machine learning algorithms for credit card fraud detection. International Journal of Information Technology.13(4):1503-11(2021).

[21] Kavadi DP, Patan R, Ramachandran M, GandomiAH Partial derivative nonlinear global pandemic machine learning prediction of covid 19. Chaos, Solitons & Fractals.1;139:110056(2020).

[22] Tuli S, Tuli S, Tuli R, Gill SS Predicting the growth and trend of COVID-19 pandemic using machine learning and cloud computing. Internet of Things.1;11:100222(2020).

[23] Pinter G, Felde I, Mosavi A, Ghamisi P, GloaguenR COVID-19 pandemic prediction for Hungary; a hybrid machine learning approach. Mathematics.8(6):890 (2020).

[24] Lalmuanawma S, Hussain J, ChhakchhuakL Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: A review. Chaos, Solitons & Fractals. 1;139:110059(2020).

[25] Singh S, Parmar KS, Makkhan SJ, Kaur J, Peshoria S, Kumar J Study of ARIMA and least square support vector machine (LS-SVM) models for the prediction of SARS-CoV-2 confirmed cases in the most affected countries. Chaos, Solitons & Fractals.1;139:110086 (2020).

[26] Ghosal S, Sengupta S, Majumder M, Sinha B Linear Regression Analysis to predict the number of deaths in India due to SARS-CoV-2 at 6 weeks from day 0 (100 cases-March 14th 2020). Diabetes & Metabolic Syndrome: Clinical Research &

Reviews.1;14(4):311-5. https://doi.org/10.1016/j.dsx.2020.03.017 (2020)

[27] Yadav RS Data analysis of COVID-2019 epidemic using machine learning methods: a case study of India. International Journal of Information Technology.12:1321-30(2020).

[28] Ardabili SF, Mosavi A, Ghamisi P, Ferdinand F, Varkonyi-Koczy AR, Reuter U, Rabczuk T, Atkinson PM Covid-19 outbreak prediction with machine learning.Algorithms.13(10):249(2020).

[29] Ahmad A, Garhwal S, Ray SK, Kumar G, Malebary SJ, BarukabOM The number of confirmed cases of covid-19 by using machine learning: Methods and challenges. Archives of Computational Methods in Engineering.28(4):2645-53(2021).

[30] Douglas.C.Montgomery "Introduction to Linear Regression Analysis, 5th Edition"Wiley series In Probability and Statistics,ch.2,7,pp38-43,428-432(2012).

[31] Christopher M. Bishop "SPARSE KERNEL MACHINES"the Pattern Recognition and Machine Learning,Microsoft Research Ltd,Cambridge:Microsoft Research Ltd U.K.,2006,ch.7,pp. 325-356(2006).