International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT206541

Statistical Analysis and Prediction of COVID-19 outbreak in India using Machine Learning

Akshar Patel[1], Dweepna Garg[2]

[1,2]Department of Computer Engineering, Devang Patel Institute of Advance Technology and Research (DEPSTAR), Faculty of Technology and Engineering (FTE), Charotar University of Science and Technology (CHARUSAT), Changa, Anand, Gujarat, India

ABSTRACT

Article Info
Volume 6, Issue 5
Page Number: 199-207
Publication Issue :
September-October-2020

Coronavirus disease globally known as COVID-19 is triggered by SARS-COV2. It is the predominant cause of an extremely dangerous disease that has bothered global health security. It is proposed that COVID-19 might be zoonotically based on the high number of people exposed in Wuhan City, China, to the wet animal market[1]. COVID-19 is a severe acute respiratory disease, transmitted by respiratory secretions and communication paths, as of WHO reports. The disease is spreading throughout the world at a faster pace. The first instance of COVID-19 was firstly discovered and found in Wuhan, Hubei Province, China in December 2019[1]. This paper analyses the outbreak of this disease until June 22, 2020, for India and other top major affected nations and also predictions were made regarding the number of cases for India over the next 17 days i.e from 23 June 2020 to 9 July 2020. Linear Regression model, Support Vector Machine Regressor (SVM) model, Autoregressive Integrated Moving Average (ARIMA) model and Facebook's Prophet model were used for prediction based on the Kaggle downloaded dataset with data collected from January 22, 2020, to June 22, 2020. By 22 June 2020, the disease has spread across more than 200 countries, reporting 12,322 confirmed cases, 45,26,333 recovered cases and 4,72,171 COVID-19 deaths. Assessment of this epidemic allows the Government to take the appropriate steps to curb the threat of this global pandemic.

Article History
Accepted : 05 Oct 2020
Published : 12 Oct 2020

Keywords: CoronaVirus, COVID-19, India, Linear Regression Model, Support Vector Machine Regressor (SVM) model, Autoregressive Integrated Moving Average (ARIMA) model, Facebook's Prophet Model, Prediction.

I. INTRODUCTION

At the end of December 2019, China's health authorities were worried by a number of cases of pneumonia of uncertain origin in Wuhan. On December 31, the Wuhan Municipal issued a warning and also a rapid response team was sent to Wuhan by the Chinese Center for Disease Control and

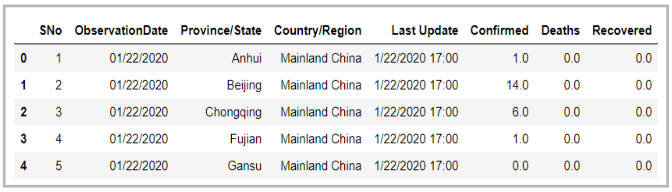
Prevention (CDC). World Health Organization (WHO) was also informed about the outbreak [2]. Fourth-one cases of unknown etiological pneumonia were identified in the provincial capital of Hubei, Wuhan City, on 31 December 2019, with a population of 11 million. It's been widely believed that the source of novel coronavirus may have originated from the Huanan Seafood Wholesale Market in Wuhan, and it has also been referred to as a zoonotic disease. The outbreak was declared an epidemic by the World Health Organization (WHO) on January 31, 2020, for public health emergencies concerns. WHO later declared it as a pandemic on 11 March 2020 [3]. The WHO initially declared this transmittable disease as Novel Coronavirus-Infected Pneumonia (NCIP), and the name 2019 novel coronavirus (2019-nCoV) was assigned to the virus. On 11 Feb 2020, COVID-19 was formally designated by the (WHO). COVID-19 stands for coronavirus disease of 2019[4]. The disease has caused various degrees of illness around the world. The patient typically exhibits various symptoms including fatigue, cough, sore throat, breathlessness, tiredness, and malaise. Usually, the virus spreads rapidly from individuals to individuals through respirational droplets released during coughing and sneezing. If people are symptomatic, it is considered most infectious while transmission can be possible before symptoms occur in patients. Period between discovery and beginning of symptoms is usually between 2-14 days, with an average of 5 days. To date, however, no vaccine was produced for the disease. A chimpanzee adenovirus vector (ChAdOx1), developed at the Jenner Institute in Oxford, has been identified as one of the most appropriate SARS-CoV-2 vaccine technology as it can produce strong immune response and cannot trigger an ongoing infection in the vaccinated individual [5]. Scientists have already begun clinical trials of the vaccine in South Africa. But currently, effective prevention steps include using the soap while washing the hands and covering the mouth while coughing, keeping a distance of 1 meter from other people as well as monitoring & maintaining self-isolation for 14 days for the individuals who find themselves contagious. The foremost COVID-19 case was registered in India on 30 January 2020 in Kerala. The Ministry of Health and Family Welfare has reported a total of 4,40,215 cases, 2,48,190 recoveries and 14011 deaths until June 22, 2020. On 19 March 2020, India's government under Prime Minister Narendra Modi announced a 21-day national shutdown, controlling and stopping the movement entire 1.3 billion Indian citizens as a protective and preventive measure against the COVID-19 pandemic [6]. Janta curfew was enforced on March 22 followed by implementation of a set of regulations in the affected regions of the country COVID-19. The lockdown was imposed when the figure of reported positive coronavirus cases in India was about 500. As the cases continued to rise and the first lockdown time was closing (14th April), state governments and other advisory committees suggested to prolonged lockdown until 3rd May. On May 1 India's government further prolonged the countrywide lockdown by two weeks until May 17. The government classified all the districts into three different zones based on the spread of the virus — green, red, and orange — with enforcing moderations according to the situation. On 17 May, the National Disaster Management Authority extended the lockdown further until 31 May. On 30 May, it was announced that the ongoing shutdown will continue to be extended in containment zones until 30 June, with services resuming phased from 8 June.

The task of researchers is to incorporate the relevant data and technology to better understand the virus and its characteristics, which can help with making the right decisions and practical action plan which will contribute to a broader image of drastic steps being taken in the future to build infrastructure, services, vaccines and restrain similar epidemics. The current research aims are as follows.

1. Top 10 worst affected countries in terms of their confirmed cases, recovered cases, death cases, mortality rate and recovery rate.
2. Predicting the number of COVID-19 cases for India for the next 17 (23 June 2020 to 9 July 2020) days using Machine Learning models like the Linear Regression Model, Support Vector Machine (SVM) model, Autoregressive Integrated Moving Average (ARIMA) model, Facebook's Prophet Model.
3. Comparison of the models listed above to find the most accurate one.

2. ANALYSIS AND CLASSIFICATION

2.1 Dataset: Time series data provided by Kaggle has been used for the empirical analysis of the result. The period of data is from 22/01/2020 to 22/06/2020[7]. The dataset includes columns like 'observation Date' which indicates the number of cases observed on that particular day, 'Province/State' and 'Country/Region' columns tells us about the place and country where the case was observed respectively and there are also 'Confirmed', 'Recovered' and 'Deaths' columns. The data includes a total of 223 countries in which the disease has spread.



	SNo	ObservationDate	Province/State	Country/Region	Last Update	Confirmed	Deaths	Recovered
0	1	01/22/2020	Anhui	Mainland China	1/22/2020 17:00	1.0	0.0	0.0
1	2	01/22/2020	Beijing	Mainland China	1/22/2020 17:00	14.0	0.0	0.0
2	3	01/22/2020	Chongqing	Mainland China	1/22/2020 17:00	6.0	0.0	0.0
3	4	01/22/2020	Fujian	Mainland China	1/22/2020 17:00	1.0	0.0	0.0
4	5	01/22/2020	Gansu	Mainland China	1/22/2020 17:00	0.0	0.0	0.0

Fig.1: COVID-19 Dataset

2.2 Information inferred from the dataset:

The figure 2. here illustrate the top 10 countries having the maximum number of confirmed cases which is inferred from the columns namely 'Country/Region' and 'Confirmed'. The USA has a maximum number of cases, which are nearly 2.3 million followed by Brazil having almost half cases as compared to the former. The number of confirmed cases depends on various factors like the number of people tested for the disease, age distribution, medical challenges and many more, however, the diagram does not illustrate the reasons behind the count of the confirmed cases.

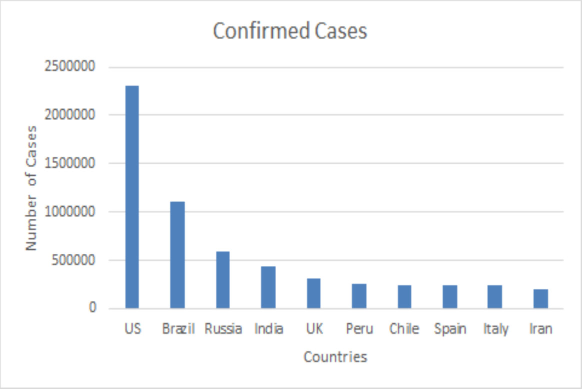


Fig.2 : Countries with the maximum number of confirmed cases

Figure 3. below depict the top 10 countries that registered the maximum number of deaths due to COVID-19. Risk of death associated with diseases is people with elderly age, poverty, respiratory diseases, diabetes and hypertension. Also, the regions with a high count of confirmed cases have reported more deaths as is the case for the USA, Brazil and some European nations.

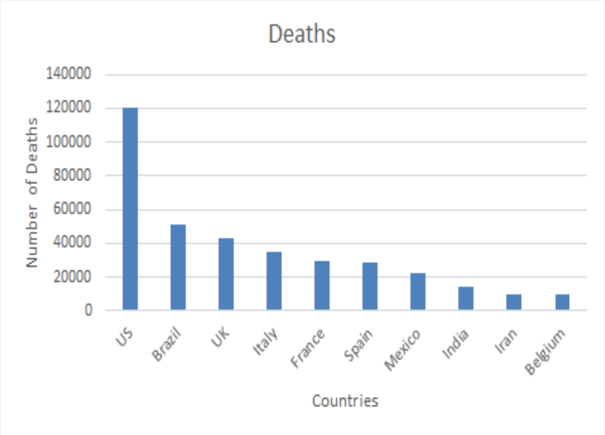


Fig.3: Countries with the maximum number of deaths

Figure 4. is a bar chart of nations having the highest death rate from COVID-19. The death rate is usually calculated as the number of known deaths in a period divide by total confirmed cases in that period.

According to estimates, Yemen appears to be highly vulnerable to the outbreak reporting the death rate above 25% which is almost double than the other top affected countries. Older age group population, poor healthcare system, imposing lockdown too late, failure to conduct rapid testing and isolating patients resulted in the higher death rate.

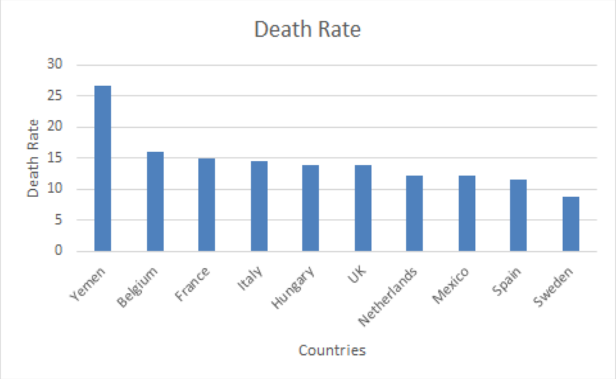


Fig.4: Countries with the highest death rate

Figure 5. reflects the countries which are having highest recovered cases from COVID-19. USA and Brazil are the are countries with the highest number of recovered cases despite having the maximum number of confirmed cases and deaths. Spain recorded the least number of recovered cases and the reason for this can be dense and largely elderly population.

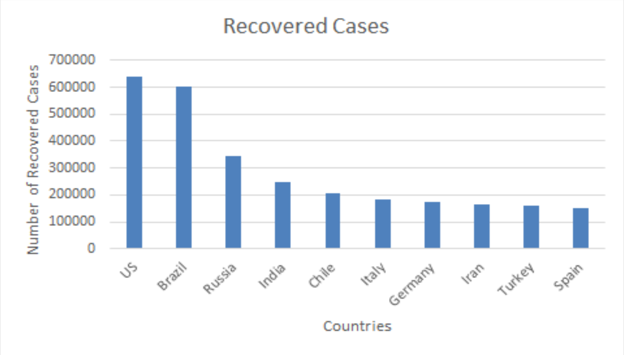


Fig.5: Countries with a maximum number of recovered cases

Figure 6. tells us about the recovery rate of the top 10 countries from COVID-19 where Iceland and New Zealand have recovery rates touching almost 99% and 98% respectively. Mainland China, where the worst of the pandemic appear to be over as the recovery rate climbs to 94%.

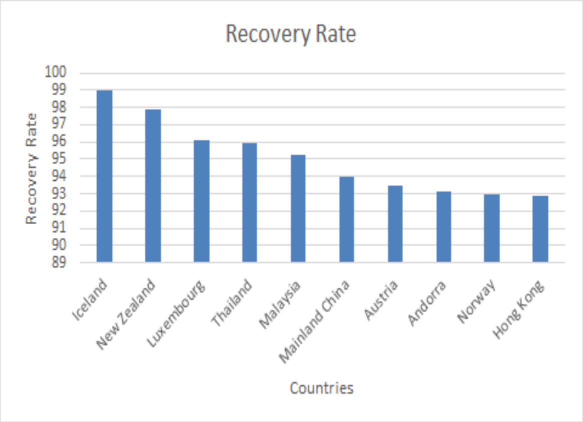


Fig.6: Countries with the highest recovery rate

2.3 Preprocessing:

Preprocessing of data involves the transformation of raw data into an understandable and usable format which can then be passed for training and testing purposes. Preprocessing of raw data is essential as it can have missing values, inconsistent values and a lot of redundant information. We need to fix these issues for more accurate output. Here we have performed data reduction by dropping the 'SNo' and 'Last Update' column as it does not provide any information and 'Province/State' column which contains too many missing values.

2.4 Training and Testing Models:

Training data is used to train our model by recognizing the patterns in data and learn from them, the validation data is used to evaluate the model, fit on training data for accurate and effective results and the test data is used to see how well the model can predict the result based on the given training data. After prediction, evaluation of our model can be done by comparing the predicted result and the actual output of testing data. Here our focus is to analyze and predict the number of confirmed COVID-19 patients for India by using the following models:

2.4.1 Linear Regression:

Linear Regression model of machine learning is based on supervised learning. Regression mockups an analytical target value, which is based on the independent variables. Generally, it is used to know the relationship between variables and forecasting. Different regression models vary in terms of the connection between the dependent and independent variables. The total number of independent variables are taken into consideration. Linear regression carries out the step of predicting a variable dependent value(y) based on a given independent variable(x). Thus, the method of regression finds a linear relation between x(input) and y(output) and is termed as Linear Regression. This model uses the value of intercept and slope to predict the output value. The equation below is the function of linear regression:

$$Y=a+bX \qquad (1)$$

Where 'Y' is the dependent variable, 'X' is the independent variable, 'b' is the slope of the line and 'a' is the intercept. Once we get the value of a and b, we can get a best-fit line. The model goals to predict 'Y' value by accomplishing the best-fit regression line so that the difference in the error value between the expected value and the true value is small and this cost function of linear regression is Root Mean Squared Error(RMSE) which is often used to calculate the difference between value predicted by the model and pragmatic (observed) value.

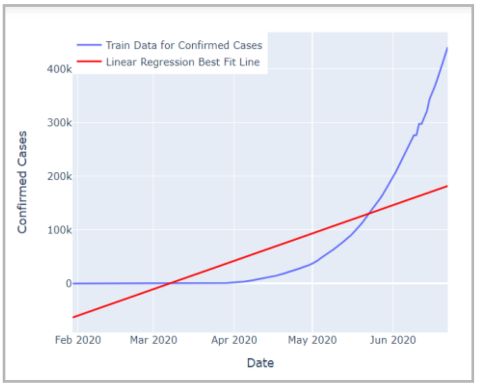


Fig.7: Training data for Linear Regression

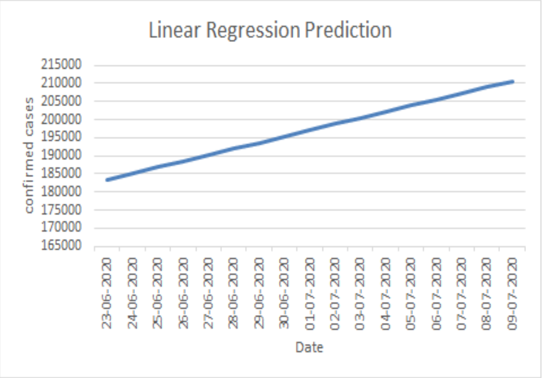


Fig.8: Predicted number of confirmed cases using Linear Regression

Figure 7. and Figure 8. gives information regarding the trained cases and number of predicted cases till 9 July 2020 respectively. However, the accuracy of this model is comparatively low, therefore, the predicted results are not the nearest to accurate probabilistic prediction. From the graph, we can understand that the number of cases is linearly increasing from around 183500 on 23 June 2020 to approximately 210500 on 9 July 2020.

2.4.2 Support Vector Machine regressor:

SVR uses the same basic concept as a classification algorithm called Support Vector Machine (SVM), but this model is used to forecast real values rather than a class. Support Vector Machine (SVM) recognizes the existence of non-linearity in the data and delivers a model capable for predictions. Support Vector Regression (SVR) operates on similar concepts as the classification of Support Vector Machine (SVM). A big advantage of consuming SVR is it's being a non-parametric procedure. The Support Vector Regression (SVR) output model is independent from the distributions of the dependent and independent variables underlying it. The Support Vector Regression (SVR) technique is based on kernel functions. The advantage of SVR is that it allows a non-linear model to be built without altering the explanatory variables, thereby helping to better understand the resulting model. The basic concept behind Support Vector Regression (SVR) is not to be concerned with the forecast provided that the error(ϵi) is less than some amount. This is called the Maximum Margin principle. The concept of maximum edge allows SVR to be regarded as a problem of convex optimization. You may also penalize the regression using a cost parameter, which is useful to prevent over-fit. Supporting the Vector Regression (SVR) is an advantageous technique that gives the user a high degree of versatility with regard to the distribution of underlying variables, the relation between independent and dependent variables and time control.

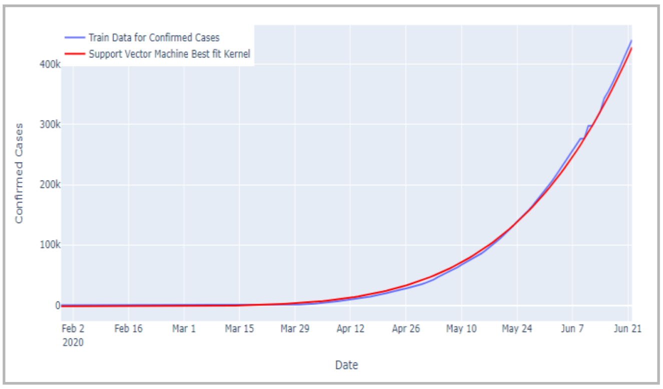


Fig.9: Training data for SVM

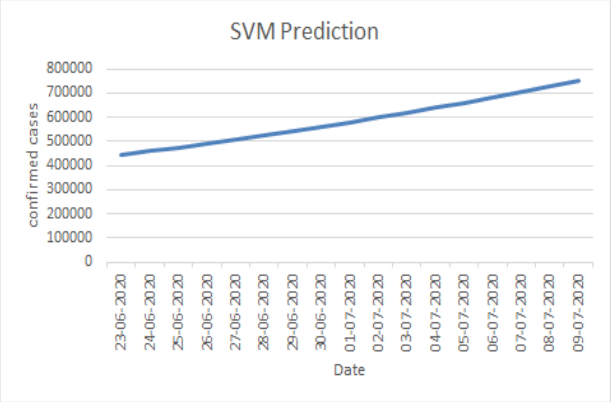


Fig.10: Predicted number of confirmed cases using SVM

Figure 9. denotes the train data for confirmed cases and figure 10. represents the confirmed cases expected for 23 June 2020 to 9 July 2020. It can be calculated from figure 10. that there will be a rise of about 300000 cases in the next 17 days.

2.4.3 Auto ARIMA:

An Autoregressive Integrated Moving Average (ARIMA) model is generally used in statistics and in the analysis of time series. This model is either tailored to time series data for greater understanding of the data or to forecast future series points. For certain cases, ARIMA models are implemented where data indicate signs of non-stationarity wherein an initial differentiation stage (corresponding to the "integrated" portion of the model) can be added one or more times to remove non-stationarity. ARIMA model has three components viz. AR (autoregressive term), I (differencing term), and MA (moving average term). The term 'AR' denotes the previous values used to determine the next value. In ARIMA, the 'AR' term is determined by the parameter 'p and the value of 'p' is set by using the PACF plot. The term 'MA' is used to describe the number of previous forecast errors used to estimate future values. In ARIMA the parameter 'q' reflects the word MA. ACF plot is used to determine the right value for 'q.' Differentiating order determines the number of times the process of the distinction is achieved on series to make it stationary. Tests such as ADF as well as KPSS can be used to assess whether the series is stationary and is capable to help defining the value 'd'. Here we have used the Auto ARIMA model as it automatically selects the value of 'p' and 'q' which was a very time-consuming task in a simple ARIMA model. The best value we got for our model ARIMA (p,d,q) is (2,2,3).

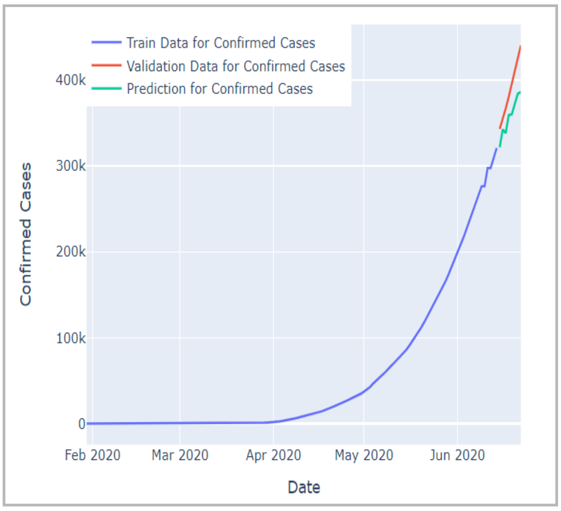


Fig.11: Train, validate and predicted data of total confirmed cases using ARIMA

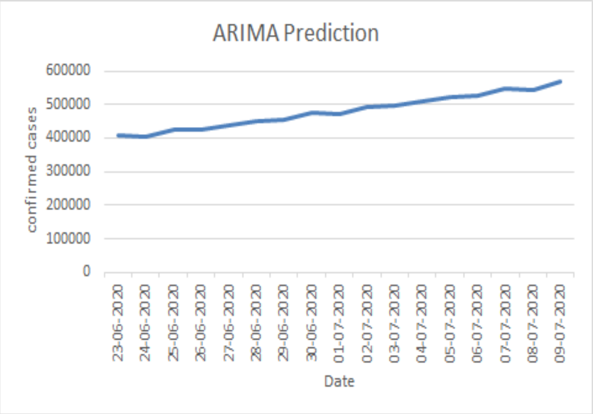


Fig.12: Predicted number of confirmed cases using ARIMA

Figure 11. represents the train data, validation data and prediction data of confirmed cases and figure 12. represent the predicted confirmed cases for the period from 23 June 2020 to 9 July 2020. From figure 12. it can be studied that there is an increase of approximately 160000 cases in the next 17 days.

2.4.4 Facebook's Prophet:

The prophet is a time-series data forecasting technique grounded on an additive model wherein non-linear patterns are fitted with annual, weekly, and regular seasonality, in addition with holiday results. This works best with time series that have powerful seasonal effects. The prophet is resilient in the absence of data, it changes the pattern and typically manages outliers well. The model makes a rational, reliable forecast much simpler. The three main components of the model are trend, seasonality, and holidays and the equation formed by combining those components is given by

$$y(t) = g(t) + s(t) + h(t) + \epsilon t\ . \qquad (2)$$

Here g(t) stands for trend function that models non-seasonal changes in the time series values, s(t) is the periodic changes and h(t) represents the effect on potentially irregular schedules of the holidays that occur over one or more days. The error term ϵt is any idiosyncratic changes that the model does not handle. This model is very easy to fit and that enables the analyst to interactively examine other model requirements.

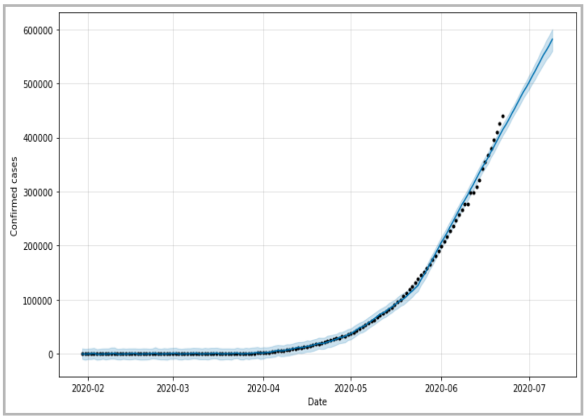


Fig.13: Trained and predicted data on the number of confirmed cases using Facebook's Prophet

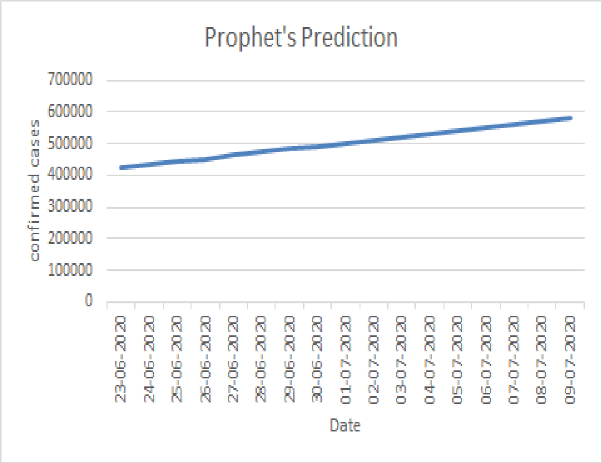


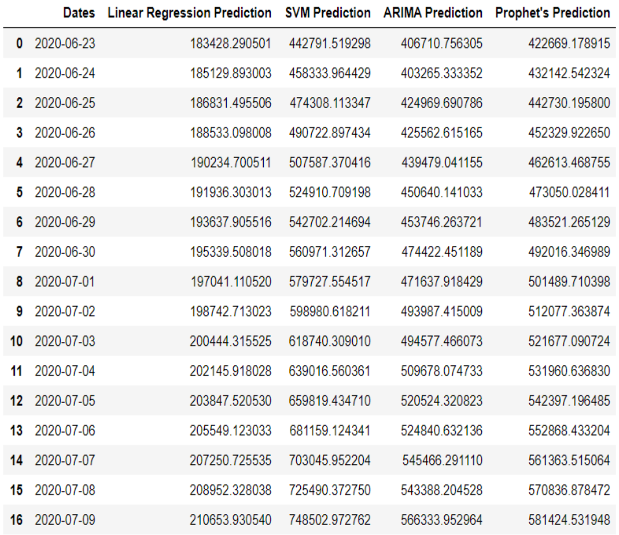
Fig.14: Predicted number of confirmed cases using Facebook's Prophet

Figure 13. and Figure 14. represents the total number of confirmed cases till 22 June 2020 and predicted cases during the period of the next 17 days. From figure 14. it can be studied that on 23 June 2020 the predicted cases are around 420000 and increase up to 580000 till 9 June 2020.

3. EMPIRICAL RESULT ANALYSIS

Table 1. and Figure 15. mentioned below compares the predicted result of all the models which are described in this paper. As depicted in figure 15. the results of SVM prediction, ARIMA prediction and Prophet's prediction are more or less similar however the result of a linear regression model is drastically different.

Table 1: Comparison of predicted values of 4 models



	Dates	Linear Regression Prediction	SVM Prediction	ARIMA Prediction	Prophet's Prediction
0	2020-06-23	183428.290501	442791.519298	406710.756305	422669.178915
1	2020-06-24	185129.893003	458333.964429	403265.333352	432142.542324
2	2020-06-25	186831.495506	474308.113347	424969.690786	442730.195800
3	2020-06-26	188533.098008	490722.897434	425562.615165	452329.922650
4	2020-06-27	190234.700511	507587.370416	439479.041155	462613.468755
5	2020-06-28	191936.303013	524910.709198	450640.141033	473050.028411
6	2020-06-29	193637.905516	542702.214694	453746.263721	483521.265129
7	2020-06-30	195339.508018	560971.312657	474422.451189	492016.346989
8	2020-07-01	197041.110520	579727.554517	471637.918429	501489.710398
9	2020-07-02	198742.713023	598980.618211	493987.415009	512077.363874
10	2020-07-03	200444.315525	618740.309010	494577.466073	521677.090724
11	2020-07-04	202145.918028	639016.560361	509678.074733	531960.636830
12	2020-07-05	203847.520530	659819.434710	520524.320823	542397.196485
13	2020-07-06	205549.123033	681159.124341	524840.632136	552868.433204
14	2020-07-07	207250.725535	703045.952204	545466.291110	561363.515064
15	2020-07-08	208952.328038	725490.372750	543388.204528	570836.878472
16	2020-07-09	210653.930540	748502.972762	566333.952964	581424.531948

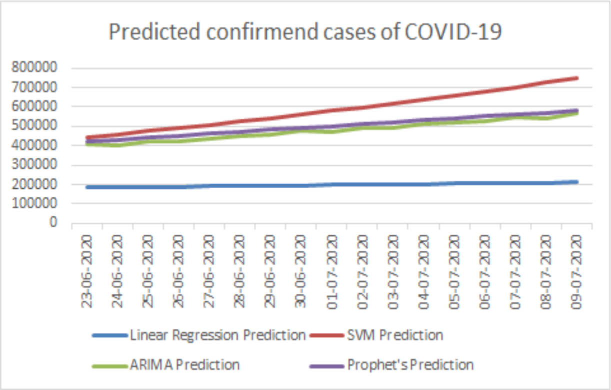
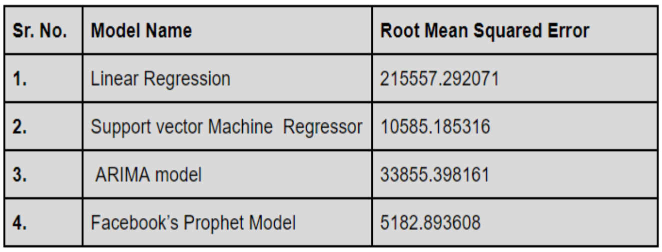


Fig.15: Graphical representation of predicted values of all models

Table 2. below contains the root mean squared error value of all the models described in this paper. The root mean square error (RMSE) is the standard deviation of the residuals and describes the errors made in the prediction. The measurement of how far the points are from the regression lines are termed as residuals. As we can see in the table the Prophet's model has the least root mean square value, therefore, we can conclude that Facebook's Prophet Model has the highest accuracy and will give the most accurate results. Support Vector Machine and ARIMA follows the Prophet model while comparing the accuracy and the Linear Regression model has the highest RMSE value therefore, the predictions are comparatively inaccurate.

Table 2: Comparison of RMSE values of all models



Sr. No.	Model Name	Root Mean Squared Error
1.	Linear Regression	215557.292071
2.	Support vector Machine Regressor	10585.185316
3.	ARIMA model	33855.398161
4.	Facebook's Prophet Model	5182.893608

4. CONCLUSION:

This study provided a detailed review of the COVID-19 outbreak in India and some worst-affected nations in the world. While we can conclude from this study that COVID-19 disease is complex and is likely to evolve rapidly. This paper has covered many aspects like Top 10 countries having the highest number of confirmed cases, recovered cases and deaths. Countries having a high mortality rate and a recovery rate was also analyzed. The paper also presents predictions for the number of confirmed cases in India for the next few days. For the prediction purpose machine learning models namely linear regression, Support Vector Machine regressor (SVM), ARIMA model, Facebook's Prophet model were used. The data analyzed that the number of Confirmed cases in India are around 4,40,000. By using the above models, we have predicted confirmed cases for the next 17 days starting from 23 June 2020 to 9 July 2020. We have measured the accuracy of our models using RMSE values. Since the RMSE value of facebook's prophet model is minimal compared to other models therefore it is the most accurate model. Further, by looking at the pattern, the number of cases will progressively increase and chances for a community transmission of disease will surge up.

II. REFERENCES

[1]. Hamid S., Mir M.Y., Rohela G.K. Novel coronavirus disease (COVID-19): a pandemic (epidemiology, pathogenesis and potential therapeutics) New Microbes New Infect.

[2]. Novel Coronavirus (2019-nCoV) SITUATION REPORT - 1 21 JANUARY 2020 https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200121-sitrep-1-2019-ncov.pdf

[3]. Timeline of WHO's response to COVID-19 https://www.who.int/news-room/detail/29-06-2020-covidtimeline

[4]. Article : https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-%28covid-2019%29-and-the-virus-that-causes-it

[5]. About the Oxford COVID-19 vaccine: https://www.research.ox.ac.uk/Article/2020-07-19-the-oxford-covid-19-vaccine

[6]. Article: https://www.nytimes.com/2020/03/24/world/asia/india-coronavirus-lockdown.html

[7]. COVID-19 dataset: https://www.kaggle.com/datasets

Cite this article as :

Akshar Patel, Dweepna Garg, "Statistical Analysis and Prediction of COVID-19 outbreak in India using Machine Learning", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 6 Issue 5, pp. 199-207, September-October 2020. Available at doi : https://doi.org/10.32628/CSEIT206541
Journal URL : http://ijsrcseit.com/CSEIT206541