

## Journal Pre-proofs

Engineering Physics and Mathematics

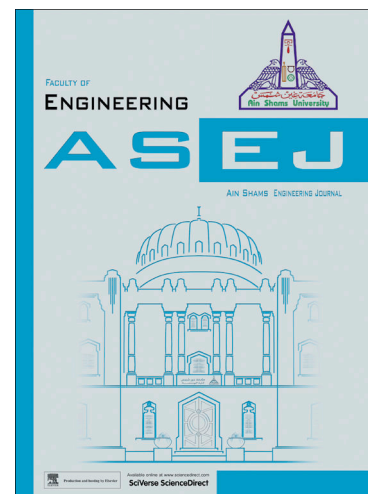
Performance Evaluation of Regression Models for COVID-19: A Statistical and Predictive Perspective

Mohammad Ayoub Khan, Rijwan Khan, Fahad Algarni, Indrajeet Kumar, Akshika Choudhary, Aditi Srivastava

PII: S2090-4479(21)00338-5  
DOI: <https://doi.org/10.1016/j.asej.2021.08.016>  
Reference: ASEJ 1574

To appear in: *Ain Shams Engineering Journal*

Received Date: 25 September 2020  
Revised Date: 3 August 2021  
Accepted Date: 31 August 2021



Please cite this article as: Ayoub Khan, M., Khan, R., Algarni, F., Kumar, I., Choudhary, A., Srivastava, A., Performance Evaluation of Regression Models for COVID-19: A Statistical and Predictive Perspective, *Ain Shams Engineering Journal* (2021), doi: <https://doi.org/10.1016/j.asej.2021.08.016>

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2021 THE AUTHORS. Published by Elsevier BV on behalf of Faculty of Engineering, Ain Shams University.

## Performance Evaluation of Regression Models for COVID-19: A Statistical and Predictive Perspective

Mohammad Ayoub Khan<sup>1</sup>, Rijwan Khan<sup>2</sup>, Fahad Algarni<sup>1</sup>, Indrajeet Kumar<sup>3</sup>, Akshika Choudhary<sup>2</sup>, Aditi Srivastava<sup>2</sup>

<sup>1</sup>College of Computing and Information Technology, University of Bisha, Bisha, 67714, Saudi Arabia

<sup>2</sup>ABES Institute of Technology, Ghaziabad (U.P.), India

<sup>3</sup>Assistant Professor, Department of Computer Science and Engineering, Graphic Era Hill University, Dehradun,

### Corresponding authors:

College of Computing and Information Technology, University of Bisha, Bisha, 67714, Saudi Arabia

Email: [ayoub.khan@ieee.org](mailto:ayoub.khan@ieee.org)

**Conflicts of Interest:** “The authors declare that they have no conflicts of interest to report regarding the present study.”

**Funding Statement:** “There is no funding available for this work”

## Performance Evaluation of Regression Models for COVID-19: A Statistical and Predictive Approach

### ABSTRACT

Research is very important in the pandemic situation of COVID-19 to deliver a speedy solution to this problem. COVID-19 has presented governments, corporations and ordinary citizens around the world with technology playing an essential role to tackle the crisis. Moderate and flexible innovation arrangements that can speed up progress towards giving critical well-being ability are proved hourly. Knowledge with the aid of creativity must be obtained, accepted and analysed in a short time frame. In this example, the machine learning model has a major role to play in predicting

the number of next positive COVID-19 cases to come. For government departments to take effective and strengthened future COVID-19 planning and innovation. The ongoing global pandemic of COVID-19 has been non-linear and dynamic. Due to the especially perplexing nature of the COVID-19 episode and its diversity from country to country, this study recommends machine learning as a convincing means to demonstrate flare-up. In this linear regression, polynomial regression, ridge regression, polynomial ridge regression, support vector regression models, the COVID-19 data set from multiple on-line tools have been evaluated. During the work process comprehensive experiments were performed and each test was evaluated with the parameters mean square error (MSE), medium absolute error (MAE), root mean square error (RMSE) and R2 score. This study also offers a path for future research using regression models based on machine learning. Precise validation and data analysis can contribute to strategies for healing and disease prevention at an early stage. A systematic comprehensive strategy is a new philosophy in which statistical data for government agencies and community can be forecast.

**Keywords:** COVID-19, Machine Learning, Regression, Prediction, Non-linear.

## 1. INTRODUCTION

Coronavirus or COVID-19 originated in Wuhan, China in the month of December 2019. Till date (12 July, 2021), there is no avowed human immunization for combating it. COVID-19 engendering is speedy when people are in vicinity [1]. Coronavirus would be the most unnerving situation faced by the whole world since the end of the second world war. Even more unnerving as the enemy is invisible. COVID-19 has affected our lives to a great extent. Work, economy, education and almost everything has come to stand still. People nowadays only focus on what is essential to live. Government and other organizations have had to take unprecedented steps to stem the trail of destruction. The reasons for its spread and thinking about its threat, practically all the nations have proclaimed either partial or complete lockdowns all through the influenced districts and regions. Since there is no endorsed medicine till now for slaughtering the infection so the legislatures of all

nations are concentrating on the precautionary measures which can stop the spread [2]. Since there is no embraced solution till now for dispatching this contamination, the authorities of all countries are paying attention to the careful steps that can minimize the spread. This virus causes respiratory tract infections that can range from mild to fatal. This infection can influence individuals of any age yet older people or individuals with previous ailments are more inclined or vulnerable to it. The most testing part of its spread is that an individual can have the infection for a long time without indicating side effects [3,4, 5, 6]. Utilization of mask, sanitizer, customary hand washing and cleanliness is the most ideal approach to keep from this illness. People are urged to believe Staying homes saves lives. COVID-19 will reshape our reality.

The World Health Organization (WHO) on February 11, 2020 came forward stating this virus as a pandemic outbreak naming it as COVID-19 [21] further stating that the virus had first taken place in China, making it to different countries. USA, Brazil, and India have been the most affected countries where the number of cases of this pandemic COVID-19 multiplied rapidly on the daily basis.

Machine learning plays a major role in better understanding and examining COVID-19 crisis as it identifies the patterns in data and uses them to automatically make predictions or decisions. In the medicinal services area, Machine Learning can be seen as an asset that has the extension to process colossal datasets past the capacity of human personalities and the derived bits of knowledge help doctors in arranging and giving consideration to get acceptable treatment. On the basis of this concept, this work has been carried out for analysis of COVID-19 cases and prediction of upcoming new positive cases [18]. In this model authors are going to look at the development of Coronavirus affirmed, demise and recouped instances of whole world and comparison of five countries that have been vigorously tainted done with rest of world. Investigation and predictions are done using Linear Regression, Ridge Regression, Polynomial ridge Regression [5], polynomial Regression and Support Vector Regression (SVM). The proposed expectation model guarantees that it follows the authentic outcome with respect to this pandemic circumstance so that tremendous financial misfortune, network spread, measure of social separation of individuals can be identifiers and furthermore precise choice can be taken likewise [7]. This strategy will ensure the administration to yield preventive evaluations reliant on our next work for foreseeing the presence of this infection in future.

The remainder of the paper comprises five sections. Section 2 present the Related Work, Section 3 presents the Materials and Methods and Staging Analysis Parameter used for investigation, Section 4 presents the evaluation with comparison table and section 5 sums up the paper and presents the conclusion with references.

## 2. RELATED WORK

Since the time the rise of COVID-19 and its resulting spread across landmasses overwhelming both progressed and creating countries, there has been a great deal of exploration papers distributions on different parts of COVID-19. So, in various research paper analysis is done on vaccination, drug therapy and also on the prediction of future infected, recovered and death cases [1,2]. Using various forecasting techniques

L. Jia et al., 2020 [12] did the prediction and analysis of COVID-19 through various models. The differential equation, classical differential equation, historical arrangement expectation model is utilized in the project. They also used an Internet-based infectious disease prediction model. They used various models which came under mathematical models such as the Logistic model, Bertalanffy model [12] and Gompertz model [12]. Model evaluation was done using regression coefficient ( $R^2$ ). Fitting and analysis of SARS and COVID-19 was done on two basis/categories: number of affirmed cases and number of passing toll. Different prediction graphs were obtained under these categories. According to this paper they concluded that this epidemic will be over presumably in the last of April but their estimation is not so accurate [12]. F. Rustam et al [2020] introduced future forecasting on Covid-19 using various machine learning models. They use four regression models namely LR, LASSO, SVM and Exponential Smoothing. They predicted their results on the basis of Evaluation Parameters used in their models. According to their results ES model is best for predicting infected cases and recovered cases and LR model best predict the death rate [9].

S. Dutta et al., 2020 [20] introduced in their paper the prediction models for confirming COVID-19 cases. They used three models in their paper namely LSTM (Long Short-Term Memory) model, GRU (Gated Recurrent Unit) model and combined LSTM-GRU Model [20]. Deep Learning Neural Network is employed. The consolidated LSTM-GRU based RNN model gives a relatively better outcomes as far as forecast of affirmed, delivered, negative, demise cases on the information [20]. Similarly, Narinder et al., 2020[19] in this paper mirrored that because the eruption of the

COVID-19 has turned into pandemic, the analysis of medicine knowledge is required to arrange the society to combat this challenge. The four stages of COVID-19 are described in this paper. The regression is trained and cross-validated on real time knowledge victimization the quantity of confirmed, recovered, and death cases. The results were displayed through graphs that pictured COVID-19 worldwide epidemic analysis victimization of machine learning and deep learning techniques/algorithms [19].

The work carried by Sohini and Sareeta in 2020 [10] proposed a model in which they have done data visualization after that they used various forecasting machine learning techniques. During this paper a comparison is formed between the expansion of Covid-19 confirmed deaths and recovered cases of Asian countries to different major countries that have conjointly been heavily infected. The various predictions are aforethought on a bar chart. Gender distribution is mirrored through a chart that reveals males are a lot of possible to be diagnosed with Covid-19. Age wise distribution is additionally shown through pie charts. They concluded that Sigmoid model is best among all the models they used [10]. Binti Hamzah et al., [8] and his fellow mates in year 2020 made corona tracker an Online stage that provides most up-to-date what's a lot of, solid news advancement, even as insights and investigation on COVID-19. They have done continuous information inquiry and imagined on their site and at that point questioned information is utilized for SEIR modelling. On the basis of day by day observations they utilized the SEIR model to figure Covid-19 episodes inside and Outside of China. Similarly, they have done investigation on the Queried News [8].

L Wynants et al., (2020) proposed a model in which they have inspected all the papers identified with covid-19 pandemic. They recognized that proposed scheme is ineffectively detailed and at high danger of predisposition, raising worry that their expectations could be untrustworthy once applied day by day [4]. R. Sujatha et al., in 2020 used three models for their predictions and did visualization on India. Using graph plotting, they try to figure out the future trend of this epidemic. From apprehended qualities and coordination cases from dataset information they concluded that MLP strategy is giving acceptable forecast outcomes than that of the LR and VAR technique [11]. L. Peng et al., (2020) using SEIR model they predict in light of the open information of National Wellbeing Commission of China from January twentieth to February sixteenth, 2020, they dependably gauge key plague boundaries and make forecasts in the curvature period and conceivable consummation time of 24 regions in Territory and 16 regions in Hubei region [17]. Li

Cuilian et al., in 2020 proposed a model to assess the forecast estimation of the Internet search information from online web indexes and web-based social networking for the COVID-19 episode in China [16]. Vinay and Lei Zhang (2020) reflected in their paper the approximate finishing purpose of this epidemic in Canada and in the world. They have developed a forecasting model using deep learning techniques. The dataset required for this purpose was collected from a University and Canadian fitness management. Long short-term memory model was displayed to forecast COVID-19 and based on the graphs depicted by LSTM they predicted that this pandemic will end by June 2020. Despite the fact that their model accomplished better execution when contrasted and other anticipating models [15].

Dong Je et al., (2020) introduced a model that clarifies the high-risk factors of coronavirus. They collected the clinical data of all the patients that were admitted to Fuyang Second People's Hospital in China between 20th January and 22nd February 2020. They divided around 208 patients under two categories- Stable group and Progressive group. Univariate and multivariate analysis depicted the autonomous high-hazard factors for COVID-19 movement. They concluded that the use of the CALL score model can be an effective resource to combat this challenge [14]. Weston C. Roda et al., (2020) summarized that after comparing SEIR and SIR models that it is not necessary that a complex model such as SEIR is more effective than a simple model SIR. They used Bayesian framework, Markov chain Monte Carlo algorithms and Akaike Information Criterion (AIC). Parameters in the SIR model and Parameters in the SEIR model were shown through tables. Comparisons were made through graphs [13]. C. Sohrabi et al., 2020 [21] tells that on 31st December 2019 few cases of coronavirus were identified in China. WHO global health emergency, global reaction, reported UK cases and British response, transmission, viral spread, preventing, diagnosing, treating, forecasting and containment methods for this virus have been addressed by the authors. The authors also compared the diagnostic criteria for the CDC against the WHO based on symptoms and travel. In the paper the authors have also described the retrospective from response to COVID-19.

R. Khan et al., 2020 [43] investigate sentiment analysis techniques for the analysis of twitter covid-19 data. They find that it has been seen by pre-handling the information utilizing the regex and coach has been a viable answer for getting out the unpredictability of the applied calculation also the information instead of straightforwardly applying on the crude information itself. By utilizing



the prepared model and further utilizing it with the classifier end up being a superior route for grouping as it decreased the time period and the size of the information outline diminishing the time unpredictability associated with the cycle. Discussing the examination, it has been seen that the number for tweets shared by the dynamic clients has been consistently more prominent when contrasted with different assessments. It implies with respect to the pandemic greatest number of individuals thought and took the choices made by the administration or the nearby experts in a positive manner. While the quantity of tainted and demised individuals continued expanding it didn't influence the psychological quality of the populace. For the 3-month investigation with respect to the Indian sub-landmass the variety among the positive, negative and nonpartisan estimations stayed consistent with the quantity of expanding cases step by step.

Gozes, Ophir, et al, 2020 [44], underlying examination, which is right now being extended to a bigger populace, exhibited that quickly created AI-based picture investigation can accomplish high precision in identification of Coronavirus just as measurement and following of sickness trouble.

Shi, Feng, et al, 2020 [45] discussed that for COVID-19 pandemic clinical imaging techniques has a significant function in battling against COVID-19. In their research authors examines how AI gives sheltered, precise and proficient imaging arrangements in Coronavirus applications. The X-ray and CT scan techniques have been utilized to exhibits the adequacy of Man-made intelligence engaged clinical imaging. It is significant that imaging just gives incomplete data about patients. This is imperative to join imaging information with clinical indications and lab assessment results to support the screening, location furthermore, finding of COVID-19. For this situation, we trust AI will exhibit its capacity in combining data from this multi-source information, for performing precise and productive conclusion, investigation and development.

Yan, Carol H., et al, 2020 [46], In ambulatory people with flu like side effects, chemosensory brokenness was unequivocally connected with COVID-19 contamination and ought to be viewed as when screening manifestations. Most will recoup chemosensory work inside weeks, resembling goal of other illness related manifestations.

Sun, Yinxiaohe, et al, 2020, [47] discusses that clinical and laboratory data can be rapidly established for individuals who are exposed to COVID-19 and allow PCR testing and containment efforts to be made a priority. For prediction models, the primary laboratory test results were important.



Oliveiros, Barbara, et al, 2020 [48] discussed that COVID-19 progression rate is projected to be slower with spring and summer. However, the two variables make up at most 18% of the transition, with the other 82% connected to other factors, such as containment steps, general health policies, population density, transport, cultural matters, etc. Furthermore, the direct effect is small: if the temperature rises by 20 ° C for example, the average time doubling is predicted to increase by 1.8 days in the best-case scenario.

Chen, Xiaofeng, et al, 2020 [49] discussed that pneumonia patients with and without COVID-19 can be differentiated on the basis of CT imaging and clinical symptoms alone. A model consisting of semantic and clinical radiological features performs very well in diagnosing COVID-19.

Dryhurst, Sarah, et al. [50] said that the risk of COVID-19 around the world clearly show the consistent connection between COVID-19 perceptions of risk and various experience- and socio-cultural factors across countries. At the same time, we notice that cultural differences in risk perception must be addressed.

## **Motivation**

As we know, data science helps us to clarify and understand the data that has been accumulated so far. Data science helps one to simulate and imagine the patterns of how coronavirus spreads including the number of patients reported daily with coronavirus or possible infected cases, i.e. it is growing or more than before and so on. Data analysis helps one to gain some valuable insight into the data. Machine Learning lets us make predictions and lets us create models that map the real world and that will take some data and predict the future. So, in our model, we first try to imagine data and then forecast future data and also try to find the best fit regression model that could help us with future predictions.

## **3. MATERIALS AND METHODS**

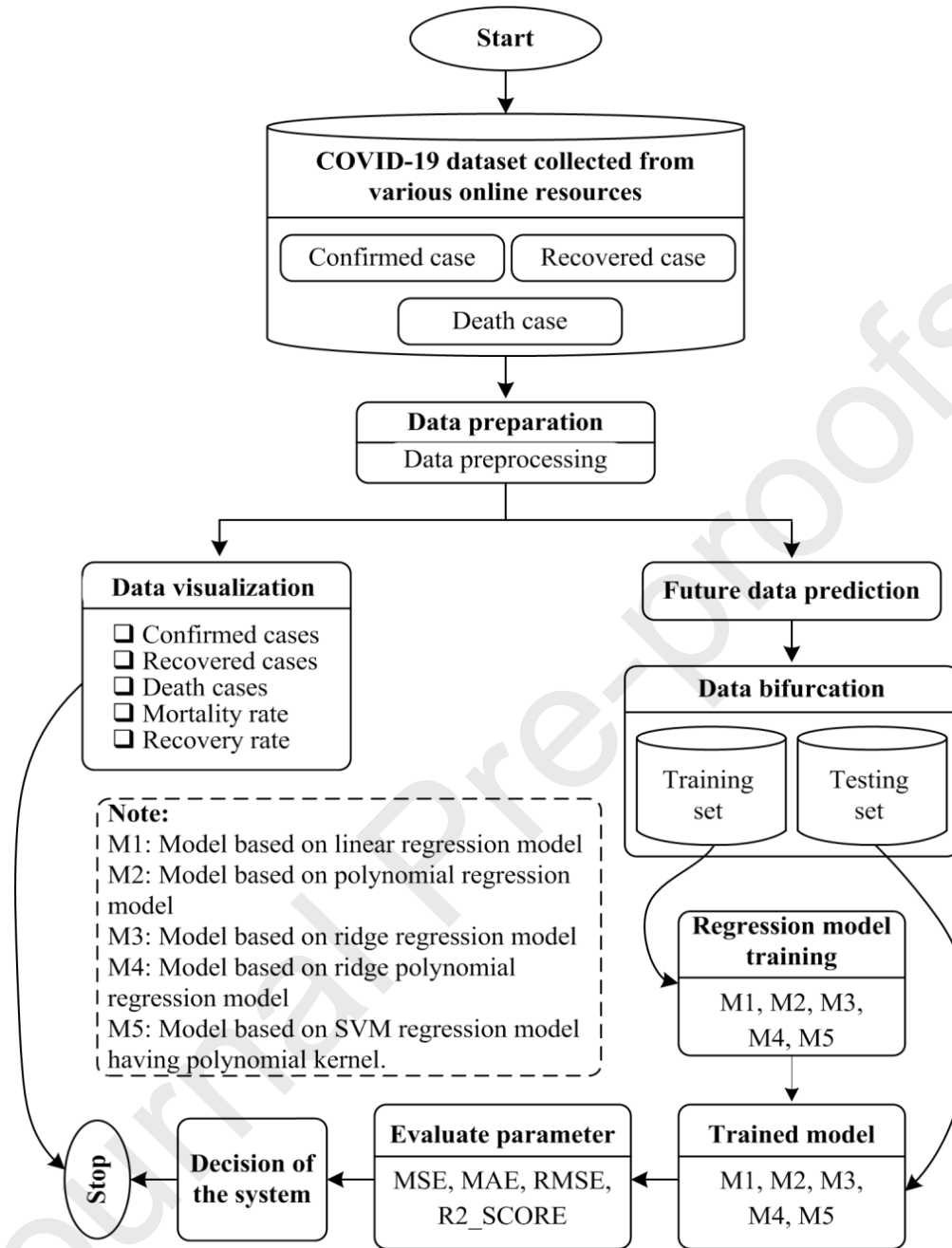
### **3.1 Dataset Preparation**

In this examination paper we have utilized the overall details of coronavirus from January 22, 2020, to July 12, 2021, was assembled from the online resources like Kaggle, Weka 3.8.4 and Orange [11]. The datasets give us the quantity of affirmed cases, recouped cases, and passing cases everywhere throughout the world. The datasets are accessible in time arrangement position with

death, month and year so fleeting parts are not neglected. We isolated our datasets into a preparing set (85%) on which our model readied and a testing set (15%) to test the exhibition of our framework.

### **3.2 Experimental work flow chart**

The experimental work flow diagram of the work is shown in Fig. 1. From the Fig. 1, it has been found that the complete work divided into two sections. Section one is used for data visualization and section two is used for future data prediction. For data visualization, progress bar, recovery rate and mortality rate are shown. In section 2, linear, polynomial, ridge, ridge polynomial and support vector machine (SVM) having polynomial kernel function is used for future case prediction. Each model is evaluated using MSE, MAE, RMSE and R2\_SCORE parameters.



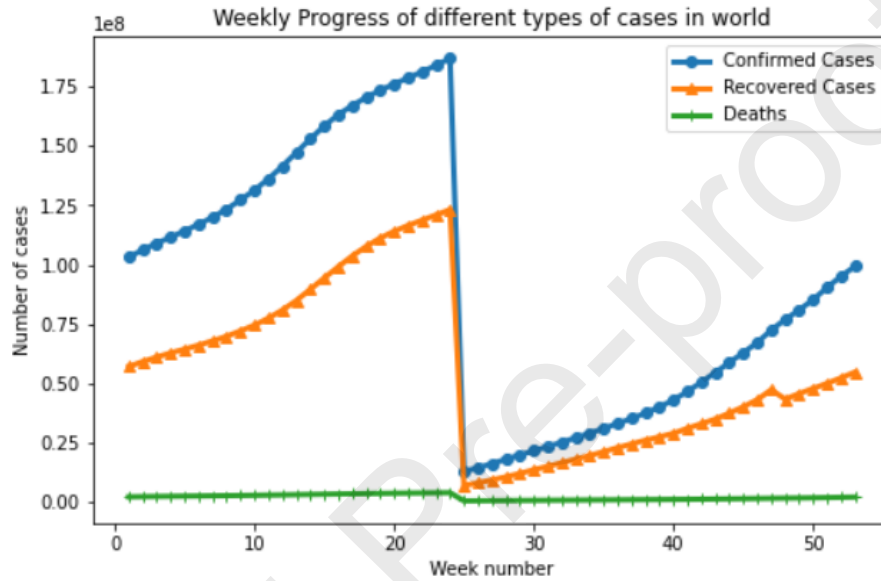
**Fig.1:** Experimental Work Flow of the Model

### 3.2.1 Data Visualization

#### 3.2.1.1 Worldwide Visualization Framework

In this work demonstrate the week after week progress of various sort of cases in world which incorporates affirmed cases, recouped cases and demise cases. From this it has been inferred that the pace of increment in affirmed cases is a lot higher than the recovered and demise rate cases.

Passing rate is low in contrast with other two cases though recouped rate is moderate. At that point we have indicated the everyday increment in affirmed cases, recouped cases and demise cases. The curve given in Fig. 2 shows the week number vs number of cases line graph shows the weekly increase of different types of cases in worldwide. From our prediction confirmed cases are increasing with a high rate and recovery rate is average as compared to number of active cases. Also, death rate is less as compared to number of active cases.



**Fig. 2:** Weekly progress of different types of cases worldwide

#### *Visualization Framework observations*

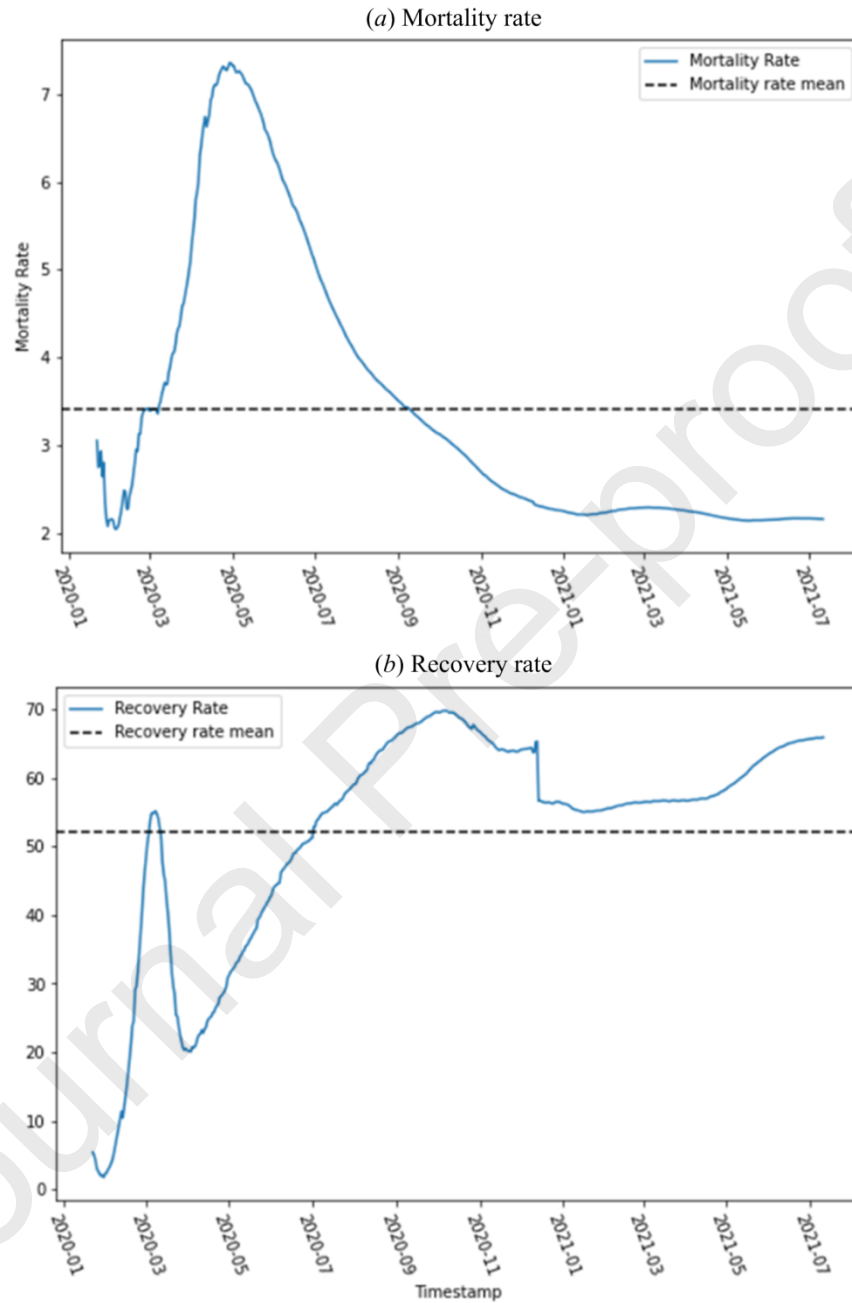
After the completion of visualization experiments, the following observations have been drawn for India cases:

- (a) It took 529 days in India to succeed in 30837222.0 Confirmed Cases.
- (b) It took 528 days in Italy to succeed in range of Confirmed cases to India.
- (c) It took 440 days in US to succeed in range of Confirmed cases to India.
- (d) It took 527 days in Kingdom of Spain to succeed in range of Confirmed cases to India.
- (e) It took 537 days in China to succeed in range of Confirmed cases to India

#### *Mortality and Recovery rate Visualization*

Mortality rate is a proportion of the quantity of deaths in a specific populace, scaled to the size of that populace, per unit of time. Recovery rate is the degree to which head and accumulated

enthusiasm on defaulted obligation can be recouped, communicated as a level of assumed worth. The monthly progress of different types of cases in world is shown in Fig. 3.

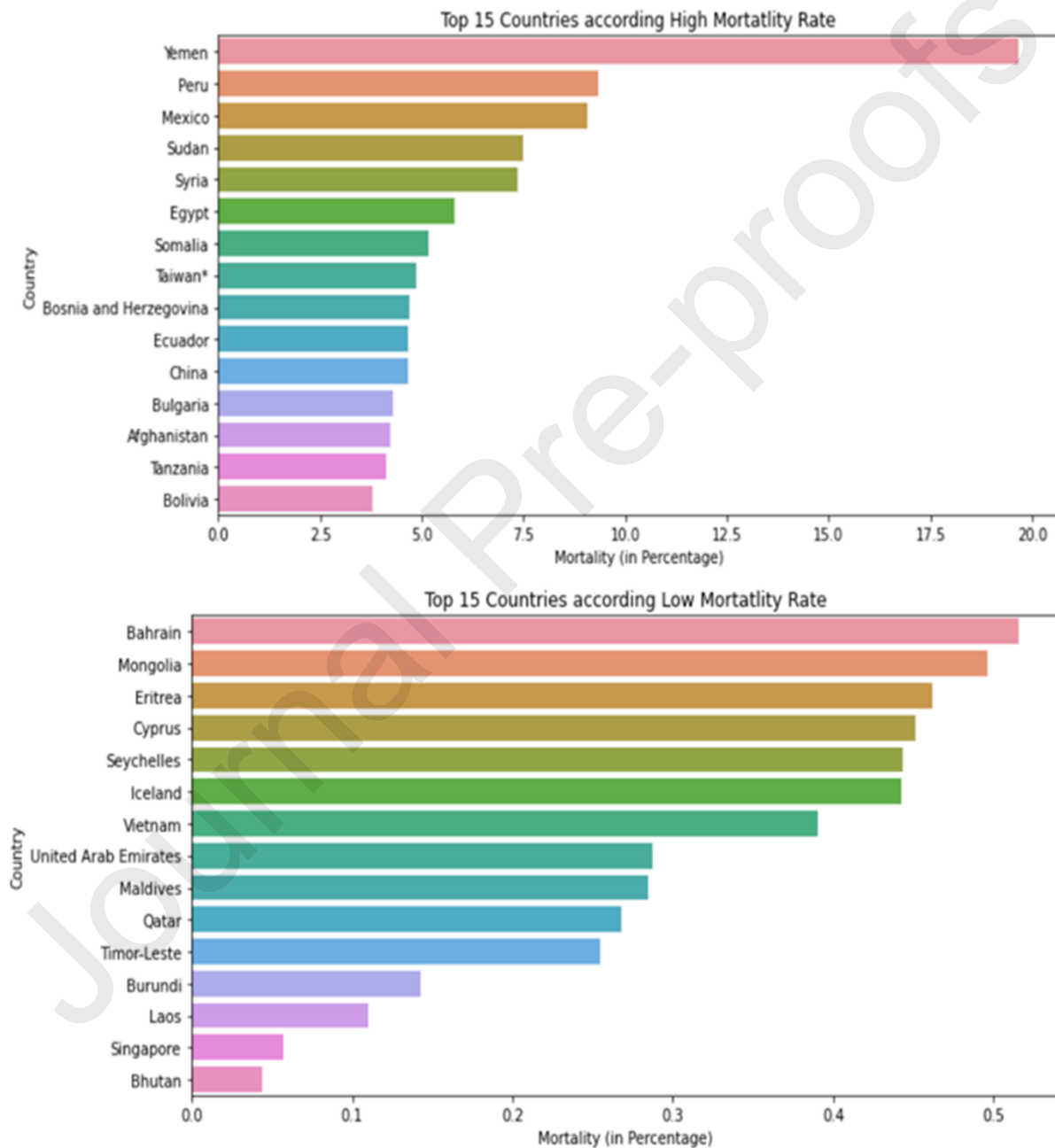


**Fig. 3:** Average monthly mortality and recovery rate (a) mortality rate (b) recovery rate

From the Fig. 3, it has been found that the mean and median of average mortality is reported as 3.413 and 2.622, respectively. Similarly, the mean and median of average recovery rate is evaluated as 52.20 and 56.61, respectively.

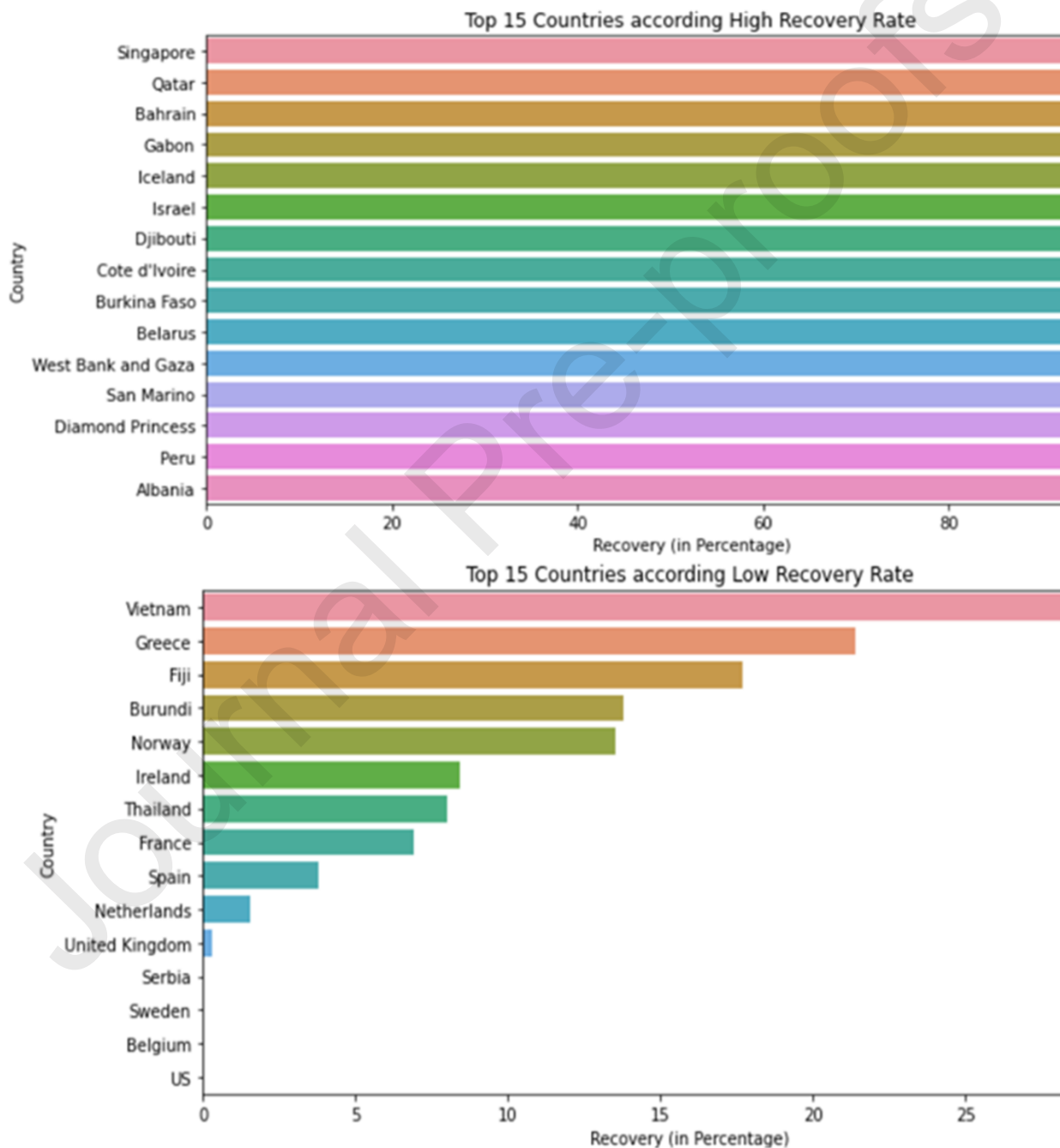
### Comparative Analysis

In this section a comparative analysis has been done according to mortality and recovery rate of top 15 countries. Fig. 4 shows the list of top 15 countries having high mortality rate and low mortality rate. According to this, Yemen has the highest mortality rate and Bhutan get the lowest position in the list of mortality rate country.



**Fig. 4:** (Top) Top 15 countries according to high mortality rate, (Bottom): 15 countries according to low mortality rate.

Fig. 5 shows the top 15 countries according to high recovery rate and low recovery rate. In this figure, it is noticed that Djibouti has highest recovery rate and Sweden has lowest recovery rate among the country list.





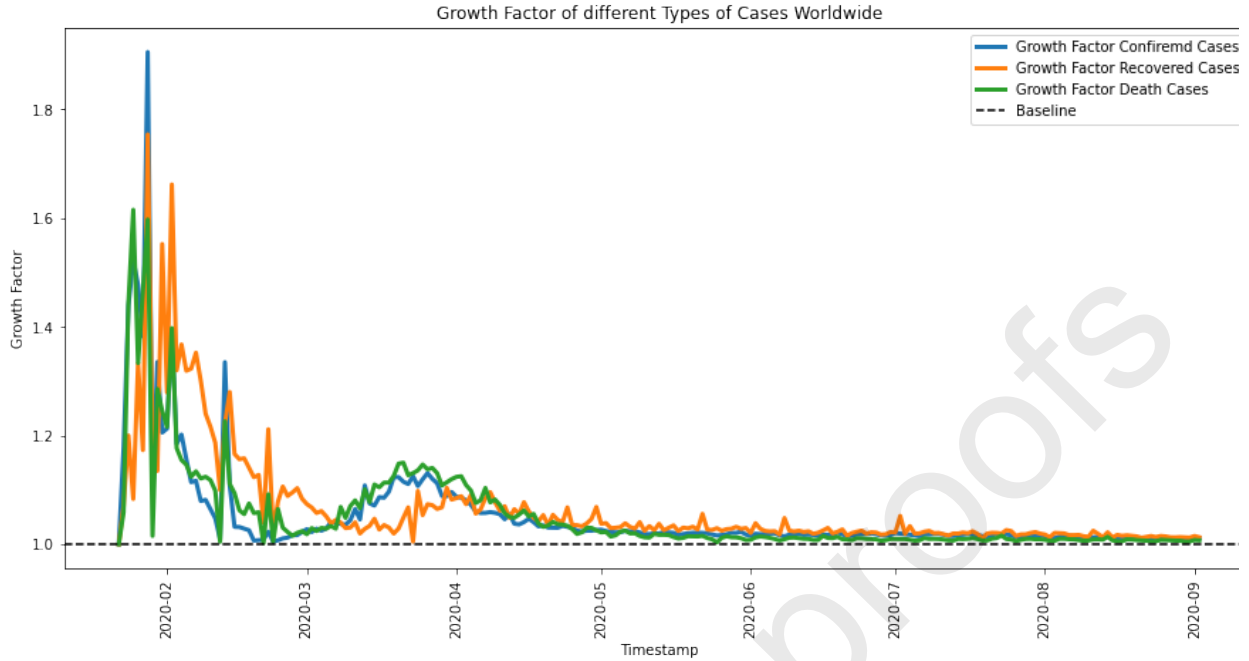
**Fig. 5:** (Top): Top 15 countries according to high recovery rate, (Bottom): 15 countries according to low recovery rate.

According to statistical analysis done by the author's, the following key facts have been observed.

- (a) Top 15 Countries according to High Mortality Rate are Yemen, Peru, Mexico, Sudan, Syria, Egypt, Somalia, Taiwan, Bosnia and Herzegovina, Ecuador, China, Bulgaria, Afghanistan, Tanzania and Bolivia. Yemen has the highest mortality rate all over the world.
- (b) Top 15 Countries according Low Mortality Rate are Bahrain, Mongolia, Entrea, Cyprus, Seychelles, Iceland, Vietnam, UAE, Maldives, Qatar, Timor-Leste, Burundi, Laos, Singapore and Bhutan. Bhutan has the lowest mortality rate all over the world.
- (c) Top 15 Countries according to the High Recovery Rate are Singapore, Qatar, Bahrain, Gabon, Iceland, Israel, Djibouti, Cote d'Ivoire, Burkina Faso, Belarus, West Bank and Gaza, San Marino, Diamond Princess, Peru and Albania. Singapore has the highest recovery rate all over the world.
- (d) Top 15 Countries according Low Recovery Rate are Vietnam, Greece, Fiji, Burundi, Norway, Iceland, Thailand, France, Spain, Netherlands, United Kingdom, Serbia, Sweden, Belgium and US. US has the lowest recovery rate all over the world.

#### *Growth factor*

Growth factor is a proportion of how rapidly the quantity of new cases is rising or falling and the basic thing to recall is that we need to keep it fewer than one. Critically, one thing that has changed since the last time the development factor was over one is that the quantity of new cases every day is low. It is simply the factor by which an amount duplicates itself after some time. Fig. 6 shows the growth factor of different types of cases worldwide.



**Fig. 6:** Growth factor of different types of cases worldwide

### *Doubling Rate*

Multiplying of cases is additionally a significant factor of COVID-19. Perceiving the doubling rate can make the government and individuals aware about the severity of the disease and hence enhance the preventive measures. The following table shows the doubling rate of coronavirus.

Table 1 shows the doubling rate of coronavirus infected cases.

**Table 1:** Doubling rate of coronavirus infected cases.

S. N	No. of cases	Days Since first Case	Number of days for doubling
1	5000	5 days	5 days
2	10000	9 days	4 days
3	20000	12 days	3 days
4	40000	17 days	5 days
5	80000	33 days	16 days
6	160000	52 days	19 days
7	320000	59 days	7 days
8	640000	65 days	6 days
9	1280000	74 days	9 days

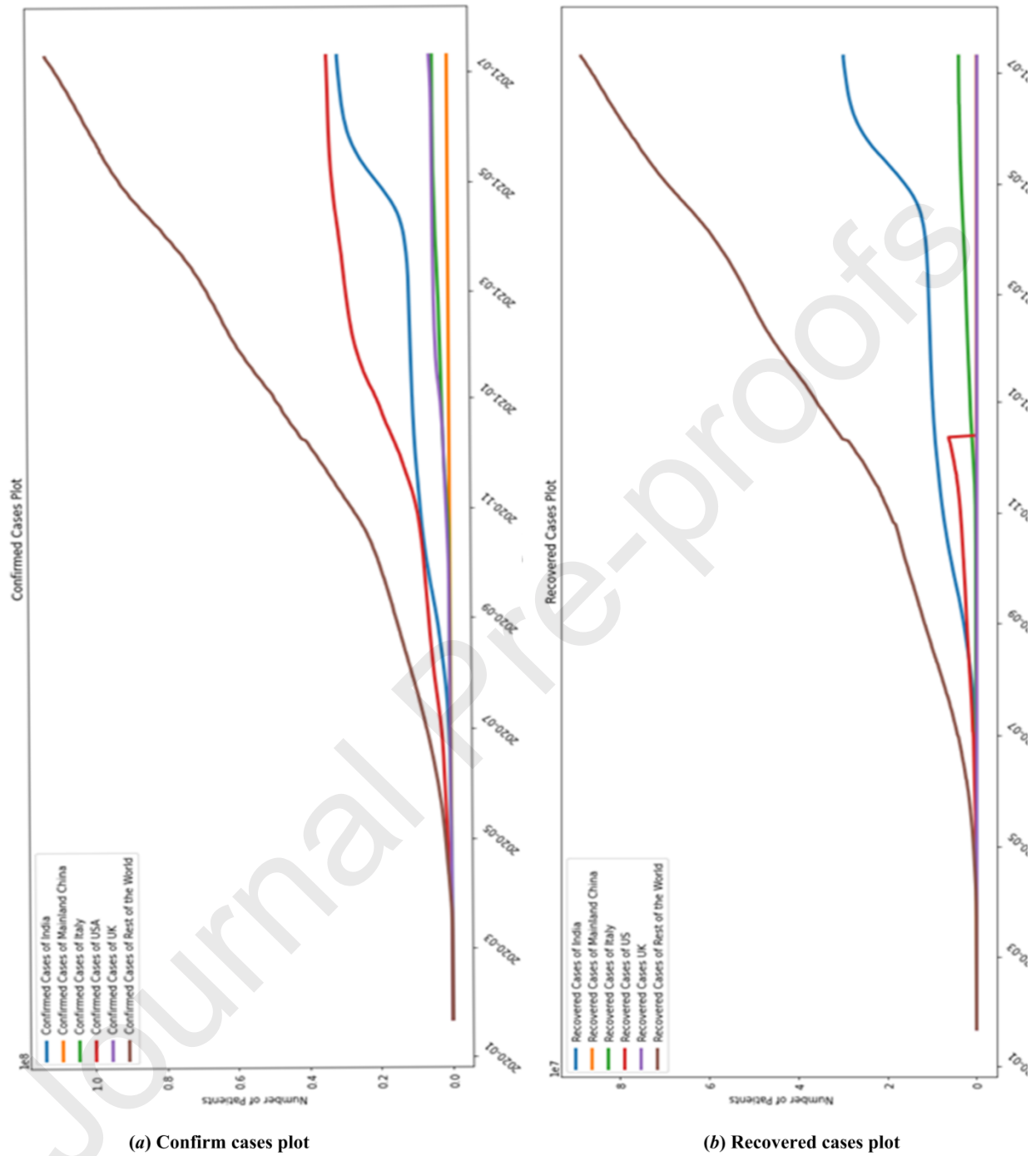
10	2560000	90 days	16 days
11	5120000	120 days	30 days
12	10240000	158 days	38 days
13	20480000	202 days	44 days
14	40960000	272 days	70 days
15	81920000	341 days	69 days
16	163840000	481 days	140 days

*Comparison of India, China, UK, Italy, US and rest of the world*

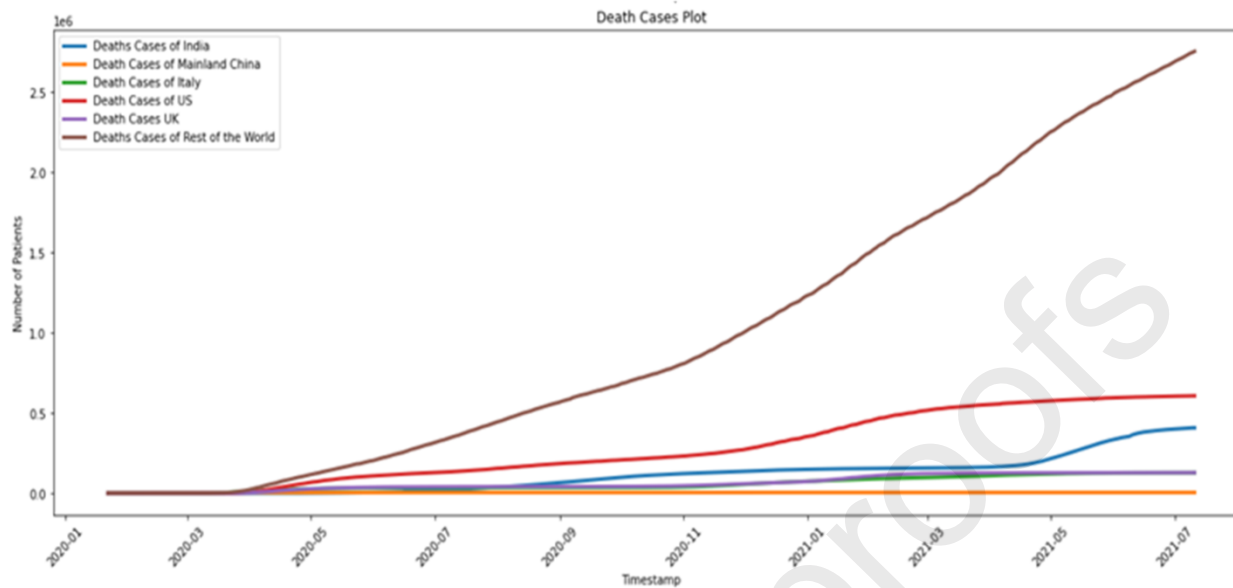
The above Fig. 7 which compares the confirmed, recovered and death cases of India, China, UK, Italy, US and rest of the world. We can infer from the graph that confirmed, recovered and death cases of the rest world is higher than the other five countries.

- (a) The blue line represents the confirmed, death and recovered cases of India. It is observed that confirmed cases and death cases- in India increased exponentially after March 2021. Recovery rate was maximum in July 2021.
- (b) The orange line represents the confirmed, death and recovered cases of Mainland China. It is observed that it maintains a constant rate.
- (c) The green line represents the confirmed, death and recovered cases of Italy. It is observed that the confirmed cases are constant but the recovered and death cases have increased after March 2021.
- (d) The red line represents the confirmed, death and recovered cases of US. It is observed that the confirmed cases have increased exponentially whereas the death cases and recovered cases have increased after January 2021.
- (e) The purple line represents the confirmed, death and recovered cases of UK. It can be inferred that the death cases have increased gradually after November 2020.

The brown line shows the confirmed, death and recovered cases of rest of the world. It is observed that confirmed, death and recovered cases have increased exponentially with time.



**Fig. 7 (a):** Comparison of 5 countries with rest of world on the basis of confirmed and recovered cases

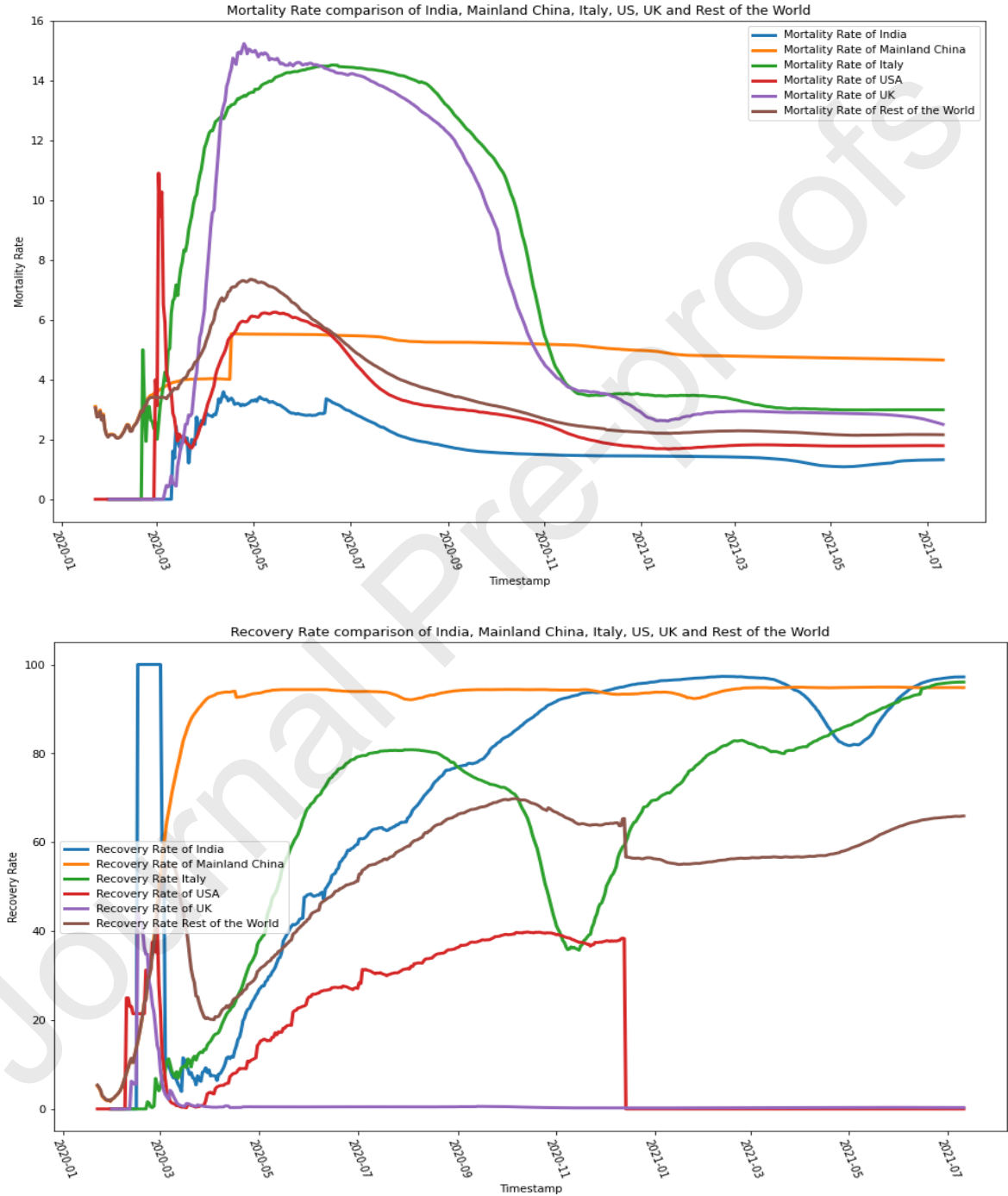


**Fig. 7 (b):** Comparison of 5 countries with rest of world on the basis of death cases

The above Fig. 7 (a) and 7 (b) which compares the confirmed, recovered and death cases of India, China, UK, Italy, US and rest of the world. We can infer from the graph that confirmed, recovered and death cases of the rest world is higher than the other five countries.

- (a) The blue line represents the confirmed, death and recovered cases of India. It is observed that confirmed cases and death cases- in India increased exponentially after March 2021. Recovery rate was maximum in July 2021.
- (b) The orange line represents the confirmed, death and recovered cases of Mainland China. It is observed that it maintains a constant rate.
- (c) The green line represents the confirmed, death and recovered cases of Italy. It is observed that the confirmed cases are constant but the recovered and death cases have increased after March 2021.
- (d) The red line represents the confirmed, death and recovered cases of US. It is observed that the confirmed cases have increased exponentially whereas the death cases and recovered cases have increased after January 2021.
- (e) The purple line represents the confirmed, death and recovered cases of UK. It can be inferred that the death cases have increased gradually after November 2020.

- (f) The brown line shows the confirmed, death and recovered cases of rest of the world. It is observed that confirmed, death and recovered cases have increased exponentially with time.



**Fig. 8:** Comparison of 5 countries with rest of world on the basis of Mortality rate and Recovery rate

The above graph shown in Fig. 8, presents the mortality rate and recovery rate comparison of 5 countries with the rest of the world. Mortality rate of U.K has increased in the timestamp whereas India has least mortality rate. Recovery rate of mainland China is highest and U.K has least recovery rate.

- (a) Blue line represents the mortality and recovery rate of India. It is observed that the mortality rate increased from mid of April 2020 till mid of September 2020 and then it started decreasing. The recovery rate decreased from mid of March to mid of May and then started increasing gradually.
- (b) Orange line represents the mortality and recovery rate of Mainland China. It is observed that the mortality rate increased steeply after May 2020. The recovery rate is represented by normal yield curve.
- (c) Green line represents the mortality and recovery rate of Italy. It is observed that mortality rate increased after April 2020 and then decreased in mid of November 2020 and recovery rate increased after November 2020 in Italy.
- (d) Red line represents the mortality and recovery rate of US. It is observed that mortality rate increased gradually from March till mid of April 2020.
- (e) Purple line represents the mortality rate and recovery rate of UK. It is observed that mortality rate has decreased after November 2020 and recovery rate is constant after mid of December 2020.
- (f) Brown line represents the mortality rate and recovery rate of rest of the world. It is observed that mortality rate has decreased after September 2020 and recovery rate has increased after September 2020.

### 3.2.2 REGRESSION MODELS

In this work unsupervised machine learning algorithm is used for the prediction of future cases of COVID-19 in terms of confirm cases, recovered cases and death cases. For this study, linear regression ((LR), polynomial Regression (PR), Ridge and Lasso regression (RLR), Ridge Regression (RR) and Support vector machine (SVM) regression models are used. The brief description of each model is given here.

#### *Linear Regression (LR)*



It is an immediate method to manage showing the association between a scalar response (or ward variable) and at any rate one instructive factor (or free variables). It was the essential kind of backslide examination to be thought completely, and to be used broadly in feasible applications. This is because models which depend legitimately upon their dark limits are less difficult to fit than models which are non-straightforwardly related to their limits and considering the way that the real properties of the resulting estimators are easier to choose. One of the information digging methods utilized for expectation undertakings is Linear Regression [22, 23, 24, 25]. In an issue with one indicator, this procedure attempts to locate the best line to fit. The formula to calculate linear regression is given in Eq. 1.

$$\hat{y} = b_0 + b_1x \quad (1)$$

Where,  $b_0$  is a constant,  $b_1$  is the regression coefficient;  $x$  is the independent variable and  $\hat{y}$  is the predicted value of dependent variable.

#### *Polynomial Regression*

In estimations, it is a kind of relapse examination wherein the association between the free factor  $x$  and the poor variable  $y$  is shown as a farthest breaking point polynomial in  $x$ . It fits a nonlinear association between the estimation of  $x$  and the relating unforeseen mean of  $y$ , implied  $E(y|x)$ . Even though polynomial regression fits a nonlinear model to the data, as a quantifiable estimation issue it is immediate, as in the backslide work  $E(y|x)$  is straight in the dark limits that are assessed from the data [26,27,28]. Subsequently, polynomial relapse is seen as an uncommon case of various direct backslides. It can be calculated according to Eq. 2.

$$\hat{y} = q_1 + q_1x_1 + q_2x_1^2 + ..... + q_nx_1^n \quad (2)$$

Where,  $\hat{y}$  is the predicted value of dependent variable,  $x_1^1, x_1^2, x_1^3, ..... x_1^n$  are independent variables and  $q_1, q_2, q_3, ..... q_n$  are all coefficients.

#### *Ridge and Lasso Regression*

For some, machine learning issues with an enormous number of highlights or a low number of perceptions, a direct model tends to overfit and variable determination is dubious. Models that utilize shrinkage, for example, Lasso and Ridge can improve the forecast precision as they diminish the estimation difference while giving an interpretable last model. Ridge and Lasso

expand on the direct model; however, their basic idiosyncrasy is regularization. The objective of these techniques is to improve the misfortune work with the goal that it depends not just on the whole of the squared contrasts yet in addition on the relapse coefficients [29, 30, 31].

The primary thing in the advancement of such a framework is the right determination of the regularization boundary. Compared to Linear Regression, Ridge and Lasso models are progressively impervious to anomalies and the spread of information. Generally speaking, their fundamental reason for existing is to forestall overfitting. The primary distinction between Ridge relapse and Lasso is the means by which they dole out a punishment term to the coefficients.

### *Ridge Regression*

The Ridge regression is a method which is specific to multilinear regression information which is multicollinearity in nature. Ridge Regression is one of the most head regularization methodologies which isn't used by various people due to the multifaceted science behind it [32, 33, 34]. On the off chance that you have a general thought regarding the idea of different relapse, it's not all that hard to investigate the science behind Ridge relapse in r. Relapse is the equivalent, what makes regularization diverse is that the way how the model coefficients are resolved. The equation for this technique is given in Eq. 3.

$$\cos t(a) = \frac{1}{2b} \sum_{p=1}^{p=b} (z_p - \hat{z}_i)^2 + \lambda \sum_q^A a_q^2 \quad (3)$$

This sets the coefficient that can be said as min (whole of square residuals +  $\lambda$  |slope|), where,  $\lambda$  |slope| is punishment term.

### *Support Vector Regression*

They are astounding yet versatile controlled AI figuring which are used both for course of action and backslide. SVMs have their exceptional strategy for execution when diverged from other AI estimations. As of late, they are incredibly notable because of their ability to manage different relentless and obvious components. A SVM model is basically a depiction of different classes in a hyperplane in multidimensional space [35, 36]. The hyperplane will be delivered in an iterative manner by SVM with the objective that the goof can be constrained. The goal of SVM is to parcel the datasets into classes to find the biggest immaterial hyperplane (MMH).

*RBF kernel:* The RBF Kernel is additionally called the Gaussian part. There's a boundless number of measurements within the feature space since it can be extended by the Taylor Arrangement. Within the arrange underneath, the  $\gamma$  parameter characterizes how much impact a single preparing illustration has. The larger it is, the closer other examples must be to be influenced [37]. It may be a general-purpose part; utilized when there's no earlier information around the data. Mathematically, it is represented as given in Eq. 4.

$$M(a, a') = \exp \left( - \frac{\|a - a'\|^2}{2\sigma^2} \right) \quad (4)$$

$\|a - a'\|^2$  denotes the squared Euclidean distance between two feature vectors.  $\sigma$  is a free parameter.

*Polynomial Kernel:* The polynomial kernel looks not as it were at the given highlights of input tests to decide their similitude [38], but too combinations of these". With  $n$  unique highlights and  $k$  degrees of polynomial, the polynomial bit yields  $n^d$  extended highlights where  $d$  is the degree of polynomial. It is prevalent in image processing. Mathematically, it is represented as given in Eq. 5.

$$a(t_c - t_d) = (t_c \cdot t_d + 1)^k \quad (5)$$

Where  $k$  is the degree of polynomial

*Sigmoid Kernel:* The Hyperbolic Tangent Kernel is additionally known as the Sigmoid kernel and as the Multilayer Perceptron (MLP) kernel [42]. The Sigmoid kernel comes from the Neural Systems field, where the bipolar sigmoid work is frequently utilized as an enactment function for artificial neurons. Mathematically, it is represented as given in Eq. 6.

$$h(a, b) = \tanh(\alpha a^D b + i) \quad (6)$$

### *Support Vector Regression*

Support Vector Regression is like Linear Regression in that the condition of the line is  $y = wx + b$  in SVR, this straight line is alluded to as a hyperplane. The data centres around either side of the hyperplane that are closest to the hyperplane are called Support Vectors which are used to plot the breaking point line.

Not at all like other Regression models that endeavour to constrain the bumble between the veritable and foreseen worth, the SVR endeavours to fit the best line inside an edge regard (Distance among hyperplane and limit line), a. Consequently, we can say that SVR model attempts fulfil the following condition:

$$-a < y < (y - wx + b) < a$$

It utilized the focuses with this limit to anticipate the worth.

### 3.3 Staging Analysis Parameters

*RMSE*: It is the standard deviation of the residuals. Residuals are a proportion of how a long way from the relapse line information focuses is. RMSE is an extent of how spread out these residuals is. Figuratively speaking, it uncovers to you how centred the data is around the line of best fit [40, 41]. Root mean square error is regularly utilized in climatology, anticipating, and relapse investigation to check test results. The formula to calculate RMSE is given in Eq. 7.

$$RMSE = \sqrt{\frac{1}{k} \sum_{m=1}^k (a_m - \hat{a}_m)^2} \quad (7)$$

Where,  $k$  is the number of observations,  $a_m$  is the observed value, and  $\hat{a}_m$  is the predicted value.

*R2\_SCORE*: It is a factual proportion of how close the information is to the fitted relapse line. It is otherwise called the coefficient of assurance, or the coefficient of various assurances for numerous relapses. It is always between 0 and 100%. 0% represents that the model clarifies none of the changeability of the reaction information around its mean and 100% shows that the model clarifies all the fluctuation of the reaction information around its mean. The mathematical expression of *R2\_SCORE* is given in Eq. 8.

$$R2\_SCORE = 1 - \frac{\sum (t_i - \hat{t}_i)^2}{(t_i - \bar{t}_i)^2} \quad (8)$$

Where,  $t_i$  is the actual cumulative confirmed cases,  $\hat{t}_i$  is the predicted cumulative confirmed cases,  $\bar{t}_i$  is the average of the actual cumulative confirmed cases.

*Mean Absolute Error*: It measures the normal size of the errors in a lot of forecasts, without thinking about their heading [39]. It's the normal over the test of the outright contrasts among

expectation and genuine perception where every individual distinction has equivalent weight. The obtained value of MAE is calculated according to Eq. 9.

$$\text{MAE} = \frac{1}{p} \sum_{q=1}^p |r_q - r| \quad (9)$$

Where, p represents the number of errors,  $|r_q - r|$  denotes the absolute errors.

*Mean Squared Error:* MSE of an estimator is the normal squared contrast between the assessed qualities and the real worth. The MSE is an extent of the idea of an estimator. It is reliably non-negative, and characteristics increasingly like zero are better. MSE is calculated according to Eq. 10.

$$\text{MSE} = \frac{1}{z} \sum_{s=1}^z (y_s - \hat{y}_s)^2 \quad (10)$$

Where, z is the number of data points,  $y_s$  represent the observed values, and  $\hat{y}_s$  represent predicted value.

## 4. EXPERIMENTS AND RESULTS

### 4.1 Dataset preparation and Experiment

To make our data perfect with sklearn design, we made another section called "Days since" which tracks the quantity of days since the initial date. We have taken four evaluation parameters namely Mean Square Error (MSE), Mean Absolute Error (MAE), Root Mean Square Error (RMSE), r2\_score for identification of the best regression model which predicts the best about future cases, recovered cases and deaths cases. The brief description of experiment carried out for the work is given in Table 2.

**Table 2:** Brief Description of Experiment

Experiment No. #	Description
Experiment 1	Experiment carried out for active cases using five different regression models.
Experiment 2	Experiment carried out for recovered cases using five different regression models.
Experiment 3	Experiment carried out for death cases using five different regression models.

This investigation endeavours to build up a framework for the future gauging of the quantity of cases influenced by COVID-19 utilizing AI strategies. The dataset utilized for the investigation contains data about the day by day reports of the number of recently contaminated cases, the quantity of recuperations, and the quantity of passing due to COVID-19 around the world. We have utilized five AI models LR, polynomial, Ridge, polynomial ridge and SVM to foresee the quantity of recently affected cases, the quantity of passing, and the quantity of recuperations. We additionally attempt to anticipate the best model for gauging affected cases, recuperated cases and the passing cases.

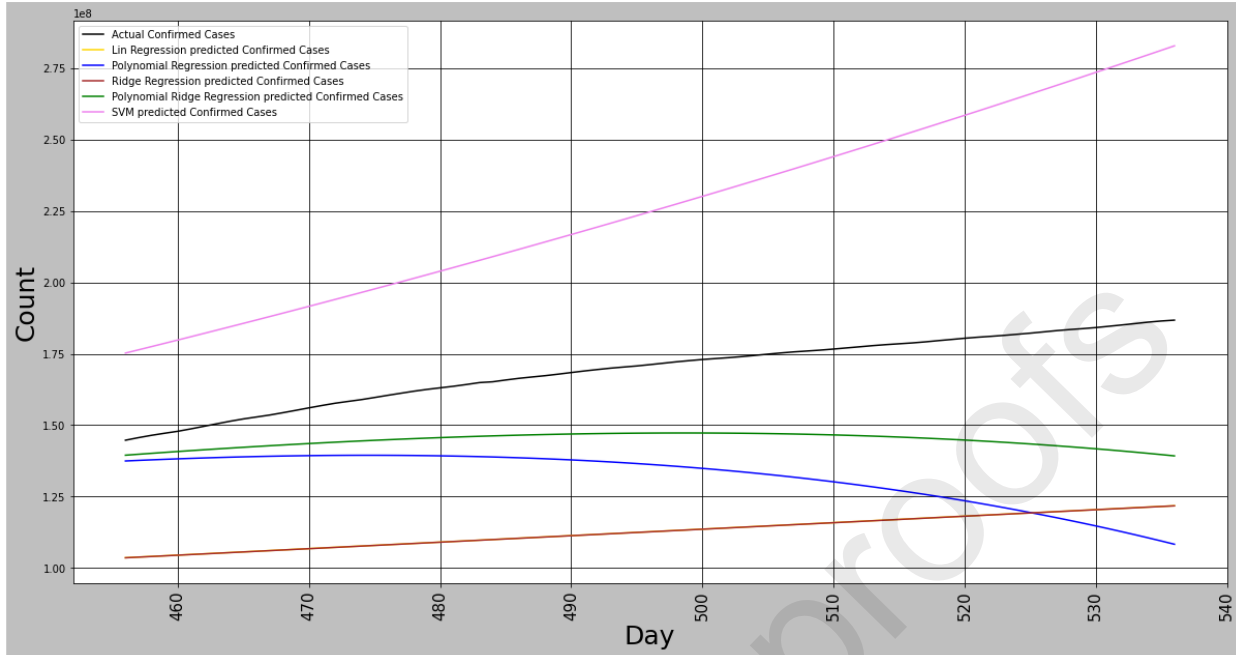
#### 4.2 Results analysis

*Experiment 1:* The investigation performs forecasts on infected cases and concurring to results polynomial ridge performs better among all the models; SVM additionally performs well and polynomial relapse performs normal score. In correlation, LR and Ridge performs most exceedingly terrible in this circumstance and accomplish nearly the same R2\_score. Be that as it may, contrasting the current confirmed cases statistics and our models' forecasts, the polynomial ridge expectation is following the patterns which are near the genuine circumstance.

**Table 3:** Obtained results for Confirmed cases using different regression model.

Model Name	MSE	MAE	RMSE	R2_score
Linear Regression	3230371601512722.5	56457125.4	56836358.0	-113.4
Polynomial Regression	1795330053135026.2	37454555.9	42371335.2	-21.0
Ridge	3233647448477261.5	56485978.2	56865169.0	-113.5
Polynomial ridge	737383324668728.4	24617677.8	27154802.9	-125.5
SVM	3645218671704054.5	56956658.6	60375646.3	-2.6

Results of the extensive experiments have been reported in Table 3 for active cases.

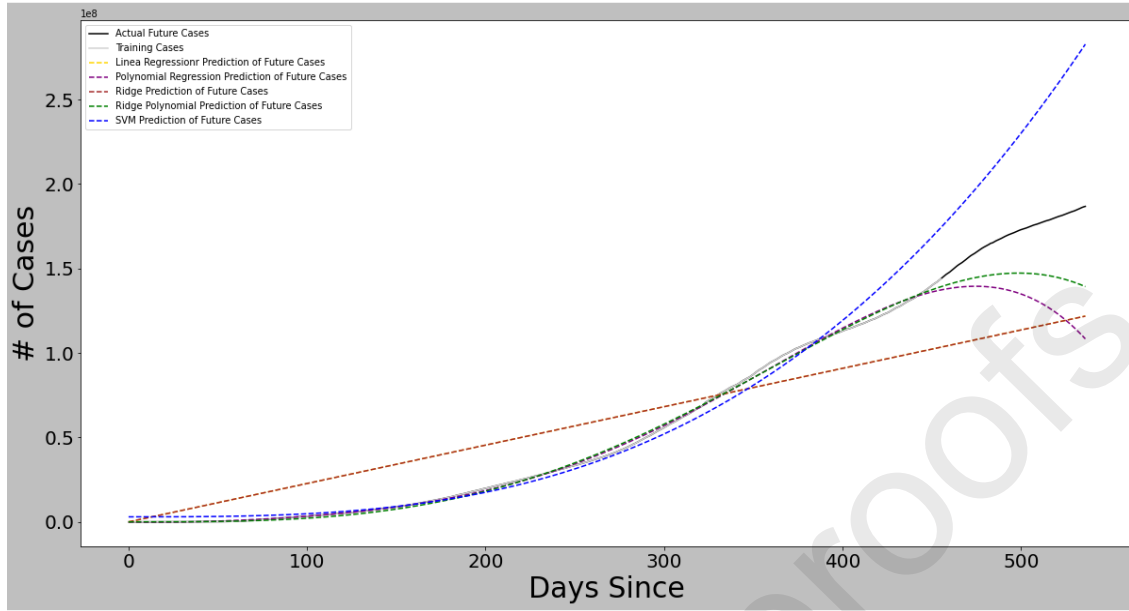


**Fig. 9:** Performance of Different Regression Models on Prediction of confirmed cases on the testing set

Fig. 9 shows the performance of all regression models on the prediction of confirmed cases on the basis of the testing set of data. From the figure we can visualize that the infected cases tend to increase day by day which is an alarming sign for all the infected countries.

- (a) Black dotted line represents the actual confirmed cases on the basis of testing set.
- (b) Yellow line represents the LR and Red line represents the ridge regression and the prediction of both the models are very much similar and overlap each other.
- (c) Purple line represents polynomial regression predicted confirmed cases on testing set.
- (d) Green line represents polynomial ridge regression predicted confirmed cases on testing set.
- (e) Blue represents the SVM (kernel=poly) predicted confirmed cases on testing set.





**Fig. 10:** Performance of Different Regression Models on prediction of future cases

Fig. 10 shows the performance of all the models on the basis of trained data. From this we have examined that polynomial ridge, SVM and polynomial predicts good result and following the pattern which are near to actual confirmed cases. LR and ridge predict similar result and does not follow the pattern of actual cases.

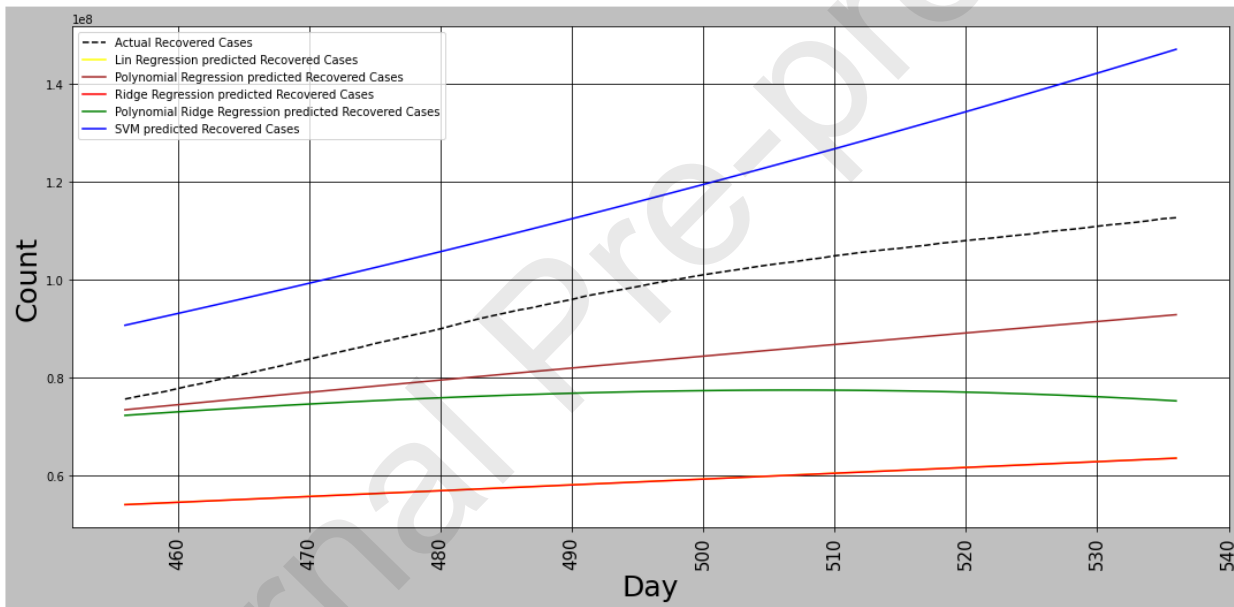
- (a) Here, black line represents actual future cases and grey line represents the training cases.
- (b) Yellow line represents the LR and Brown line represents the ridge regression and the future prediction of both the models are very much similar and overlap each other.
- (c) Purple line represents the prediction of confirmed cases using polynomial regression.
- (d) Green line represents the prediction of confirmed cases using polynomial ridge regression.
- (e) Blue line represents the prediction of confirmed cases using SVM.

*Experiment 2:* The investigation performs forecasts on Recovered cases and concurring to results polynomial ridge performs better among all the models; polynomial Regression additionally performs well and SVM performs normal score. In correlation, LR and Ridge performs most exceedingly terrible in this circumstance and accomplish nearly the same  $R^2$  score.

**Table 4:** Obtained results for recovered cases using different regression model.

Model Name	MSE	MAE	RMSE	R2_score
Linear Regression	1524067870205389.0	38118611.9	39039311.8	-197.1
PolynomialRegression	217065302458256.3	13627274.2	14733136.2	-5.7
Ridge	1525190339272752.2	38133183.5	39053685.3	-197.3
Polynomial ridge	539545477473680.5	20924557	23228118.2	-263.5
SVM	451722886562148.9	20420131.9	21253773.4	-0.6

Table 4 depicts the result of performance of all models and polynomial ridge gives best fit values of MSE, MAE, RMSE and R2\_score.

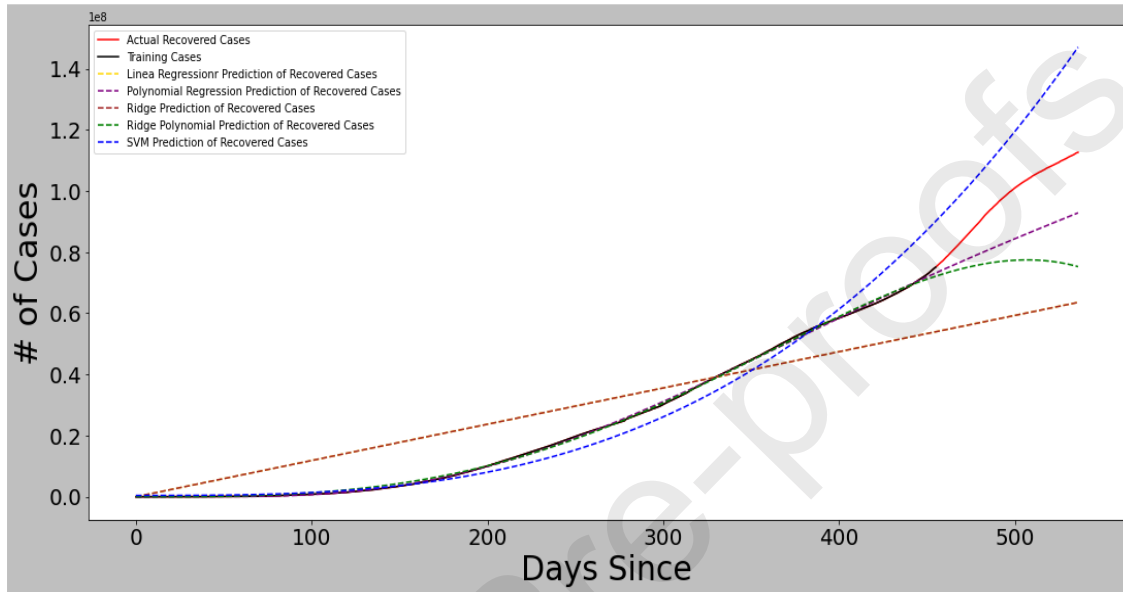


**Fig. 11:** Performance of Different Regression Models on Prediction of Recovered cases on the testing set

Fig. 11 shows the performance of LR, polynomial, Ridge, polynomial ridge and SVM in the form of graph. Be that as it may, contrasting the current recuperation insights and our models' expectations, the polynomial ridge forecast is following the patterns which are near the genuine circumstance.

- (a) Black dotted line represents the actual recovered cases on the basis of testing set.
- (b) Yellow line represents LR and red represents ridge regression having similar pattern of prediction of recovered cases.

- (c) Brown line represents polynomial regression recovered cases prediction on testing set.
- (d) Green line represents polynomial ridge prediction of recovered cases on the basis of testing set.
- (e) Blue line represents SVM prediction of recovered cases on the basis of testing set.



**Fig. 12:** Performance of Different Regression Models on Prediction of recovered cases

From Fig. 12 we observe that polynomial and polynomial ridge following the pattern which is almost similar to actual recovered cases and SVM gives average performance but the pattern of linear and ridge regression is almost different from actual recovered cases.

- (a) Here, Crimson line represents actual recovered cases and Black line represents the training cases.
- (b) Yellow line represents the LR and Brown line represents the ridge regression and the prediction of recovered cases of both the models are very much similar and overlap each other.
- (c) Purple line represents the polynomial prediction of recovered cases.
- (d) Green line represents the polynomial ridge prediction of recovered cases.
- (e) Blue line represents the SVM prediction of recovered cases.

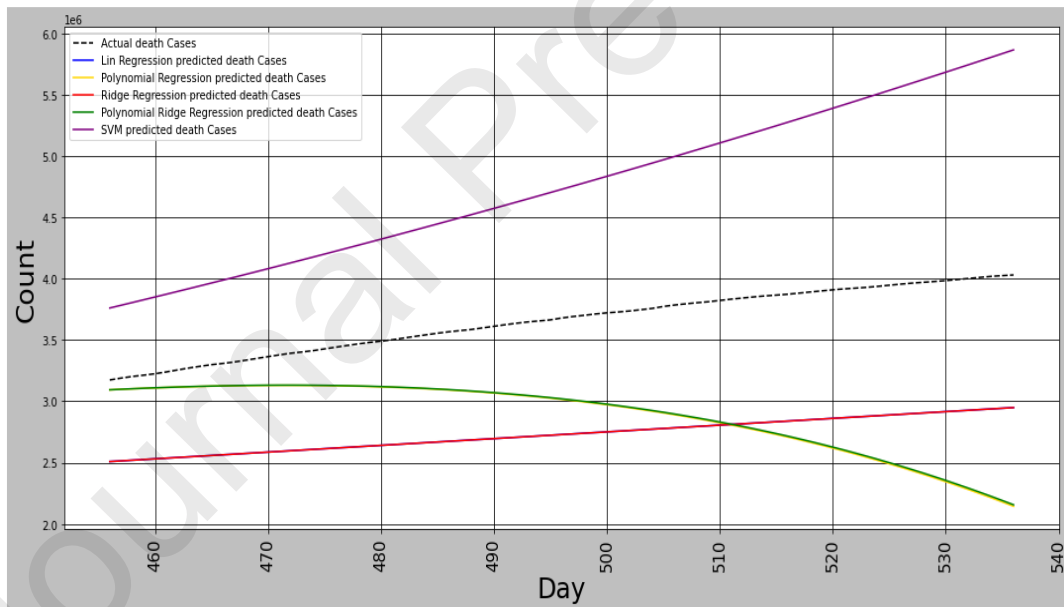
*Experiment 3:* In death rate future estimation of the polynomial ridge performs better among the various models. Every single other model performs ineffectively, the request for execution from

best to most noticeably awful is polynomial ridge is best trailed by LR, Ridge, polynomial and SVM because of the nature of accessible time-arrangement information. From our observation we also observed that no model can predict the best results for death rate prediction.

Table 5: Obtained results for death cases using different regression model.

Model Name	MSE	MAE	RMSE	R2_score
Linear Regression	866756405380.5	922497.6	930997.5	-51.4
Polynomial Regression	865638169852.0	770365.6	930396.8	-9.4
Ridge	867288071.97.1	922783.9	931283.0	-51.4
Polynomial ridge	854566130116.6	765500.4	924427.4	-9.5
SVM	1359199418014.9	1106760.1	1165847.0	-2.5

Table 5 depicts the result of performance of all models and linear and ridge gives best fit values of MSE, MAE, RMSE and R2\_score.

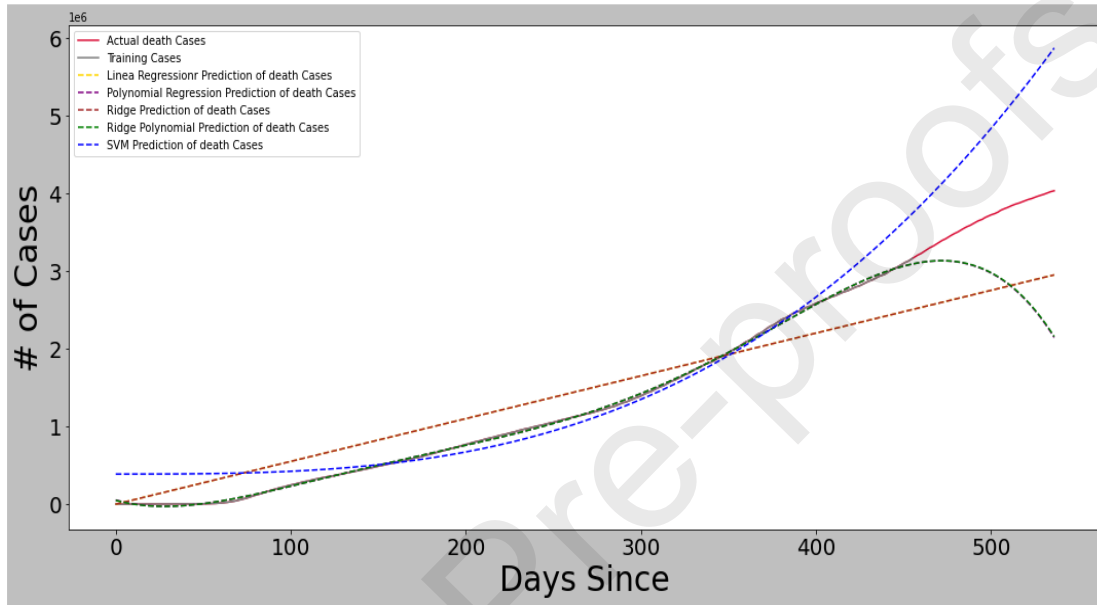


**Fig. 13:** Performance of Different Regression Models on Prediction of Death cases on the testing set

Fig. 13 shows the performance of Different Regression Models on Prediction of Death cases on the testing set. From this we conclude that not a single regression model gives us best result.

a) Black dotted line represents the actual death cases on the basis of testing set.

- b) Blue line represents LR and red represents ridge regression having similar pattern of prediction of death cases.
- c) Yellow line represents polynomial regression death cases prediction on testing set.
- d) Green line represents polynomial ridge prediction of death cases on the basis of testing set.
- e) Blue line represents SVM prediction of death cases on the basis of testing set.



**Fig. 14:** Performance of Different Regression Models on Prediction of death cases

From Fig. 14 we concluded that till now we cannot any model which follows the pattern of actual death cases. Only polynomial ridge model comes out with better result.

- (a) Here, Red line represents actual death cases and Black line represents the training cases.
- (b) Yellow line represents the LR and Brown line represents the ridge regression and the prediction of death cases of both the models are very much similar and overlap each other.
- (c) Purple line represents the polynomial prediction of death cases.
- (d) Green line represents the polynomial ridge prediction of death cases.
- (e) Blue line represents the SVM prediction of death cases.

Linear and Ridge Regression models' results are very similar to each other and they predict very similar  $r^2\_score$  value in all the three cases.

#### 4.3 Comparative Analysis with state of art

Table 6 represent the comparative analysis of the work carried out for this study with previously published work.

**Table 6:** Comparative analysis of the work carried out for this study with previously published work

Parameters	F. Rustam et al., 2020, [9]	S. Sengupta et al., 2020, [10]	R. Sujath et al., 2020, [11]	Present study
World Data Visualization on the basis of Different Factors	NO	Yes	They do only for India	YES
Comparison of 5 countries with rest of the world	NO	NO	NO	YES
Growth Factor of different type of Cases World-wide	NO	YES	NO	YES
Data analysis of India and Comparison With other countries	NO	Done but not Compare with Other Countries	Done but not Compare with Other Countries	YES
Time Taken for Doubling of Infected Cases	NO	NO	NO	YES
Future Prediction using 5 Models	Only with four models	Only with three models	ONLY with three models	YES

Finally, comparative analysis of the present with previously published work [9, 10, 11] is performed and the major findings of the study is given as:

- In the study, visualization of Active Cases, closed cases (recovered and death) all over the world has been reported.
- The study also performs the visualization of weekly and daily progress of different types of cases in the world.
- The work also highlights the country wise confirmed cases, Death cases, Mortality rate, Recovery Rate.
- Comparison of India, China, US, UK, Italy has been done with rest of the world.
- The study has also visualized the growth factor of different cases World-wide.

- f.* Analysis on India and comparison with other countries to find out how much time other countries took to reach a certain number of confirmed cases in comparison to India has been done.
- g.* Prediction of the time taken for doubling the number of confirmed cases has also been done.
- h.* Future prediction using five regression models and analysis on the premise of Mean square Error, Mean Absolute Error has been done.
- i.* From the comparison Model it can be inferred that no alternative has done as several things as through this model.

## 5. Conclusion

This examination introduced current patterns of COVID-19 occurrence from twenty-second February 2020 to 2<sup>nd</sup> September 2020 as envisioned in our task. The quickly expanding number of latest COVID-19 cases day by day worldwide has placed an overwhelming weight on clinical assets in nations with enormous flare-ups. Therefore, prediction of future confirmed cases became necessary. The size of information accessible is gigantic associate degree gathering knowledge and obtaining an intriguing example out of the cumulated information could be a tough trip. With the common info regarding confirmed, recovered and ending across Republic of India for over the time term helps in anticipating and deciding the not thus distant future. Our outcomes imply that there is a sure comprehensiveness in the time development of Covid-19.

This proposes a nation that turns into the venue of a scourge flood can be regarded, in any event in first guess, as a very much mixed synthetic reactor, where various populaces associate as indicated by mass-activity like principles with little association with topographical varieties.

In light of itemized examination of the overall information, we gauge a few key boundaries for COVID-19, similar to the inactive time, the isolate time and the compelling propagation number in a moderately dependable manner, and foresee the inflection point, conceivable completion time and final absolute tainted cases for everywhere throughout the world. We have done prediction by using 5 regression models so that we can conclude which model is best and on the basis of that model we are able to tell the rate of increase in number of infected, recovered and death cases in future. The experiments have been completed with best MSE value as 737383324668728.4, MAE as 24617677.8, RMSE as 27154802.9 and R2\_score as -125.5 using polynomial Ridge model for



confirmed cases. For recovered cases best MSE value as 1524067870205389.0, MAE value as 38118611.9, RMSE as 39039311.8 and R2\_score as -197.1 using linear regression model. Also, for Death cases best MSE value as 1359199418014.9, MAE value as 1106760.1, RMSE value as 1165847.0 and R2\_score as -2.5 using SVM model. From our investigation we concluded that polynomial ridge is the best model for predicting confirmed cases, linear regression model for recovered cases and SVM regression model for death cases.

## References:

- [1] F. Petropoulos and S. Makridakis, "Forecasting the novel coronavirus COVID-19", *PLOS ONE*, vol. 15, no. 3, p. e0231236, 2020, doi: 10.1371/journal.pone.0231236.
- [2] M. A. A. Al-qaness, A. A. Ewees, H. Fan, and M. Abd El Aziz, "Optimization Method for Forecasting Confirmed Cases of COVID-19 in China", *Journal of Clinical Medicine*, vol. 9, no. 3, p. 674, 2020, doi: 10.3390/jcm9030674.
- [3] D. Fanelli and F. Piazza, "Analysis and forecast of COVID-19 spreading in China, Italy and France", *Chaos, Solitons & Fractals*, vol. 134, p. 109761, 2020, doi: 10.1016/j.chaos.2020.109761.
- [4] M. Sperrin, S. W. Grant, and N. Peek, "Prediction models for diagnosis and prognosis in Covid-19", *BMJ*, p. m1464, 2020, doi: 10.1136/bmj.m1464.
- [5] R. Gupta, G. Pandey, P. Chaudhary, and S. K. Pal, "SEIR and Regression Model based COVID-19 outbreak predictions in India", 2020, doi: 10.1101/2020.04.01.20049825.
- [6] M. Perc, N. GorišekMiksić, M. Slavinec, and A. Stožer, "Forecasting COVID-19", *Frontiers in Physics*, vol. 8, 2020, doi: 10.3389/fphy.2020.00127.
- [7] C. Menni, "Real-time tracking of self-reported symptoms to predict potential COVID-19", *Nature Medicine*, vol. 26, no. 7, pp. 1037-1040, 2020, doi: 10.1038/s41591-020-0916-2.
- [8] F. A. B. Hamzah, "CoronaTracker: World-wide COVID-19 Outbreak Data Analysis and Prediction", 2020, doi: 10.2471/blt.20.255695.
- [9] F. Rustam, "COVID-19 Future Forecasting Using Supervised Machine Learning Models", *IEEE Access*, vol. 8, pp. 101489-101499, 2020, doi: 10.1109/access.2020.2997311.

- [10] S. Sengupta, S. Mugde, and G. Sharma, "Covid-19 Pandemic Data Analysis and Forecasting using Machine Learning Algorithms", 2020, doi: 10.1101/2020.06.25.20140004.
- [11] R. Sujath, J. M. Chatterjee, and A. E. Hassanien, "A machine learning forecasting model for COVID-19 pandemic in India", *Stochastic Environmental Research and Risk Assessment*, vol. 34, no. 7, pp. 959-972, 2020, doi: 10.1007/s00477-020-01827-8.
- [12] Prediction and analysis of Coronavirus Disease 2019 Lin Jia, Kewen Li, Yu Jiang, Xin Guo, Ting zhaoh <https://arxiv.org/abs/2003.05447v2>
- [13] W. C. Roda, M. B. Varughese, D. Han, and M. Y. Li, "Why is it difficult to accurately predict the COVID-19 epidemic?", *Infectious Disease Modelling*, vol. 5, pp. 271-281, 2020, doi: 10.1016/j.idm.2020.03.001.
- [14] D. Ji, "Prediction for Progression Risk in Patients With COVID-19 Pneumonia: The CALL Score", *Clinical Infectious Diseases*, 2020, doi: 10.1093/cid/ciaa414.
- [15] V. K. R. Chimmula and L. Zhang, "Time series forecasting of COVID-19 transmission in Canada using LSTM networks", *Chaos, Solitons & Fractals*, vol. 135, p. 109864, 2020, doi: 10.1016/j.chaos.2020.109864.
- [16] C. Li, L. J. Chen, X. Chen, M. Zhang, C. P. Pang, and H. Chen, "Retrospective analysis of the possibility of predicting the COVID-19 outbreak from Internet searches and social media data, China, 2020", *Eurosurveillance*, vol. 25, no. 10, 2020, doi: 10.2807/1560-7917.es.2020.25.10.2000199.
- [17] L. Peng, W. Yang, D. Zhang, C. Zhuge, and L. Hong, "Epidemic analysis of COVID-19 in China by dynamical modeling", 2020, doi: 10.1101/2020.02.16.20023465.
- [18] D. Ivanov, "Predicting the impacts of epidemic outbreaks on global supply chains: A simulation-based analysis on the coronavirus outbreak (COVID-19/SARS-CoV-2) case", *Transportation Research Part E: Logistics and Transportation Review*, vol. 136, p. 101922, 2020, doi: 10.1016/j.tre.2020.101922.
- [19] N. S. Pun, S. K. Sonbhadra, and S. Agarwal, "COVID-19 Epidemic Analysis using Machine Learning and Deep Learning Algorithms", 2020, doi: 10.1101/2020.04.08.20057679.

- [20] S. K. Bandyopadhyay and S. Dutta, "Machine Learning Approach for Confirmation of COVID-19 Cases: Positive, Negative, Death and Release", 2020, doi: 10.1101/2020.03.25.20043505.
- [21] C. Sohrabi, "World Health Organization declares global emergency: A review of the 2019 novel coronavirus (COVID-19)", *International Journal of Surgery*, vol. 76, pp. 71-76, 2020, doi: 10.1016/j.ijssu.2020.02.034.
- Hayes, Andrew F., and Amanda K. Montoya. "A tutorial on testing, visualizing, and probing an interaction involving a multicategorical variable in linear regression analysis." *Communication Methods and Measures* 11.1 (2017): 1-30.
- [23] Schuld, Maria, Ilya Sinayskiy, and Francesco Petruccione. "Prediction by linear regression on a quantum computer." *Physical Review A* 94.2 (2016): 022342.
- [24] Schleich, Maximilian, Dan Olteanu, and Radu Ciucanu. "Learning linear regression models over factorized joins." *Proceedings of the 2016 International Conference on Management of Data*. 2016.
- [25] Groeneboom, Piet, and Kim Hendrickx. "Current status linear regression." *The Annals of Statistics* 46.4 (2018): 1415-1444.
- [26] Cheng, Xi, et al. "Polynomial regression as an alternative to neural nets." *arXiv preprint arXiv:1806.06850* (2018).
- [27] Aguirre-Urreta, Miguel I., Mikko Rönkkö, and Jiang Hu. "Polynomial Regression and Measurement Error: Implications for Information Systems Research." *ACM SIGMIS Database: the DATABASE for Advances in Information Systems* 51.3 (2020): 55-80.
- [28] Li, Hongran, and Shigeru Yamamoto. "Polynomial regression-based model-free predictive control for nonlinear systems." *2016 55th Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE)*. IEEE, 2016.
- [29] Yang, Xiaoxing, and Wushao Wen. "Ridge and lasso regression models for cross-version defect prediction." *IEEE Transactions on Reliability* 67.3 (2018): 885-896.
- [30] Melkumova, L. E., and S. YaShatskikh. "Comparing Ridge and LASSO estimators for data analysis." *Procedia engineering* 201 (2017): 746-755.

- [31] Pereira, Jose Manuel, Mario Basto, and Amelia Ferreira da Silva. "The logistic lasso and ridge regression in predicting corporate failure." *Procedia Economics and Finance* 39 (2016): 634-641.
- [32] Chowdhury, Agniva, Jiasen Yang, and Petros Drineas. "An iterative, sketching-based framework for ridge regression." *International Conference on Machine Learning*. 2018.
- [33] Kibria, B. M., and Shipra Banik. "Some ridge regression estimators and their performances." *Journal of Modern Applied Statistical Methods* 15.1 (2016): 12.
- [34] Pereira, Jose Manuel, Mario Basto, and Amelia Ferreira da Silva. "The logistic lasso and ridge regression in predicting corporate failure." *Procedia Economics and Finance* 39 (2016): 634-641.
- [35] Santamaría-Bonfil, Guillermo, A. Reyes-Ballesteros, and C. Gershenson. "Wind speed forecasting for wind farms: A method based on support vector regression." *Renewable Energy* 85 (2016): 790-809.
- [36] Cheng, Kai, et al. "Global sensitivity analysis using support vector regression." *Applied Mathematical Modelling* 49 (2017): 587-598.
- [37] Gopi, ArepalliPeda, et al. "Classification of tweets data based on polarity using improved RBF kernel of SVM." *International Journal of Information Technology* (2020): 1-16.
- [38] Lin, Shaobo, and Jinshan Zeng. "Fast learning with polynomial kernels." *IEEE transactions on cybernetics* 49.10 (2018): 3780-3792.
- [39] De Myttenaere, Arnaud, et al. "Mean absolute percentage error for regression models." *Neurocomputing* 192 (2016): 38-48.
- [40] Chai, Tianfeng, and Roland R. Draxler. "Root mean square error (RMSE) or mean absolute error (MAE)?—Arguments against avoiding RMSE in the literature." *Geoscientific model development* 7.3 (2014): 1247-1250.
- [41] Chai, Tianfeng, and Roland R. Draxler. "Root mean square error (RMSE) or mean absolute error (MAE)?." *GMDD* 7.1 (2014): 1525-1534.
- [42] Camps-Valls, Gustavo, et al. "Fuzzy sigmoid kernel for support vector classifiers." *Neurocomputing* 62 (2004): 501-506.

- [43] Khan, R., Shrivastava, P., Kapoor, A., Tiwari, A., & Mittal, "A. social media analysis with AI: sentiment analysis techniques for the analysis of twitter covid-19 data", *Journal of Critical reviews* 7.9 (2020),: 2761-2774.
- [44]. Gozes, Ophir, et al. "Rapid ai development cycle for the coronavirus (covid-19) pandemic: Initial results for automated detection & patient monitoring using deep learning ct image analysis." *arXiv preprint arXiv:2003.05037* (2020).
- [45] Shi, Feng, et al. "Review of artificial intelligence techniques in imaging data acquisition, segmentation and diagnosis for covid-19." *IEEE reviews in biomedical engineering* (2020).
- [46] Yan, Carol H., et al. "Association of chemosensory dysfunction and Covid-19 in patients presenting with influenza-like symptoms." *International forum of allergy & rhinology*. 2020.
- [47] Sun, Yinxiaohe, et al. "Epidemiological and clinical predictors of COVID-19." *Clinical Infectious Diseases* 71.15 (2020): 786-792.
- [48] Oliveiros, Barbara, et al. "Role of temperature and humidity in the modulation of the doubling time of COVID-19 cases." *medRxiv* (2020).
- [49] Chen, Xiaofeng, et al. "A diagnostic model for coronavirus disease 2019 (COVID-19) based on radiological semantic and clinical features: a multi-center study." *European radiology* (2020): 1.
- [50] Dryhurst, Sarah, et al. "Risk perceptions of COVID-19 around the world." *Journal of Risk Research* (2020): 1-13.

**Dr. Ayoub Khan** (Senior Member, IEEE) received a Ph. D (Electrical Engg.) from Jamia Millia Islamia, New Delhi, India and a Master of Technology (Computer Science and Engineering) from Guru Gobind Singh Indraprastha, New Delhi, India. He worked with many leading organizations like C-DAC (Ministry of IT and Communications), Noida, Sharda University. Presently, he is an Associate Professor at the University of Bisha, Saudi Arabia with interests in the Internet of Things, RFID, wireless sensor networks, ad hoc networks, smart cities, industrial IoT, signal processing, NFC, routing in network-on-chip, real time and embedded systems. He has more than 15 years of experience in his research areas. He has published many research papers and books in reputed journals and international conferences. He contributes to the research community by undertaking various volunteer activities in the capacity of editor for many journals and as a conference chair.






**Dr. Rijwan Khan** has obtained Ph. D in Computer Engg. From Jamia Millia Islamia, New Delhi, India. He is presently work as head of department & professor at ABES Institute of Technology (Affiliated to AKTU), Ghaziabad in the Department of Computer Science & Engineering. He more than 50 publications in his area of research.

**Fahad Algarni**, Dean of the College of Computing and Information Technology at the University of Bisha, Saudi Arabia. He received a PhD from Clayton School of Information Technology, Monash University, Melbourne, Australia. He was awarded a bachelor degree from the Department of Computer Science at King Abdul-Aziz University with First Class Honors. He also graduated with a Master of Information Technology (Computer Networks) from La Trobe University, Melbourne, Australia. His research interests include wireless sensor networks, cloud computing, systems' reliability, IoT, and cybersecurity.

**Dr Indrajeet Kumar** is working as assistant professor with graphic era hill university, Dehradun. His research interest includes IoT, Ad hoc network, data analytics etc.

**Akshika Choudhary** is a scholar at ABES IT Engg. College, Ghaziabad with research interest in IoT, Ad hoc network, data analytics etc.

**Aditi Srivastava** is a scholar at ABES IT Engg. College, Ghaziabad with research interest in data analytics and machine learning.

Dr. Mohammad Ayoub Khan	
Dr Rizwan Khan	
Dr Fahad Algarni	
Indrajeet Kumar	
Akshika Choudhary	

Aditi Srivastava



**Conflicts of Interest:** “The authors declare that they have no conflicts of interest to report regarding the present study.”