International Journal on Emerging Technologies **11**(2): 000-000(2020)

ISSN No. (Print): 0975-8364
ISSN No. (Online): 2249-3255

Predicting the Rate of Growth of the Novel Corona Virus 2020

Geetika Vashisht[1] and Ravi Prakash[2]
[1]Assistant Professor, Department of Computer Science, Delhi University, Delhi, India.
[2]Student, Department of Computer Science, Delhi University Delhi, India.

(Corresponding author: Geetika Vashisht)
(Received 16 March 2020, Revised 02 April 2020, Accepted 06 April 2020)
(Published by Research Trend, Website: www.researchtrend.net)

ABSTRACT: The deadly pandemic of 21st century has spread its wings across the globe with an exponential increase in the number of cases in many countries. The novel corona virus, responsible for the pandemic originated from Hubei province of China and the highly contagious nature of the virus soon crippled the world. This work presents a framework to predict the growth rate of the corona virus disease in the upcoming week in China and henceforth predict the rate of spread of the disease in the upcoming week.Using available data about the confirmed cases, death cases and recovery cases, this study defines the status of the disease in terms of Total Active Cases percentage till a given date, cumulatively.

Keywords: Corona, Covid-19, pandemic, SVM-k, Linear & Polynomial Regression algorithms, RMSE, R^2, kNN.

Abbreviations: RMSE, Root Mean Squared Error; R^2, R-Squared; kNN, k-Nearest Neighbor; SVM, Support Vector Machine.

I. INTRODUCTION

A virus that badly hit the world in mid-November 2019 is thought to have originated from an animal-to-human spillover event linked to seafood and live-animal markets. The virus derived its name 'corona' from the Latin word corona that means crown in English. Since a similar kind of virus was noticed in 2002, this particular virus is termed as novel corona virus. The infection originated in Wuhan, China. Although laboratory testing for corona virus quickly ramped up in China, information on individual patients remain scarce and official datasets have not been made publicly available. Patient-level information is important to estimate key time-to-delay events such as the incubation period and interval between symptom onset and visit to a hospital, analyze the age profile of infected patients, reconstruct epidemic curves by onset dates and infer transmission parameters but individual-level data is not available. Considering this, the parameters considered for this work is not based on the details of the individual patients but on the total number of confirmed cases, total number of deaths and the total number of recovered cases. The aim of this paper is to analyze that how the status of the pandemic will change over a period of a week in China and hence how long it might take China to have control over this disease.

This paper is divided into six sections, and the outline of each section is as follows:

– *Contribution of the work*: To address the motivation and the contribution of this paper.
– *History of pandemics:* To present the background knowledge of the pandemics.
– *Definition and formulation:* To describe in detail the key terms and the related work.
– *Data and methodology:* To introduce the research methods, algorithms and procedures.
– *Results and Discussions*: To explain the quantified results while discussing the prediction of the rate of spread of the disease.
– *Conclusions and Future Work:* To summarize the achievements of this research and possible applications in the future.

II. CONTRIBUTION OF THE WORK

Tracking the trend of pandemics such as the novel corona virus disease using machine learning algorithms supports the government and health organizations to implement several strategic decisions in time that can be a catalyst in saving lives worldwide. This paper contributes to the study of the trend of the growth pattern of the disease by analysing the time-series data. This estimate can help the country to manage and plan its resources well such as the likelihood of the requirement of new beds or hospitals within the next few days. The study can also aid in identifying the places where the focus has to be driven first together with an estimate of more precise information for estimating test kits, other medical equipment and similar such considerations. The country considered in this work is China though the model can be applied to other countries as well.

III. HISTORY OF PANDEMICS

The pandemics have hit each century badly. The notorious ones are listed in this section.

Great Plague of Marseille: **1720**

As per the records Great Plague of Marseille started when a ship called Grand-Saint-Antoine docked in Marseille, France, carrying a cargo of goods from the eastern Mediterranean. It continued for the next three years, killing up to **30%** of the population of Marseille.

First Cholera Pandemic: **1820**

In 1820 the *First Cholera Pandemic* occurred, in Asia, affecting countries, mainly Indonesia, Thailand and the Philippines. This pandemic has also killed about **1,00,000** as declared officially. It was a bacterial infection caused due to the contaminated water.

Spanish Influenza/Flu: **1920**
Spanish Flu was the most recent and the most unrelenting pandemics. It has infected about half a billion people and killed **100 million**. The Spanish flu holds the official record for the deadliest pandemic officially recorded in history.
Novel Corona virus Disease: **2020**
In the 21st century, novel corona virus disease has appeared as the most severe pandemic. The disease is discussed in detail in the next section.

IV. DEFINITION AND FORMULATION

A. Covid-19- a close view

On Feb. 11th 2020, the World Health Organization (WHO) gave the disease an official name: COVID-19. WHO has also declared the COVID-19 as a pandemic [6, 10]. During November 2019 - a severe viral infection was noticed in **Wuhan**, a city in Hubei provinces of China. On November 17th 2019, the first case of this infection was reported. Doctors initially took it for normal fever or cold but when a wide range of patients reported similar kind of symptoms, a doctor, **Dr. Li Wenliang** of Chinese Academy of Sciences (CAS) Lab, claimed that it is a type of *severe acute respiratory syndrome*, spreading through a new corona virus. In January, Dr. Li himself fell a prey to this virus. As per the sources, in 2003, the same lab found first deadly SARS Corona virus, leading to 813 casualties, all over the world, within two months! Initially, it was named as "Wuhan Virus". But, as it started spreading rapidly from Hubei to the whole mainland of China, its name got replaced by term "China Virus".
The disease is spreading rapidly as it has the Reproduction Number (R-naught) 2.5 [8, 9, 12]. In more generalized language, on an average, from every COVID-19 patient, virus is getting transmitted to at least 2.5 fit persons. For Ebola, this number is approximately 1.7-1.9 [14]. In order to have control on any epidemic, the reproduction number has to be decreased as much as possible. Scientists are still trying to create the vaccine as soon as possible, but for now, no final cure is up.

B. Symptoms

The severe acute respiratory syndrome *COVID-19* **(SARoS-CoV-19)** is spread by the novel corona virus that tends to attack the respiratory system of the patient. Though research is still going on, some of the most common symptoms of COVID-19, known so far are fever, dry cough, fatigue, shortness of breath and nasal congestion. These are the symptoms that are reported by most of the patients. Apart from these, as per the data from China, few patients also found to develop aches and getsniffles while some didn't exhibit any symptoms at all.

C. Related Work

Since the COVID-19 is spreading at a lightning speed, many researchers and organizations worldwide are working on the same. Remarkable work in this area is presented in this section.
The Situation Report –22 of World Health Organization [11] came up with the official name of disease [10], presented a detailed report on *COVID-19* and declared the disease a global pandemic. The symptoms and other basic information of this disease are also shared through this portal.
The *Transmissibility of 2019-nCoV.* at Abdul Latif Jameel Institute for Disease and Emergency Analytics (J-IDEA) [7, 15] described the Reproduction Number and rate of this disease. The same is also supported by WHO whereas the data about the Reproduction Number of Ebola (used for comparison) was used as mentioned in the report by Special Team of WHO [14]. R. Varalakshmi presented a probabilistic mathematical model based on Bayes theorem to predict the number of people likely to be affected in the upcoming week [17].

D. Monitoring the Pandemic

Since the outbreak of COVID-19, over 1,90,000 people have been infected throughout the world and over 7,500 people have lost their lives till March 18th [13]. The data used is time series data. Time series data is a type of data where a variable have a set of observations on values collected over different points of time. They are usually collected at fixed intervals, such as daily, weekly, monthly, quarterly or annually. Fig. 1 presents the rapid growth of COVID-19 across the world since 21st January 2020 to 18th March 2020.

Fig. 1. Map showing the spread of COVID-19 from 21st January 2020 to 18th March 2020.

Among the several pandemics that have been studied till date it has been found that every pandemic's spread attains a pyramid like structure. In this pyramid, the mild cases are arranged at the bottom whereas the serious cases or deaths occupy the peak of this pyramid.

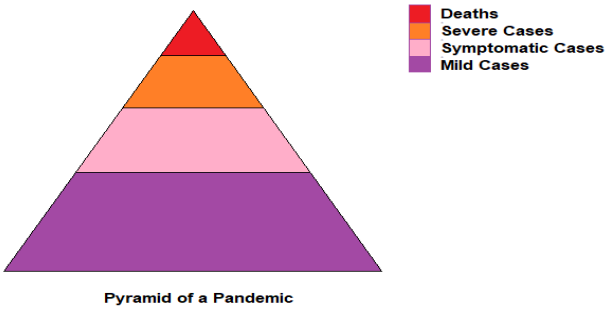

Fig. 2. Pyramid shape attained by pandemic.

It's indeed challenging to identify the base of the pyramid i.e. the actual number of people who are infected but don't exhibit the symptoms and are not identified as victims [16]. In this situation, if the total number of active cases are identified correctly then it can play a vital role in breaking the chain of the disease. Under the exploratory data analysis, following graphs & world cloud shows the growth of confirmed cases and death cases in the world.

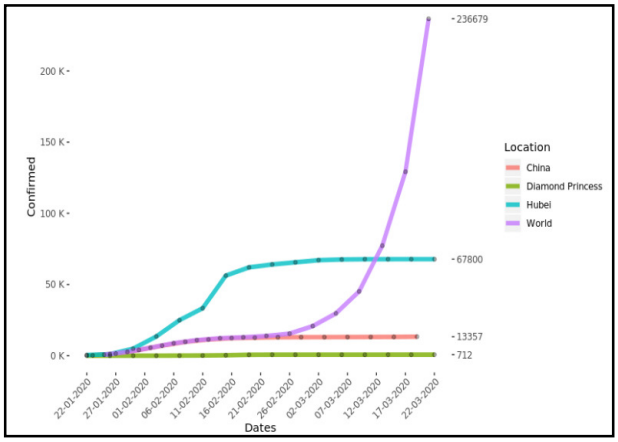

Fig. 3. Confirmed cases since 22nd January 2020.

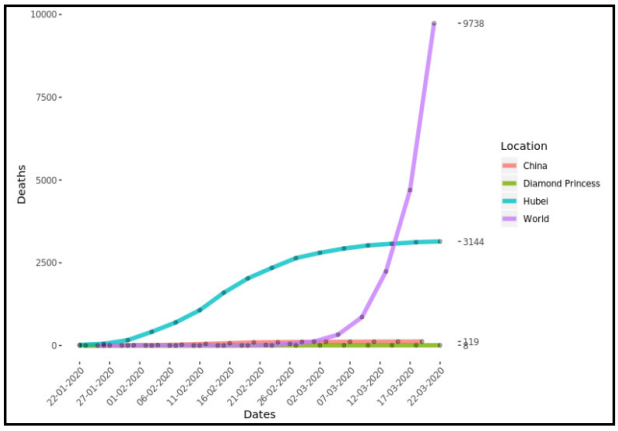

Fig. 4 (a) Death cases since 22nd January 2020.

Fig. 4 (b) World cloud presenting the death rate across world till 18th March 2020.

V. DATA AND METHODOLOGY

This section throws light on the datasets, data preparation and the algorithms used to establish a pattern of interrelationships among the number active cases, number of deaths and the number of confirmed cases to predict the rate of spread of the pandemic.

A. Dataset used

The time-series datasets used in this analysis is collected from the GitHub portal of Johns Hopkins University [13] as well as the Situation Report –58 of World Health Organization [9]. There are three dedicated databases for data about the total number of confirmed cases, death cases and recovery cases, all around the world. All the three datasets have 468 observations of 62 variables each. The brief description of the attributes is as follows:

Province/State:
Data-type: factor they can be specific, unique and valid names
Data-description: Holds name of City/Province/State, where the data is coming from
Example: *Hubei*
Country/Region:
Data-type: factor they can be specific, unique and valid names
Data-description: Holds name of the country, in which the reported area comes
Example: *China* (Hubei is a Province of China)
Lat:
Data-type: numeric (i.e. can have values in decimals, too)
Data-description: Holds the Latitude position of the given place
Example: *Latitude* position of Hubei = 30.9756
Long:
Data-type: numeric (i.e. can have values in decimals, too)
Data-description: Holds the longitude position of the given place
Example: *Longitude* position of Hubei = 112.2707
Col. 5 to 62:
Data-type: integer (i.e. discrete) and remains *always positive* as it determines the *no, of individuals*
Data-description: It's a time series data where the data is collected at various interval of time.

Each datum value is represented, based on the different days in series (from 22/01/2020). The constant entity is the location, whose data is represented in every row

B. Data Preparation

The data-preparation enriches the data and is considered to be the most time-consuming phase of any analysis task. It makes the data ready for the analysis by cleaning, transforming or reducing the set of attributes.

Raw-data has a lot of the vulnerabilities. Most often, these are *NA*s and *NaN*s, missing data values, inconsistent data or incorrect data values. The dataset used has noise like blanks in the place of states' name and data of a Cruise Ship among countries' data. To deal with these issues, data-cleaning is performed. Since available data is the time-series dataset populated with discrete data values, storing the aggregate count of the total people having COVID-19 confirmed cases, death cases or recovered cases, the missing values cannot be replaced by using the mean of the column because in this dataset, the data value can remain CONSTANT or can INCREASE, on the very next day. The missing values or NAs are replaced with the maximum value up to a day before the current day. It means that the values are carried constant for the next day whose data is missing. Fig. 5 presents the growth rate graph of COVID-19 in the world.

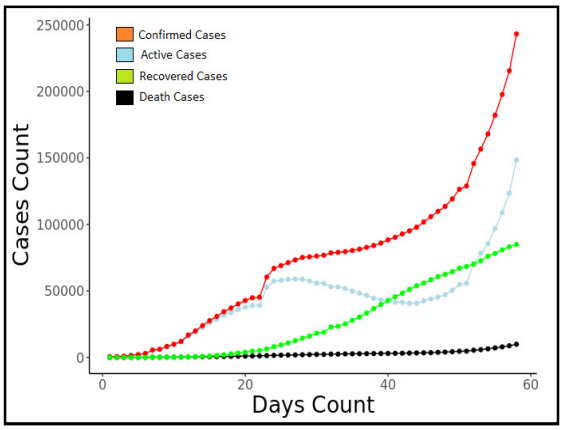

Fig. 5. Exponential growth of the COVID-19 worldwide.

Table I gives the statistics of the distribution of the available data of the confirmed cases, death cases and recovered cases.

Table 1: Mean(μ) of each category of cases.

Confirmed cases	519.7308
Death cases	21.14316
Recovered cases	181.3953
Active cases (Derived)	374.9399142

From Table 1, it's clear that the mean (μ) values of the Confirmed Cases, Death Cases & Recovered Cases are all greater than 1. It means that their reciprocal (λ) is always remains positive.

Fig. 6 presents the growth rate graph of COVID-19 in China and the box-plot in Fig. 7 clearly shows that China is the most affected country till this date.

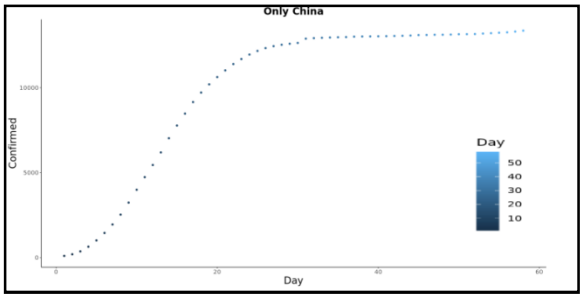

Fig. 6. Example Growth of COVID-19 in China.

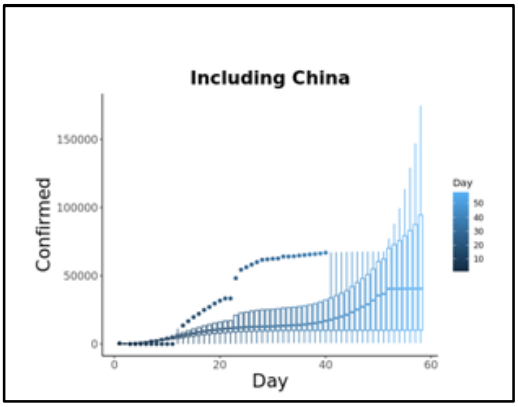

Fig. 7. Growth of COVID-19 worldwide, including China as outlier.

As per the trend and the conclusion of the Table 1 and the graphs in Figs. 5, 6 and 7, it is clear that Confirmed Cases, Death Cases & Recovered Cases data is distributed exponentially.

The redundant attributes such as longitude and latitude positions of the affected states or provinces in the available datasets are dropped. The data from Confirmed Cases, Death Cases & Recovered Cases merged into a single dataset. The new dataset is used for data transformation. In this phase initially the two attributes i.e. *Date* & *Day* were added having the data type as factor, for both. Along with these two, four other columns were added to this dataset using Confirmed Cases, Death Cases & Recovered Cases as their generator. These new attributed were Active Cases, Closed Cases, Active Cases (%), Closed Cases(%).

The percent of cases was calculated by using Total Confirmed cases as the aggregate. Active and Close Cases were calculated as follows:

Closed Cases = Death Cases + Recovered Cases

Active Cases = Confirmed Cases - Closed Cases

This is a quantitative analysis and hence regression approach is implemented. The value of total number of Active Cases is an absolute figure as the number of these cases cannot be a floating-point number. An absolute number is not a good measure to determine the current status of pandemic in a region. It is because this is the cumulative sum and thus this number depends upon the past records and cannot be used to predict whether the pandemic is spreading or is under control. Hence, a new measure 'Active Case (%)' and hence 'Closed Case (%)' is introduced which is defined as:

Closed Case (%) = (Total Closed Cases * 100)/Total Confirmed Cases
Active Case (%) = 100 - Closed Case (%)
This measure is able to provide the current status of the pandemic in the provided region. Here the aim is to estimate the Active Cases(%).

C. Regression algorithms used

Regression algorithms are used to establish a pattern of interrelationships among the number active cases, number of deaths and the number of confirmed cases in a region to predict the rate of change of the pandemic in the upcoming weeks. The four algorithms used are Support Vector Machine - kernel Regression, k-Nearest Neighbor Regression [5], Linear Regression [1] and Polynomial Regression [3].

The evaluation metrics used are RMSE and R^2.RMSE (Root Mean Squared Error) [2] tells that how far the actual data points are from the regression curve whereas the R^2 (R-Squared) [4] tells about the closeness of the predicted and actual data points. RMSE (Root Mean Squared Error) tells that how far the actual data points are from the regression curve whereas the R^2 (R-Squared) tells about the closeness of the predicted and actual data points. R-Squared is also known for identifying the goodness of fit for any regression model. To attain a good accuracy, the RMSE must be decreased while the R^2 values must be increased.

Support Vector Machine - kernel Regression: kSVM supports class-probabilities output and confidence intervals for regression. As per this algorithm, the model seems to attain following summary:

Table 2: Evaluation of the SVM-k Model.

RMSE	56.3856139
R2	0.9164658

Here the radial basis function (RBF) kernel is used inthis model training. It is so because the localized & finite response is required. Also, the RBF is the most used kernel because it's known for giving localized and finite response along the **entire x-axis.** Fig. 8 shows the graphical representation of a test dataset fed to the model trained on kSVM.

As the RMSE value for this algorithm is quite high, kNN algorithm is tried next for training.

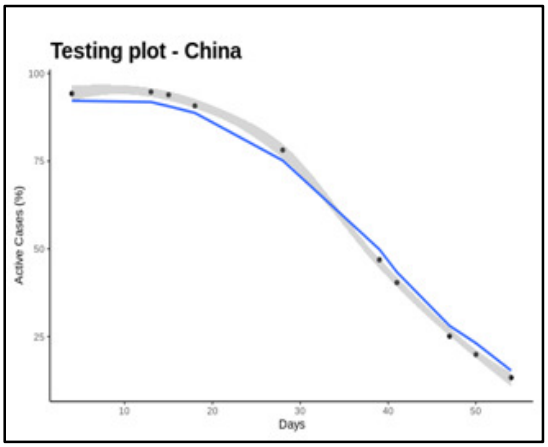

Fig. 8. Testing model for SVMK algorithm.

k-Nearest Neighbor Regression: After plugging different values of k, for k = 3, the kNN model gives an acceptable estimate. The summary of the model trained on kNN algorithm is given in Table 3. Fig. 9 shows the graphical representation of a test dataset fed to the model trained on kNN.

Table 3: Evaluation of the kNN regression Model.

RMSE	58.3887824
R^2	0.9141718

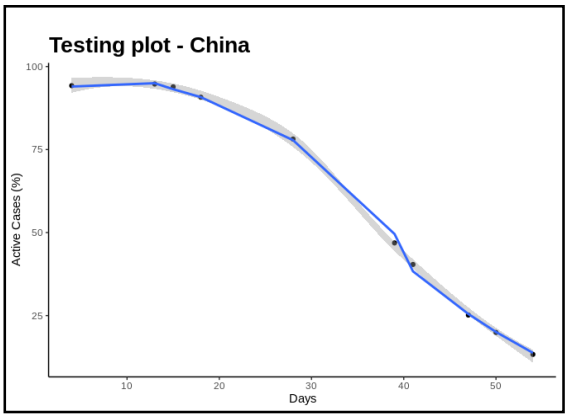

Fig. 9. Testing model for kNN algorithm.

Here both of the models i.e. trained on the SVMK and the kNN algorithms, RMSE value was very high. To train the model with best possible accuracy that should have a relatively low RMSE score, the Linear Regression is used.

Linear Regression: After the linear regression, the summary of the model appears as follows:

Table 4: Evaluation of the linear regression Model.

RMSE	4.9293341
R^2	0.9131809

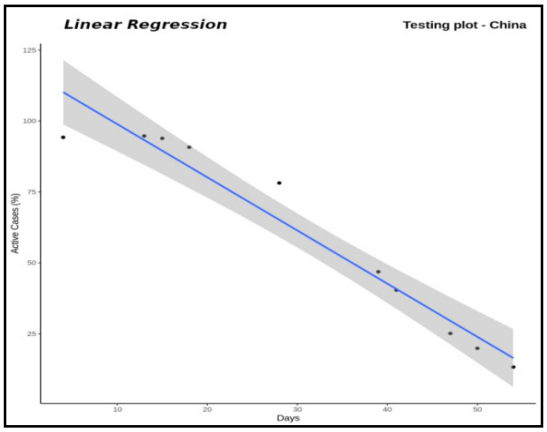

Fig. 10. Testing model for Linear Regression.

Fig. 10 shows the graphical representation of a test dataset fed to the model trained on linear regression. As in case of the Linear Regression, the regression line is straight whereas for the actual values, there are many fluctuations in the actual data. That is why the Polynomial Regression is selected for further comparison, that can fit a very large range of degrees.

Polynomial Regression: After all these different regression models, for the degree level equals to 11, the polynomial regression best fits into the scenario. For Degree 11, the summary of polynomial regression is noted in Table 5. Fig. 11 shows the graphical representation of a test dataset fed to the model trained on polynomial regression.

Table 5: Evaluation of the polynomial regression Model.

RMSE	0.5819927
R^2	0.9996465

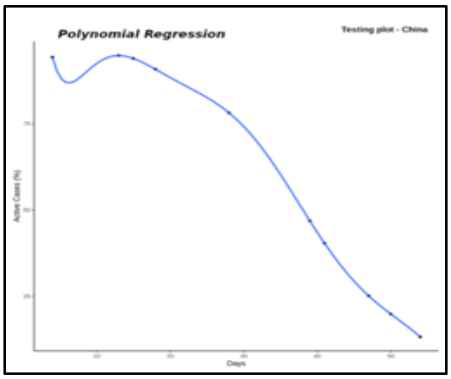

Fig. 11. Testing model for Polynomial Regression.

Finally, this graph as well as the its evaluation matrix appear to be the best among all the four different algorithms that were compared.

VI. RESULTSAND DISCUSSIONS

Out of the four regression algorithms used to train the model, polynomial regression with degree level 11 came out to be the best. An estimate of the total cases that shall be *active* by the end of the following day using polynomial regression model is presented in Table 6.

Table 6: Font Sizes for Papers.

Date	Active Cases (%)
22/03/2020	8.400352
23/03/2020	8.311558
24/03/2020	8.699806
25/03/2020	9.441795
26/03/2020	10.174679
27/03/2020	10.159197
28/03/2020	8.096655
29/03/2020	1.889395

Representing this result graphically in Fig. 12.

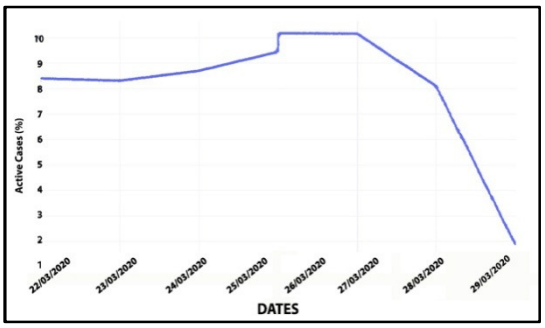

Fig. 12. Graphical representation of result (estimation).

The graphical representation of the results of this study, as in Table 6 shows a rapid drop in the active cases out of every 100 confirmed cases (i.e. active cases percent) or in other words, a rapid rise in resolved (Closed) cases' percentage.

After this whole statistical analysis, from Table 6 and Fig. 12, it becomes clear that China seems to be very successful in controlling the growth of the disease in its region. As per the results, it seems that the share of resolved cases is going to be far greater than those which are not resolved (i.e. Active cases) or confirmed recently, among all the Confirmed Cases, till date. Thus, it indicates that by the end of the March, this pandemic can be controlled at up to a satisfactory level.

VII. CONCLUSION AND FUTURE SCOPE

The year 2019 ended with the inception of the novel corona virus. The highly contagious nature of the virus wrecked the health care systems of the nations as the hospitals ran out of beds and ventilators that led to a national health emergency. Measures were taken to quarantine the infected ones from the rest of the population and the need to monitor the COVID-19 outbreak surfaced. In such a situation, this study proves to be very helpful in prediction of the growth of COVID-19. The models were trained on four regression algorithms namely kSVM, kNN, linear regression and polynomial regression. Among these four algorithms, the polynomial regression came up with the most promising result with the least RMSE value 0.5819927 and the best R^2 value 0.9996465. In future work, predictions can be made for any other country as well. In this particular study, the cases are divided in two categories i.e. Active & Closed. This work can be extended to the next level by further dividing the closed cases by finding the Death rate and recovery rates (in terms of percentage), separately. This can be helpful in understanding whether the closed cases are a success or a failure, depending upon the condition of most of the closed cases resulted into death or recovery. Estimations for the possible deaths can be made using Death rate. In case of having the number of days for different countries, in which the cases get doubled, the possible number of deaths that are not reported, might also be estimated. Having the real time data, this study can be used to take even better decisions in case of the prevention of the disease.

Conflict of Interest. No.

REFERENCES

[1]. Yan, X., & Su, X. (2009). *Linear regression analysis: theory and computing*. World Scientific, 1-2.

[2]. Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International journal of forecasting*, *22*(4), 679-688.

[3]. Gergonne, J. D. (1974). The application of the method of least squares to the interpolation of sequences. *Historia Mathematica*, *1*(4), 439-447.

[4]. Anderson-Sprecher, R. (1994). Model comparisons and R 2. *The American Statistician*, *48*(2), 113-117.

[5]. Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, *46*(3), 175-185.

[6]. WHO, World Health Organization (2020). [Online]. Available: https://who.int/. [Accessed February 2020].
[7]. Imai, N., Cori, A., Dorigatti, I., Baguelin, M., Donnelly, C. A., Riley, S., & Ferguson, F. M. (2020). Transmissibility of 2019-nCoV. *2020-01-25]. https://www. imperial. ac. uk/mrc-global-infectious-disease-analysis/news-wuhan-coronavirus*.
[8]. Coursera, "Science Matters: Let's Talk About COVID-19, Week 1-3," [Online]. Available: https://www.coursera.org/. [Accessed February 2020].
[9]. Ilaria Dorigatti+, Lucy Okell+, Anne Cori, Natsuko Imai & Team, "Report 4: Severity of 2019-novel coronavirus (nCoV)," (J-IDEA), p. 2, February 2020.
[10]. WHO, "Novel Coronavirus(2019-nCoV), Situation Report –58, 18 March," WHO, 2020.
[11]. WHO, "Novel Coronavirus(2019-nCoV), Situation Report -22, 11 February," WHO, p. 1, 2020.
[12]. WHO, "Novel Coronavirus(2019-nCoV) , Situation Report –22, 11 February," WHO, p. 2, 2020.
[13]. WHO-China, "Joint Missionon Coronavirus Diesease 2019 (COVID-19), 16-24 February," WHO, 2020.
[14]. Johns Hopkins University (2020) CSSE at Johns Hopkins University (COVID-19)," 2020. [Online]. Available: https://github.com/CSSEGISandData/COVID-19/. [Accessed January 2020].
[15]. WHO Ebola Response Team. (2016). After Ebola in West Africa—unpredictable risks, preventable epidemics. *New England Journal of Medicine*, *375*(6), 587-596.
[16]. Abdul Latif Jameel Institute for Disease and Emergency Analytics, "(J-IDEA)," [Online]. Available: https://www.imperial.ac.uk/jameel-institute/. [Accessed February 2020].
[17]. V. R., (2020). Mapping of Corona Virus Transmission in India with a Mathamatical Approach. *International Journal on Emerging Technologies, 11*(2), 245-250, 2020.

How to cite this article Vashisht, G. and Prakash, R. (2020). Predicting the Rate of Growth of the Novel Corona Virus 2020. *International Journal on Emerging Technologies*, *11*(2): 000–000.