

Forecasting Pandemic Diseases with Regression Techniques

Dr Shantala Giraddi¹, Dhanashree R I², Dr Satyadhyan Chikkerur³, Preeti T⁴ and Dr Suvarna G K⁵

¹⁻⁵KLE Technological University, Hubli, India

¹shantala@kletech.ac.in, ²dhanashree.ingalahalli5@gmail.com, ³chikkerur@kletech.ac.in,

⁴preeti@kletech.ac.in, ⁵suvarna_gk@kletech.ac.in

Abstract—As of 16th July there are 13,338,364 confirmed cases and 579,319 deaths. Since then the disease is spreading its tentacles to every corner of the world, every country is struggling to manage the disease by channelizing all their resources. South economic disaster of the century. Early predictions of such epidemics help in ramping up efforts to contain the spread. Planning is very much essential in lessening the sudden and potentially catastrophic impact on society. In this paper, we discuss Linear China Morning Post reported the first case of Covid-19 on 17 November 2019 in Hubei province. Subsequently Hubei province became Covid epicenter. Almost 81,000 of the cases occurred in China, mostly in the province of Hubei. The first case outside of China was from Thailand on 13 January 2020. First case of COVID-19 was reported in India in January CORONA is the greatest and SVR regression, Polynomial regression techniques in predicting the COVID 19 in different countries.

Index Terms— COVID, Forecasting, Disease prediction, linear regression, support vector regression.

I. INTRODUCTION

Community well being is a significant issue. Health care providers should be updated about the health of people and outbreak of diseases. This will assist them offer enhanced services in an effective manner and at the appropriate time. Prediction of epidemics helps in taking preventive actions and plan effective treatment at an initial stage. Novel Corona virus (2019- nCoV), now COVID-19 appeared in Wuhan, China and spread all over the world. A large chunk of global citizenry is affected by COVID. World Health Organization declared CORONA as pandemics on 30 January 2020. Figure 1. shows the spread of CORONA in India as on 22nd June. In this work, authors have conducted a study on forecasting future evolution of pandemic using Linear and SVR regression techniques. The objective of this study was to build an effective regression model with high accuracy.

A. Support Vector Machine

Support Vector machine is supervised machine learning model. It can be used for both classification and regression. It has broadly two versions to support linear and non linear problems. For Linearly separable data, linear SVM is used. SVM with kernels are used when the solution is not linearly separable. The kernel functions transpose the data into a higher dimensional feature space. Now the data becomes linearly distinguishable and a linear curve will classify the labels in the hyper plane. Amalgamation of Capacity parameter C , the kernel function parameter ϵ , and the kernel parameter σ controls the complexity and accuracy the model.

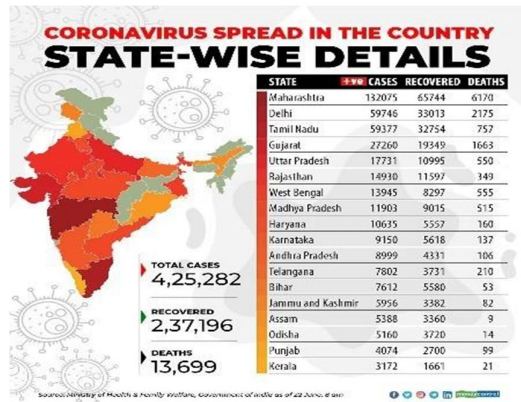


Fig 1. Highly infected states in India as on 22nd June 2020

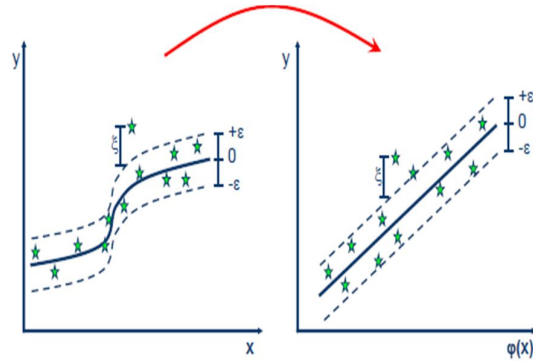


Fig 2. Non Linear data transformed in to linear data by kernel function

II. RELATED WORK

Shawni Dutta et al (2020) proposed and conducted study on Covid detection using Deep-learning techniques. Two techniques are used together and they are Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU). The disease has exponential growth and with the existing poor doctor patient ratio, it would be impossible to handle a huge number of patients. The detection techniques are much useful in the present scenario given the highly infectious nature of the disease.

Pavan Kumar et al., (2020) conducted a study on Forecasting COVID-19 impact using an Auto regression integrated with moving average model (ARIMA) and Richard's model. Authors have predicted the number of incidence of new cases to be 5200 (ARIMA model) and 6378(Richard model) by April 2020.

Support vector machines are also being used for regression. Bing Zhang et al. (2019), proposed a system for blood pressure prediction with SVR. Traditional blood pressure prediction measurement was given at certain intervals but there is an efficient blood pressure prediction method based on the Support Vector Machine regression (SVR) algorithm. This algorithm measures continuous values. The results show that the SVR model is more accurate than linear regression and BPNN and effective to predict the blood pressure.

Linear Regression models are used for regression H. ROOPA et al., (2019), different diseases are predicted using various classification methods. In this paper, eight features are extracted for diabetes prediction which includes glucose, blood pressure, skin thickness, age, BMI etc. Using principal component analysis the data is projected and then modeled by applying a linear regression model. By this method, accuracy of 82.1% for diabetes prediction is achieved over the classification methods. Logistic Regression is also used for prediction whether a person is infected with malaria. Ellis Kobina Paintsil et al., (2019), In Ghana the overall prevalence of malaria is about 45.3% between the age groups of 5 and 14. To predict the number of people infected, binary logistic model is developed by considering the following predictors like age, hemoglobin, platelet and lymphocytes. Diagnosis of malaria is improved and waiting time is reduced using Logistic model compared to RDT(Rapid Diagnostic Test). Anirudh Ravi et al., (2019), Malaria is an infectious disease among the human community. The objective is prediction of malaria based on the individual parameters using logistic regression model. Disease detection is done with maximum accuracy, speed and high reliability in this proposed method of detecting infectious disease. (Zlatan et al., (2020) conducted study on spread of COVID multilayer perceptron (MLP) and artificial neural network (ANN). Authors used a publicly available dataset. Grid search algorithm used for hyper parameter setting. Neural network with 4 hidden layers yielded best results.

Ogundokun et al., (2020) conducted study with 3 models Support Vector Regression, Neural Network, and Linear Regression on number, size, and length of COVID-19 cases. Authors conclude that CORONA can spread through air and water. Preventive measures such as wearing masks, social distancing, wearing gloves are important in slowing down the spread.

Rustum et al.(2020) made three types of predictions confirmed cases, deaths and recovered cases. Statistical models linear regression (LR), least absolute shrinkage and selection operator (LASSO), exponential smoothing (ES) as well as machine learning models such as support vector machine have been used.

Pandey et al., (2020) conducted study in Indian context with two models SEIR model and Regression model. Public data from John Hopkins University repository used for the study. Authors predict between 5000-6000 in subsequent two weeks. Shahid et al., (2020) conducted a study on prediction of confirmed cases, deaths and recoveries of COVID-19 using several forecast models like AutoRegressive Integrated Moving Average (ARIMA), Support Vector Regression (SVR), Long Short Term Memory (LSTM) and Bidirectional Long Short Term Memory (Bi-LSTM). The models ranked according to the performance i.e., Bi-LSTM, LSTM, GRU, SVR and ARIMA. According to authors Bi-LSTM spawns least MAE and RMSE values of 0.0070 and 0.0077 for deaths in China while ARIMA and SVR models yielded high prediction error.

Sengupta et al., (2020) conducted comparative analysis of statistical as well as machine learning models for forecasting the number of active cases. The models considered for the study are Linear Regression, Support Vector Regression and Polynomial Regression and LSTM (Long short-term memory). Statistical regression model with sigmoid function is surpassing all other models. Authors estimated expected day of peak of actives cases with logistic regression..

Yadav et al., (2020) conducted broad study with multiple objectives of Predicting the spread , analyzing the growth rates, types of mitigation, transmission rate of the virus, relation between corona virus and weather conditions using Support Vector Regression method . SVR has demonstrated its efficiency as well as accuracy over other models. Neve et al., (2020) in the study conducted using Support vector regressions, Polynomial regression, Random forest regression and Bayesian regression with large data of 280 days. Support Vector Regression has yielded accurate results. Due to highly correlated feature variables, Random forest regression results in over fitting followed by poor Bayesian regression. Authors concluded that Support Vector Regression is best among all methods.

Shantala et al., (2019) used deep neural network with transfer learning for prediction of fungal deceases in agricultural plants. Shrinivas et al., (2019) carried a study on heart disease detection using logistic regression and backpropogation neural network

III. METHODOLOGY

The study is conducted using three different regression models i.e. linear regression, Support vector regression and Polynomial regression. Methodology is shown in Figure 3.

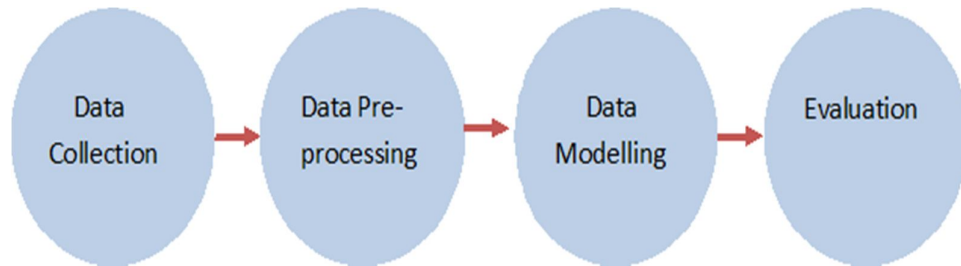


Fig. 3. Proposed methodology for forecasting the confirmed cases

Dataset: Authors made use of Dataset from a publicly available repository operated by the Johns Hopkins University Center for Systems Science and Engineering. Data from 1st April 2020 to 30th November 2020 is considered for training the regression models.

Data Preprocessing: The live data is converted into array format using NumPy library. The field's latitude and longitude which are not necessary are removed.

Data modeling and Evaluation: Study conducted using three models linear regression, Support Vector Regression and Polynomial regression.

A. MODEL 1: Linear Regression

Regression task is performed using Linear Regression, a supervised machine learning algorithm .Given the independent variable X(which is input) , linear regression performs the task to predict the dependent variable value Y(which is output). Linear Regression mainly uses fit- intercept parameter which is by default set to TRUE. Next, the model will fit to the X_train and Y_train dataset and predict the result using test data. The data till 15th November 2020 is used for modeling, and then the model is used to forecast the confirmed cases. Figure 4 shows prediction of new confirmed cases using linear regression.

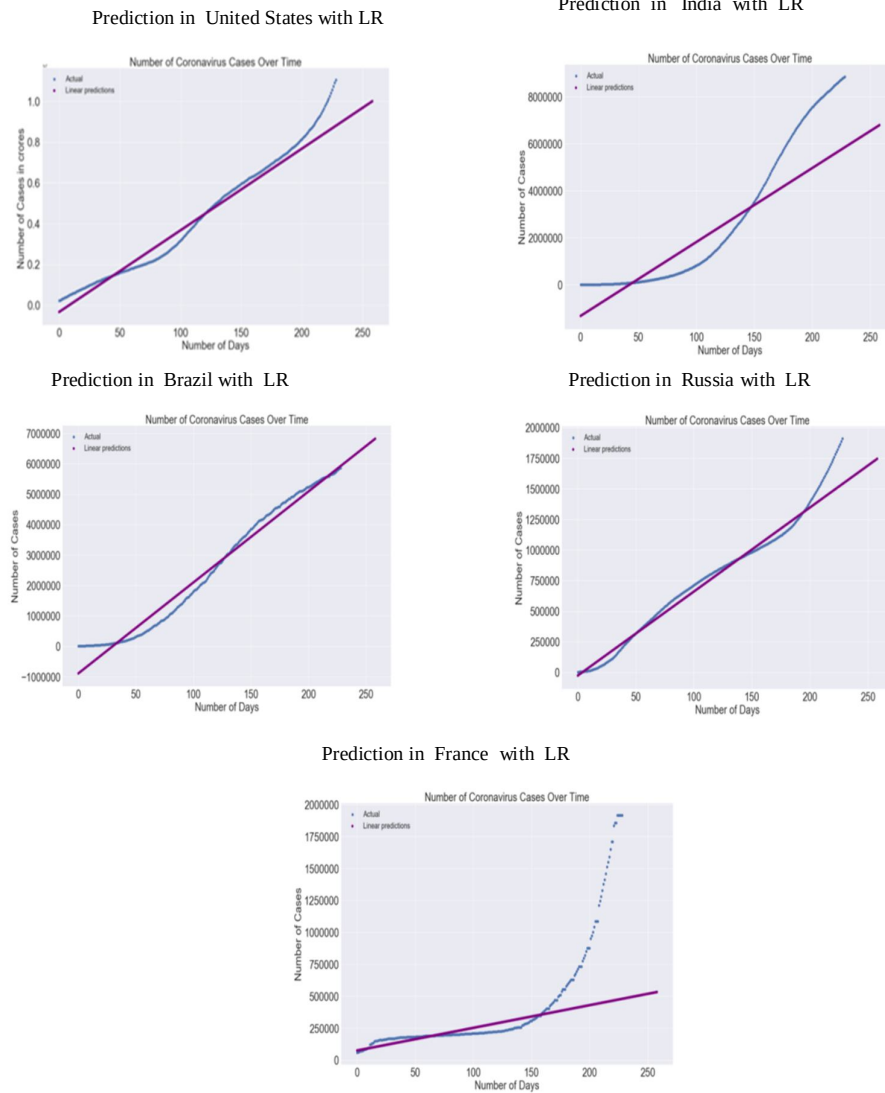


Fig 4. Training and forecasting the number of confirmed cases for the top five countries using LR

B. MODEL 2: SVR Regression

Parameter tuning is an important step in tuning the performance of algorithms. There are two methods for selecting the optimal parameters and they are Grid search and randomized search. Randomized search takes less time. Authors have used randomized search. Figure 5. Shows the 5 cross validation procedures for optimal parameter searching.

The optimal parameters are, Polynomial kernel function used, C: 1, Gamma: 1, Epsilon: 0.01 and Shrinking: True. After finding the optimal parameters, the training till June 30th is used to train the SVR and authors have used this data to predict the number of confirmed cases for the period 16 November to 15 December 2020.

Authors conducted study with two goals, to predict number corona confirmed cases country wise. The United States, Brazil, India, Russia and France emerged as top five countries during forecasting. With this forecasting, authors conducted experimentation on forecasting the confirmed cases in these five countries.

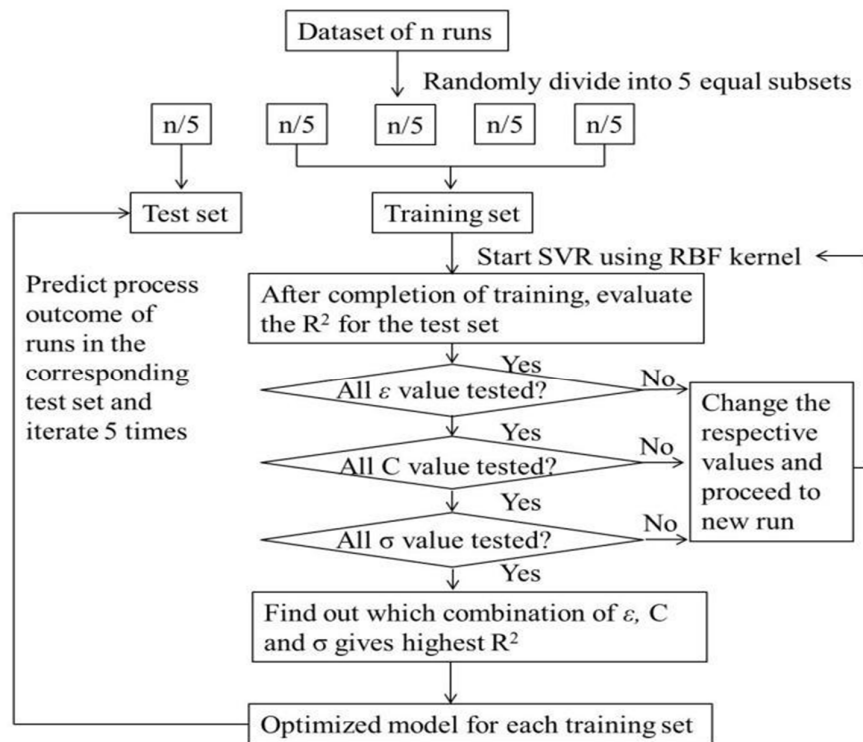
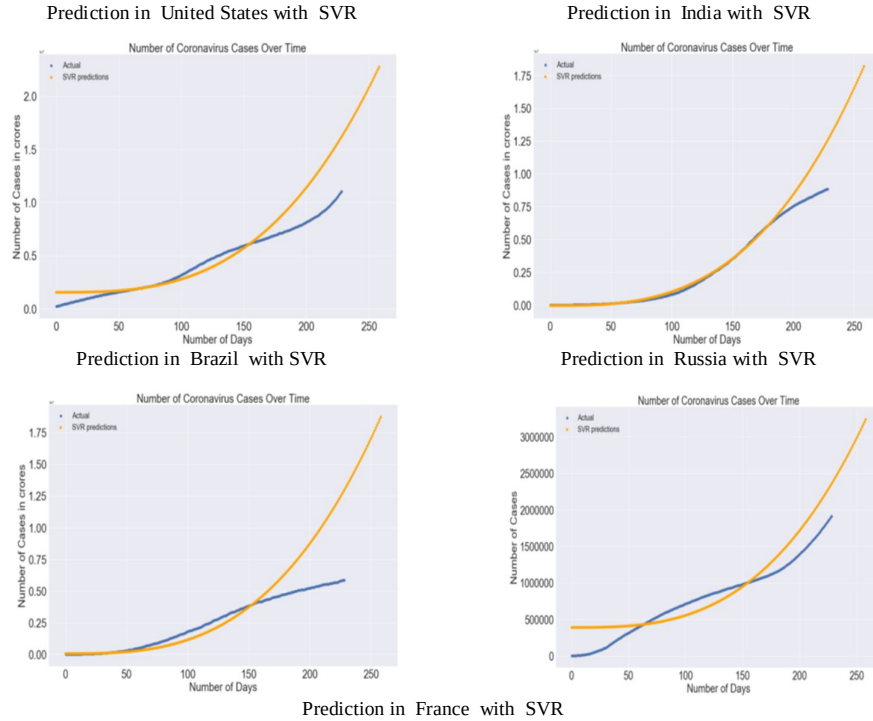


Fig 5. Support vector machine hyper parameter searching grid search



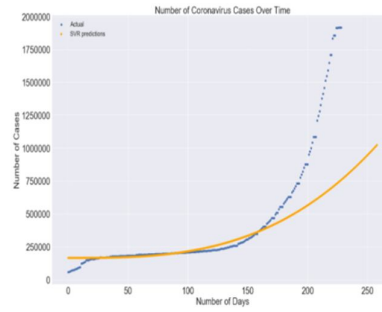


Fig 6. Training and Forecasting the number of cases for the top five countries using SVR

C. MODEL 3: Polynomial Regression

Using low value of error it is very difficult to fit a linear regression line, so we can try using the polynomial regression to fit a polynomial line which will help achieve a minimum error or minimum cost function. Polynomial regression fits a polynomial equation on the data with a curvilinear relationship between the target variable and the independent variables. It is a special case of Linear regression. Based on the relationship between target and predictor, we can choose the degree of polynomial. The 1-degree polynomial is a simple linear regression; therefore, the value of degree must be greater than 1. The complexity of the model increases with the increasing degree of the Polynomial, hence the value of n must be precisely chosen. If this value is low, then the model won't be able to fit the data properly and if high, the model will over fit the data. The data till 30th November 2020 is used for modeling, and then the model is used to forecast the confirmed cases.

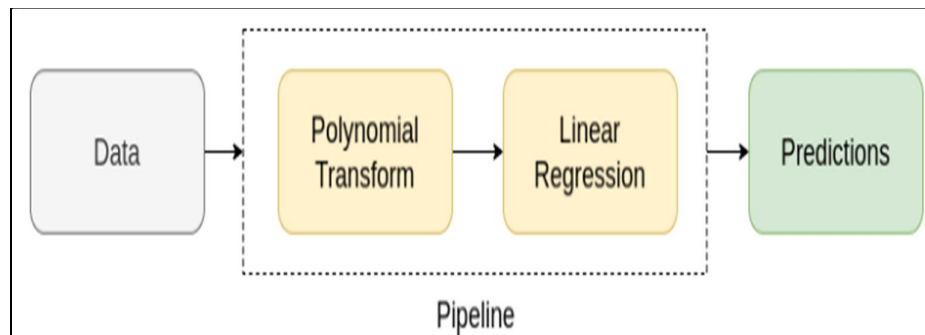
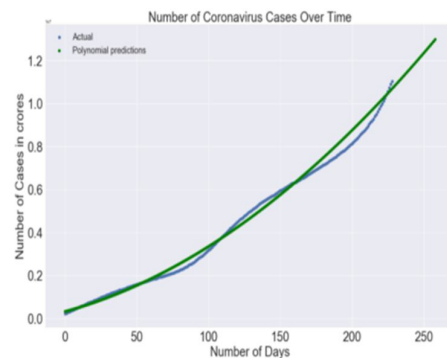
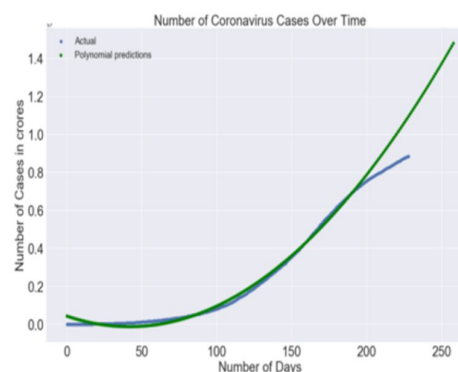


Fig 7. Polynomial regression process

Prediction in USA with Polynomial regression

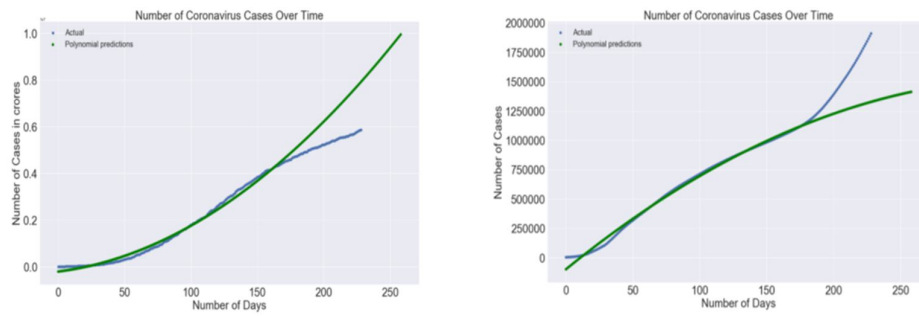


Prediction in India with Polynomial regression



Prediction in Brazil with Polynomial regression

Prediction in Russia with Polynomial regression



Prediction in France with Polynomial regression

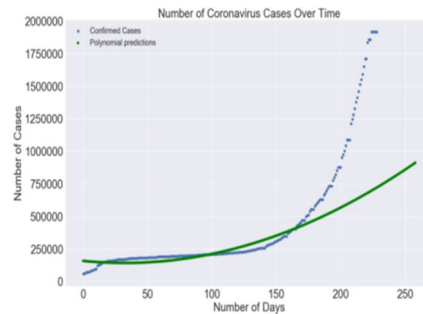
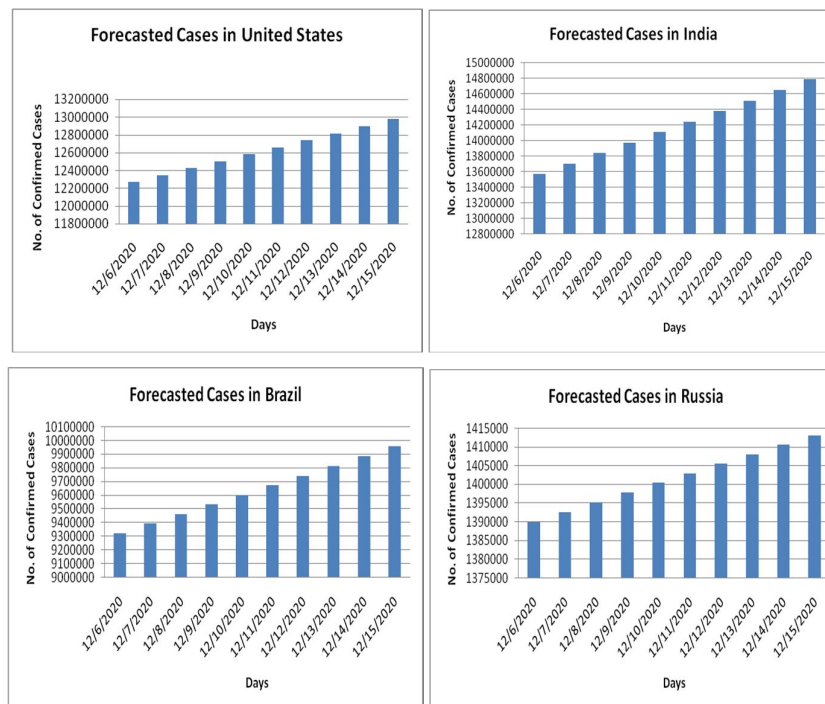


Fig 8. Training and Forecasting the number of cases for the top five countries

Forecasting of Covid Cases using Polynomial Regression

Table 1 shows R^2 error obtained with three regression models. R^2 measures how much of the variation in y is described by the model. It's a measure used to measure the performance of regression model. Polynomial model has the highest R^2 value, hence it is being used to forecast the confirmed cases from December 6th to December 15th 2020. Figure 9 shows forecasting confirmed cases using polynomial regression.



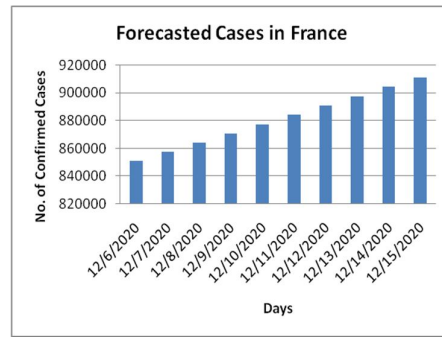


Fig 9 .Forecasting the number of confirmed cases using Polynomial regression in top five countries

TABLE I .R² ERRORS FOR LINEAR, SUPPORT VECTOR REGRESSION, POLYNOMIAL

Countries	Linear	SVR	Polynomial
United States	0.7765780049654872	0.83403401598029845	0.9850249031005904
India	0.7286877513778904	0.84598937965293941	0.948078764513735
Brazil	0.8591444286328828	0.81091318746468388	0.936680791748393
Russia	0.7199915129437456	0.85764399915129437	0.9186130441856067
France	0.690584629233765	0.8092356541196548	0.8994599270702346

IV. CONCLUSION

In this paper, authors have conducted a study on forecasting of highly infectious disease COVID-19 with regression models linear, SVR and Polynomial. These models were trained using day wise cumulative count of past confirmed cases. The tuned models are then used for future forecasting. These models are used to predict the infections in the top 5 countries. The data till 30 November 2020 is being used to tune, and the model is used to forecast the confirmed cases in the next seven days. Disease forecasting is highly challenging due to changing behavior of humans, uncertainties with data.

REFERENCES

- [1] Bandyopadhyay, Samir Kumar, and Shawni Dutta. "Machine learning approach for confirmation of covid-19 cases: : Positive, negative, death and release." *medRxiv* (2020).
- [2] Bing Zhang, Huihui Ren, Guoyan Huang, Yongqiang Cheng and Changzhen Hu "Predicting blood pressure from physiological index data using the SVR algorithm", Zhang et al. *BMC Bioinformatics* (2019) 20:109
- [3] Car, Zlatan, Sandi Baressi Segota, Nikola Andelić, Ivan Lorencin, and Vedran Mrzljak. "Modeling the Spread of COVID-19 Infection Using a Multilayer Perceptron." *Computational and Mathematical Methods in Medicine* 2020
- [4] Ellis Kobina Paintsil ,Akoto Yaw Omari-Sasu , Matthew Glover Addo, and Maxwell AkwasiBoate "Analysis of Haematological Parameters as Predictors of Malaria Infection Using a Logistic Regression Model: A Case Study of a Hospital in the Ashanti Region of Ghana".
- [5] H. ROOPA AND T. ASHA, "A Linear Model Based on Principal Component Analysis for Disease Prediction", August 14, 2019, Digital Object Identifier 10.1109/ACCESS.2019.2931956.
- [6] Kumar, Pavan, Ram Kumar Singh, Chintan Nanda, Himangshu Kalita, Shashikanta Patrairiya, Yagya Datt Sharma, Meenu Rani, and Akshaya Srikanth Bhagavathula. "Forecasting COVID-19 impact in India using pandemic waves Nonlinear Growth Models." *medRxiv* (2020).
- [7] Mr. Anirudhh Ravi , Dr. D. Devaraj , Mr. Varun Gopal, Mr. Pranav Chandran, Dr. J. Preetha Roselyn, and Ms. R. Sai Madhura "Detection of Infectious Disease using Non- Invasive Logistic Regression Technique ".

- [8] Ogundokun, Roseline Oluwaseun, and Joseph Bamidele Awotunde. "MACHINE LEARNING PREDICTION FOR COVID 19 PANDEMIC IN INDIA." *medRxiv* (2020).
- [9] Pandey, Gaurav, Poonam Chaudhary, Rajan Gupta, and Saibal Pal. "SEIR and Regression Model based COVID-19 outbreak predictions in India." *arXiv preprint arXiv:2004.00958* (2020).
- [10] Rustam, Furqan, Aijaz Ahmad Reshi, Arif Mehmood, Saleem Ullah, Byungwon On, Waqar Aslam, and Gyu Sang Choi. "COVID-19 Future Forecasting Using Supervised Machine Learning Models." *IEEE Access* (2020).
- [11] Shahid, Farah, Aneela Zameer, and Muhammad Muneeb. "Predictions for COVID- 19 with deep learning models of LSTM, GRU and Bi-LSTM." *Chaos, Solitons & Fractals* 140 (2020): 110212.
- [12] Sengupta, Sohini, Sareeta Mugde, and Garima Sharma. "Covid-19 pandemic data analysis and forecasting using machine learning algorithms." *medRxiv* (2020).
- [13] Yadav, Milind, Murukesan Perumal, and M. Srinivas. "Analysis on novel corona virus (COVID-19) using machine learning methods." *Chaos, Solitons & Fractals* 13 (2020): 110050.
- [14] Neve, Dishita, Honey Patel, and Harsh S. Dhiman. "On Modeling of COVID-19 for the Indian Subcontinent using Polynomial and Supervised Learning Regression." *medRxiv* (2020).
- [15] Giraddi, Shantala, Shrinivas Desai, and Anupama Deshpande. "Deep Learning for Agricultural Plant Disease Detection." In *ICDSMLA 2019*, pp. 864-871. Springer, Singapore, 2020.
- [16] Desai, Shrinivas D., Shantala Giraddi, Prashant Narayankar, Neha R. Pudakalakatti, and Shreya Sulegaon. "Back-propagation neural network versus logistic regression in heart disease classification." In *Advanced computing and communication technologies* pp. 133-144. Springer, Singapore, 2019.
- [17] Shamarao Jahagirdhar Divyashree Naik , Ramsha Shaikh , Sahana Shetti , Praveen K , Suvarna G Kanakaraddi "Detection and Quantification of Disease in Cabbage using Clustering and RGB Colour Features", *International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE)*, Vol14, issue2, pp-194-199.