COVID-19: EDA and Predictive Analysis

A DISSERTATION

submitted in partial fulfillment of the requirements

for the award of the degree of

Master of Technology

in

INFORMATION TECHNOLOGY

by

Md Afzal Ansari

(MIT2019072)

Ramji Jaiswal

(MIT2019106)

Under the supervision of

Dr. Krishna Pratap Singh

Associate Professor

DEPARTMENT OF INFORMATION TECHNOLOGY

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY

ALLAHABAD – 211015, (UP) INDIA

May, 2020

DECLARATION

This is to certify that the project entitled **"COVID-19: EDA and Predictive Analysis"** Submitted by **Md Afzal Ansari** Roll No. **MIT2019072**, **Ramji Jaiswal** Roll No. **MIT2019106** is a report of bonafide work carried out by us, in partial fulfillment of the requirement of the Degree of Master of Technology in Department of Information Technology, **Indian Institute of Information Technology, Allahabad.**

Md Afzal Ansari
(MIT2019072)

Ramji Jaiswal
(MIT2019106)

CERTIFICATE

This is to certify that the project entitled **"COVID-19: EDA and Predictive Analysis"** Submitted by **Md Afzal Ansari** Roll No. **MIT2019072**, **Ramji Jaiswal** Roll No. **MIT2019106** has been carried out under my/our supervision and that this work has not been submitted elsewhere for a degree*.

Signature of Supervisor

Date: 02-05-2020

Dr. Krishna Pratap Singh

Associate Professor

ACKNOWLEDGEMENT

We would like to express my gratitude to the mercy of Almighty, that given me enough strength to complete this project work. We are very thankful to my supervisor Dr. K P Singh, Associate Professor, Department of Information technology, IIIT-Allahabad to guide me throughout the project work and always motivated me to work hard. The regular discussions and extensive comments helped to improve my project and a direct impact can be seen on the project.

We are thankful to Dr. Vrijendra Singh, Head of Department (IT), IIIT-Allahabad, for providing valuable guidance at beginning of the project. We are also grateful to all the faculty members and supporting staff who had been very supporting in providing resources to complete this project.

We are also thankful to the TA's of our supervisor who assisted us and shared their experience while working on the project. Last but not the least, we wish to thanks all my friends who always been a great support to me and a source of motivation to work forward.

(Md Afzal Ansari)

(Ramji Jaiswal)

ABSTRACT

The project deals with the exploratory data analysis and predictive analysis of the outbreak of global pandemic COVID-19. It's first outbreak reported in Wuhan, China and with the passage of time it has covered more than 50 countries. In concern with the COVID-19, WHO declared it as Public Health Emergency and advised to deal with this pandemic with more concern. The COVID-19 is a infectious disease that is spreading rapidly all over the world and it requires immediate and right steps to be taken in order to deal with such pandemic. In view of the above problems, COVID-19 EDA and prediction analysis is required in order to know the approx. rate of increase of the pandemic. The project consist of exploratory data analysis of the data of countries and provinces/states within the country. This project helps in viewing the expected convergence of the disease and visualizing the growth of the pandemic in next few months. The project aims to predict and visualize the COVID19 cases which comprise of number of recoveries, number of deaths, number of confirmed cases, and total number of active cases that can happen in the next few years. The model helps to understand the patterns of sentiment analysis on COVID cases, and assess influence of the spread of the corona virus.

Contents

List of Figures

Key Terms

- COVID-19
- data analysis
- predictive modelling
- exploratory analysis
- pandemic

Chapter 1

INTRODUCTION

The first case of COVID-19 was reported on 31 December 2019 in Wuhan, China. While outside China the first case reported on 13 january 2020, in Thailand. From that time the outbreak had spread to more than 50 countries. This situation has been declared as public Health Emergency situation by WHO and it is a matter of grave concern to the world. There are more than 30 lacs cases have found of corona virus all over the world and its still increasing. The is an infectious disease that usually not related to a particular place or community or period of time. COVID-19 spreads rapidly due which it is causing huge loss to the humanity all over the world. There is a need to implement immediate actions in order to prevent this disease from spreading all over the world. The COVID-19 is caused by a new type of virus which was earlier indentified as 2019-nCoV by the WHO. The problems caused by the virus includes fever, shortness of breath, cough and nausea. In case of extreme condition it can upto pneumonia and even death if proper ventilators are not provided. It takes 14 days fot the symptoms to appear and it is a latent type of disease. The virus spreads through personal contact or can be through respiratory droplets and interaction.

1.1 Objective

The proposed model aims to process the COVID-19 dataset in order to do the exploratory and predictive analysis. As the COVID-19 disease is spreading at exponential rate, it is necessary to view the extent or the spread of the disease. The

model also predicts the peak of the global pandemic of specific states and country based on input. Then model also tries to find out the expected days of convergence of the disease or the it can be said that the expected number of days after which the pandemic ends.

1.2 Motivation

The motivating factor behind this project is to combat the current situation of COVID-19 through Machine Learning Models. The COVID-19 is highly infectious disease which had covered around more than 50 countries. There is a need to understand the pattern through which the disease is spreading so that necessary steps can be taken in order to deal any critical situation. The project model predicts the COVID-19 spread in the country or the particular provinces of the country. It helps to understand the pattern of spread of the disease.

Chapter 2

LITERATURE REVIEW

Machine learning algorithms has a wide applications in different fields. The ML algorithms also find applications in the medical field that can be related to a medical disease or to find major cause of occurrence of any medical condition. In this project the ML algorithms are implemented in order to predict the occurrence of the infectious disease called as COVID-19. The applicability of the ML model formed can be extended to various other analysis related to this disease.

COVID-19 came with an overabundance of information that is related to the outbreak. But all the information related to the disease is not accurate and can be misleading. Due to which it becomes very hard for the people to understand the disease and to find out the information related to it. As with increase of the outbreak, need for the reliable sources is also increasing so that people can get guidance when they require it. The dataset related to the outbreak authorised by various sources are available on different sites like kaggle. But there is need to understand the dataset which is present. For the common people it is very difficult to understand the complex structure of the data. Hence here the need for a ML EDA analysis for the COVID-19 outbreak originates. This project with extensive and exploratory analysis of the outbreak based on the data present on authentic sites. It helps the people to understand how long will the disease can retain so it eliminates any of fear and panic among the people. This also rise awareness among the people regarding the corona virus and its infectious nature. It opens gateway for the people to understand the pattern through which the disease is spreading across the world and it's impact on the humanity.

In this project we have proposed a comprehensive framework which will represent as a tool for dealing with the epidemics and framing health policy. The project focuses on processing the data-sets in order to drive out meaningful information which is carried forward to form basis of prediction for the pandemic situation.

Chapter 3

METHODOLOGY

The proposed model is framed to work for two major applications:

- **Prediction**

  In this prediction, the model predicts the spread of the COVID-19 in future with specified number of days.

- **Exploratory**

  In this Exploratory data analysis, the model analyses the COVID-19 dataset in order to find the peak of the outbreak and to predict the convergence of the disease.

3.1 SVR Model

Support vector Regression(SVR) works similar to Support Vector Classifier but it is used for regression while SVC is used for classification. SVR uses the algorithm of Support Vector Machine algorithm so as to predict a continuous variable. Unlike other regression models which tries to minimize the error between predicted and actual value, but SVR tries to fit the best line within a predefined or threshold error value. One of the advantage of SVR algorithm that it consist of non-parametric technique. SVR model does not depend on distributions of the given dependent and independent variables. The SVR method depends on the type of kernal used in the model. In order to avoid an over-fit condition the regression can be penalized with a cost parameter. SVR is a useful technique provides the user with high flexibility

in terms of distribution of underlying variables, relationship between independent and dependent variables and the control on the penalty term.

3.1.1 Cost Function

The cost function for the SVR can be given into two forms:

- **Without the cost parameter 'C'** :

$$min \ \frac{1}{2}||w||^2 \tag{3.1}$$

Constraints:

$$y_i - wx_i - b \leq \epsilon \tag{3.2}$$

$$wx_i + b - y_i \leq \epsilon \tag{3.3}$$

- **With the cost parameter 'C'** :

$$min \ \frac{1}{2}||w||^2 \ + \ C\sum_{i=1}^{N}(\epsilon_i \ + \ \epsilon_i^*) \tag{3.4}$$

Constraints:

$$y_i - wx_i - b \leq \epsilon \ + \ \epsilon_i \tag{3.5}$$

$$wx_i + b - y_i \leq \ + \ \epsilon_i \tag{3.6}$$

$$\epsilon_i, \ \epsilon_i^* \geq 0 \tag{3.7}$$

3.1.2 Kernel types

There are various kernals functions can be used in the SVR Algorithm. These can be given as follows:

- **Polynomial Functions**: It is given as follows:

$$k(x_i, x_j) \ = \ (x_i.x_j)^d \tag{3.8}$$

- **Gaussian radial basis Functions**: It is given as follows:

$$k(x_i, x_j) \ = \ \exp(-\frac{||x_i - x_j||^2}{2\sigma^2}) \tag{3.9}$$

3.2 Types of SVR Models

The SVR models tries to manipulate the data in order to fit into a feature space that can able to perform the linear separation. The SVR model can be classified as:

- **Linear SVR**: It is given as follows:

$$y = \sum_{i=1}^{N}(\alpha_i - \alpha_i^*) < x_i, x_j > + b \tag{3.10}$$

- **Non-Linear SVR**: The kernel functions transforms the data into a high dimensional feature space to make it possible to perform the linear separation. It is given as follows:

$$y = \sum_{i=1}^{N}(\alpha_i - \alpha_i^*)K(x_i, x_j) + b \tag{3.11}$$

3.3 Method

The algorithm of the model can be given as follows:

1. Reading of the dataset using pandas library.
2. Pre process the dataset in order to fetch the required data or specific data.
3. For prediction purpose using kernal = poly and degree = 3.
4. After prediction, using the EDA analysis, plotting the graph and finding the peak and convergence of the outbreak

Chapter 4

DATASET

The data-set consist of COVID-19 outbreak in various countries. There are two dataset used in this project which are given below:

1. The first dataset consist of all the COVID-19 outbreak cases with the different states of India.
   The number of columns in the dataset = 20000.
   The number of rows in the dataset = 6

2. The second dataset consist of all the COVID-19 outbreak cases with the different countries wrt date.
   The number of columns in the dataset = 19000.
   The number of rows in the dataset = 5

Chapter 5

PREDICTION

Prediction of spread of COVID-19 disease in terms of confirmed cases for Indian states and Countries are given below:

5.1 Prediction for Delhi

Days Since 02/03/20 For Delhi State of India

	Date	Predicted Number of Confirmed Cases
0	2020-05-01	3484.198994
1	2020-05-02	3664.239377
2	2020-05-03	3850.280559
3	2020-05-04	4042.420915
4	2020-05-05	4240.758818
5	2020-05-06	4445.392641
6	2020-05-07	4656.420759
7	2020-05-08	4873.941546
8	2020-05-09	5098.053375
9	2020-05-10	5328.854620

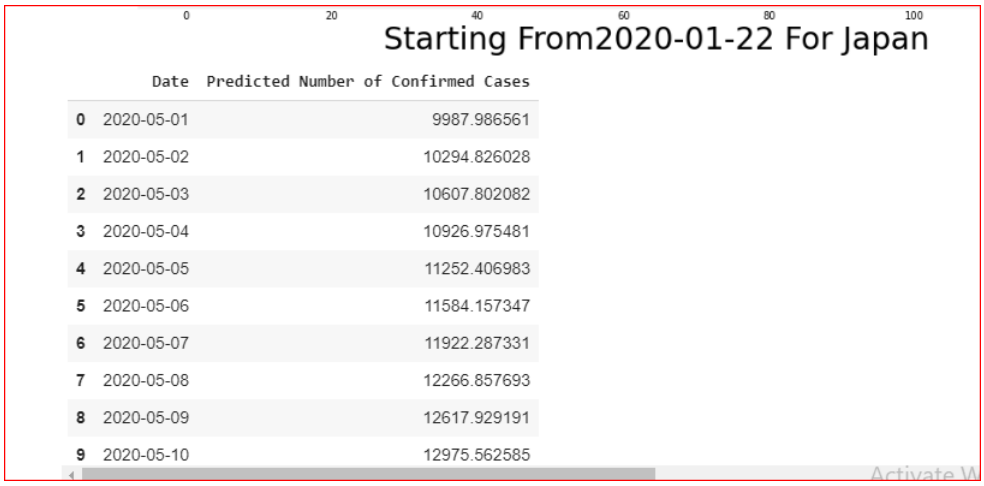
Figure 5.1: prediction for future spread

5.2 Prediction for Kerala

	Date	Predicted Number of Confirmed Cases
0	2020-05-01	582.428111
1	2020-05-02	601.723563
2	2020-05-03	621.438465
3	2020-05-04	641.577328
4	2020-05-05	662.144660
5	2020-05-06	683.144974
6	2020-05-07	704.582778
7	2020-05-08	726.462583
8	2020-05-09	748.788900
9	2020-05-10	771.566238

Figure 5.2: prediction for future spread

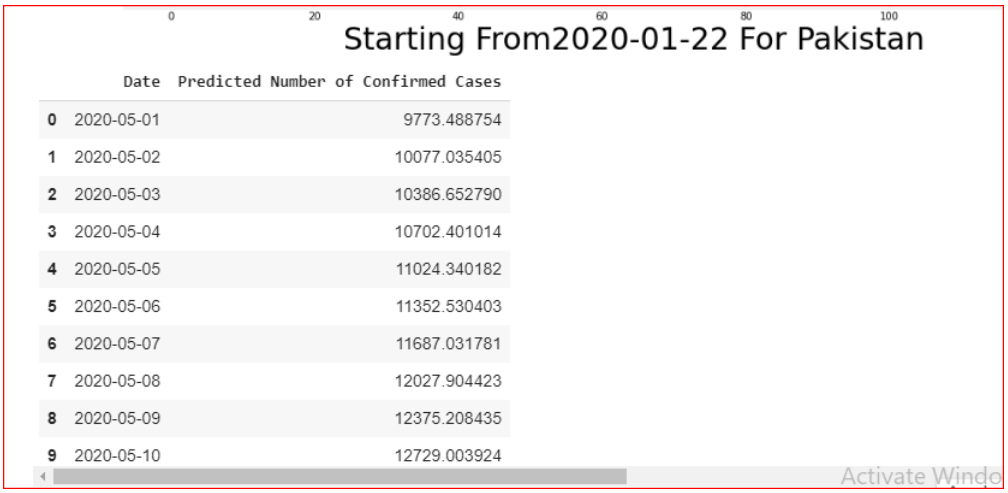
5.3 Prediction for Japan



	Date	Predicted Number of Confirmed Cases
0	2020-05-01	9987.986561
1	2020-05-02	10294.826028
2	2020-05-03	10607.802082
3	2020-05-04	10926.975481
4	2020-05-05	11252.406983
5	2020-05-06	11584.157347
6	2020-05-07	11922.287331
7	2020-05-08	12266.857693
8	2020-05-09	12617.929191
9	2020-05-10	12975.562585

Figure 5.3: prediction for future spread

5.4 Prediction for Pakistan



	Date	Predicted Number of Confirmed Cases
0	2020-05-01	9773.488754
1	2020-05-02	10077.035405
2	2020-05-03	10386.652790
3	2020-05-04	10702.401014
4	2020-05-05	11024.340182
5	2020-05-06	11352.530403
6	2020-05-07	11687.031781
7	2020-05-08	12027.904423
8	2020-05-09	12375.208435
9	2020-05-10	12729.003924

Figure 5.4: prediction for future spread

Chapter 6

EXPLORATORY DATA ANALYSIS

The analysis of the COVID-19 data-set is done on basis of country-wise and state-wise data.

6.1 India COVID dataset

- **Data representing COVID-19 cases in India**

	Sno	Date	Time	State/UnionTerritory	ConfirmedIndianNational	ConfirmedForeignNational	Cured	Deaths	Confirmed
0	1	30/01/20	6:00 PM	Kerala	1	0	0	0	1
1	2	31/01/20	6:00 PM	Kerala	1	0	0	0	1
2	3	01/02/20	6:00 PM	Kerala	2	0	0	0	2
3	4	02/02/20	6:00 PM	Kerala	3	0	0	0	3
4	5	03/02/20	6:00 PM	Kerala	3	0	0	0	3
...	...	...	...	...	...	...	...	...	...
1473	1474	30/04/20	5:00 PM	Telengana	-	-	367	26	1012
1474	1475	30/04/20	5:00 PM	Tripura	-	-	2	0	2
1475	1476	30/04/20	5:00 PM	Uttarakhand	-	-	36	0	55
1476	1477	30/04/20	5:00 PM	Uttar Pradesh	-	-	513	39	2203
1477	1478	30/04/20	5:00 PM	West Bengal	-	-	124	22	758

1478 rows × 9 columns

Figure 6.1: state-wise data

- **Comparison of confirmed cases for multiple states in India**

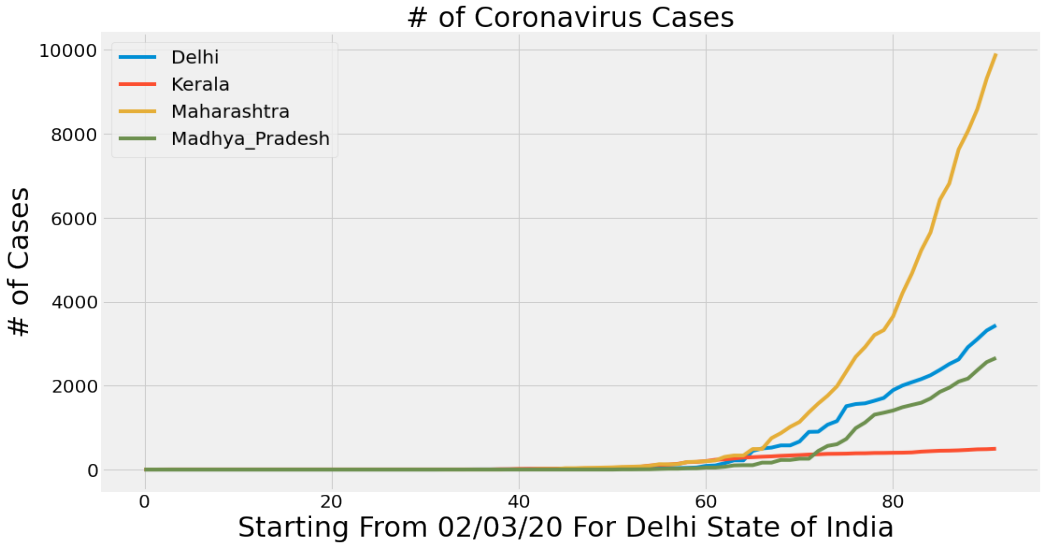


Figure 6.2: state-wise data comparision

6.2 Analysis on Delhi's COVID-19 dataset

6.2.1 Data-set of Delhi

- **As first case found on 02/03/2020**

[25] (60, 6)

	Date	Time	State/UnionTerritory	Cured	Deaths	Confirmed
34	02/03/20	6:00 PM	Delhi	0	0	1
38	03/03/20	6:00 PM	Delhi	0	0	1
42	04/03/20	6:00 PM	Delhi	0	0	1
45	05/03/20	6:00 PM	Delhi	0	0	2
51	06/03/20	6:00 PM	Delhi	0	0	3
62	07/03/20	6:00 PM	Delhi	0	0	3
71	08/03/20	6:00 PM	Delhi	0	0	3
84	09/03/20	6:00 PM	Delhi	0	0	4
87	10/03/20	6:00 PM	Delhi	0	0	4
98	11/03/20	6:00 PM	Delhi	0	0	5
109	12/03/20	6:00 PM	Delhi	0	0	6
122	13/03/20	6:00 PM	Delhi	0	0	6
135	14/03/20	6:00 PM	Delhi	1	1	7

Figure 6.3: Delhi state data

6.2.2 Analysis of Increase rate of Delhi cases

- **Analysis of Increase rate of Confirmed cases in Delhi**

  We found maximum Increase at Day 42

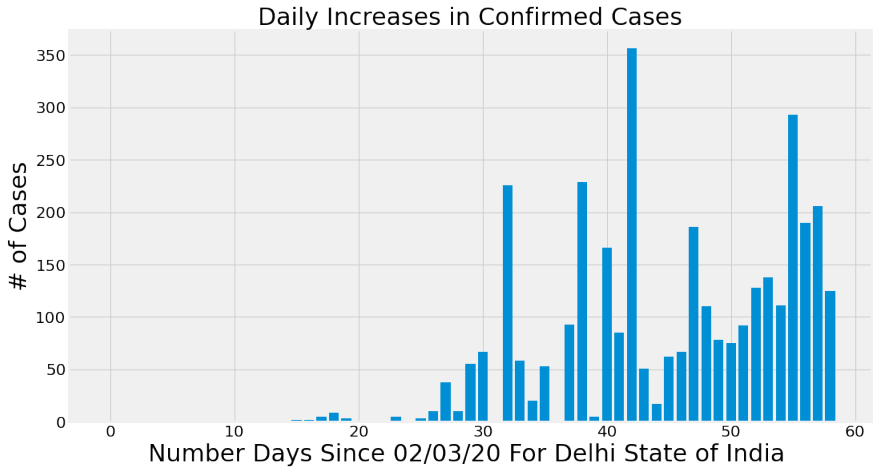


Figure 6.4: Bargraph of increasing confirmed cases

- **Analysis of Increase rate of cured cases in Delhi**

  We found maximum Increase at Day 48

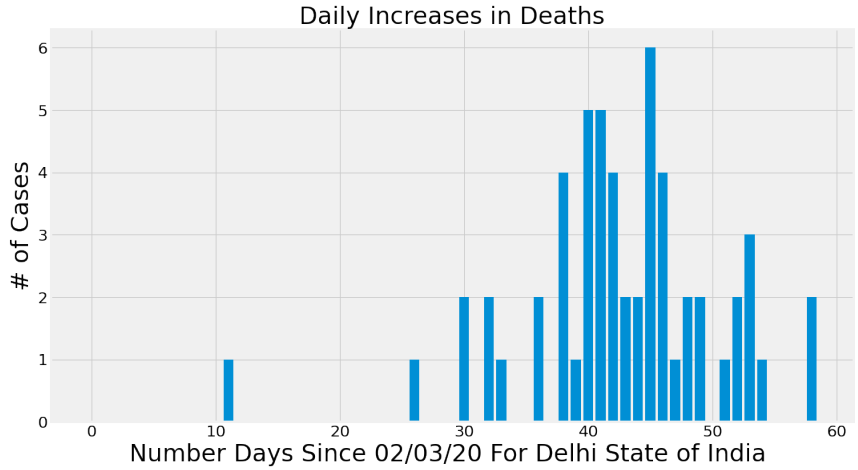


Figure 6.5: Bargraph of increasing cured cases

- **plot Increase rate of cured cases in Delhi**

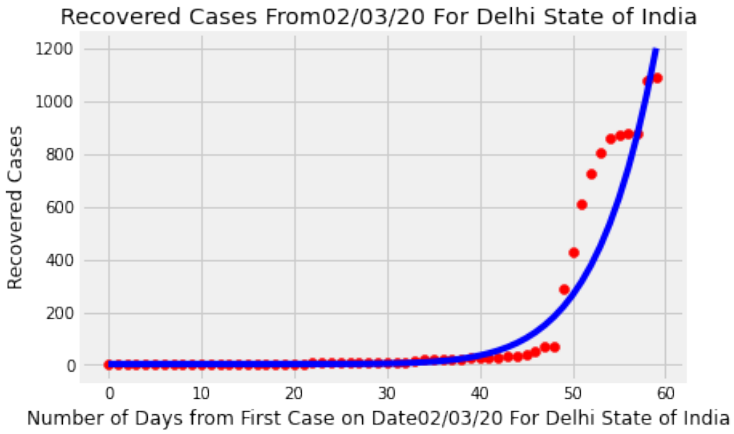


Figure 6.6: plot of increasing cured cases

- **plot Increase rate of death cases in Delhi**

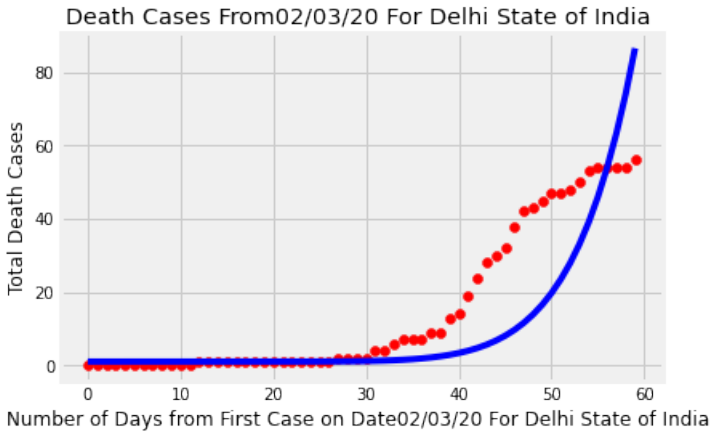


Figure 6.7: plot of increasing death cases

6.2.3 Convergence Analysis

- We found maximum Active cases at day

  Day = 58, Active Cases = 2194.6613328962985

  Day = 70, Active Cases = 0

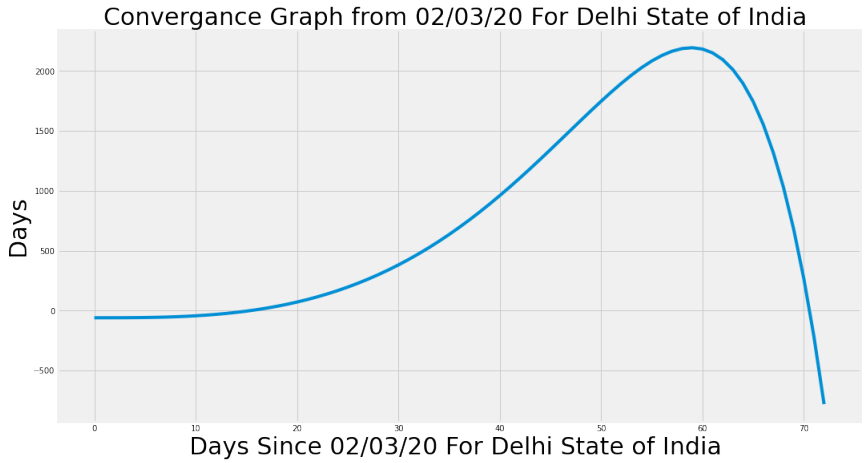


Figure 6.8: convergence analysis of delhi

6.3 Analysis on Kerala's COVID-19 dataset

6.3.1 Data-set of Kerala

- **The data-set is shown as:**

	Date	Time	State/UnionTerritory	Cured	Deaths	Confirmed
0	30/01/20	6:00 PM	Kerala	0	0	1
1	31/01/20	6:00 PM	Kerala	0	0	1
2	01/02/20	6:00 PM	Kerala	0	0	2
3	02/02/20	6:00 PM	Kerala	0	0	3
4	03/02/20	6:00 PM	Kerala	0	0	3
...	...	...	...	...	...	...
1333	26/04/20	5:00 PM	Kerala	338	4	458
1365	27/04/20	5:00 PM	Kerala	342	4	469
1397	28/04/20	5:00 PM	Kerala	355	4	482
1429	29/04/20	5:00 PM	Kerala	359	4	486
1461	30/04/20	5:00 PM	Kerala	369	4	496

92 rows × 6 columns

Figure 6.9: Kerala state data

6.3.2 Analysis of Increase rate of Kerala cases

- **Analysis of Increase rate of Confirmed cases in Kerala**

  We found maximum Increase at Day 56

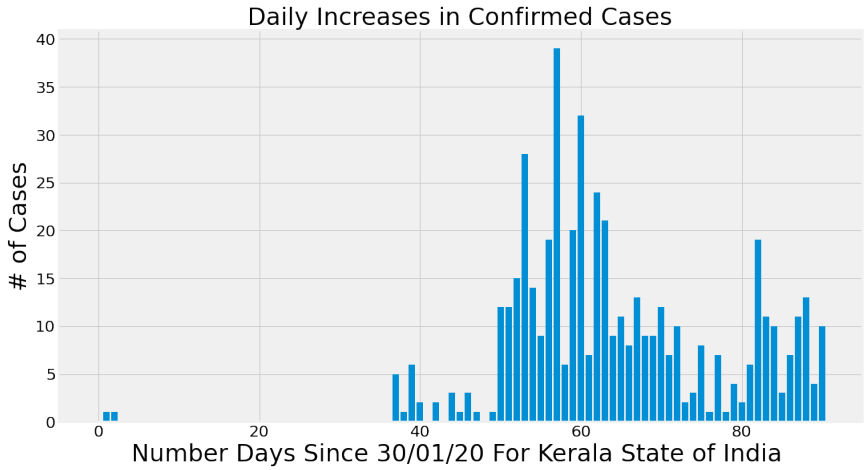


Figure 6.10: Bargraph of increasing confirmed cases

- **Analysis of Increase rate of cured cases in kerala**

  We found maximum Increase at Day 74

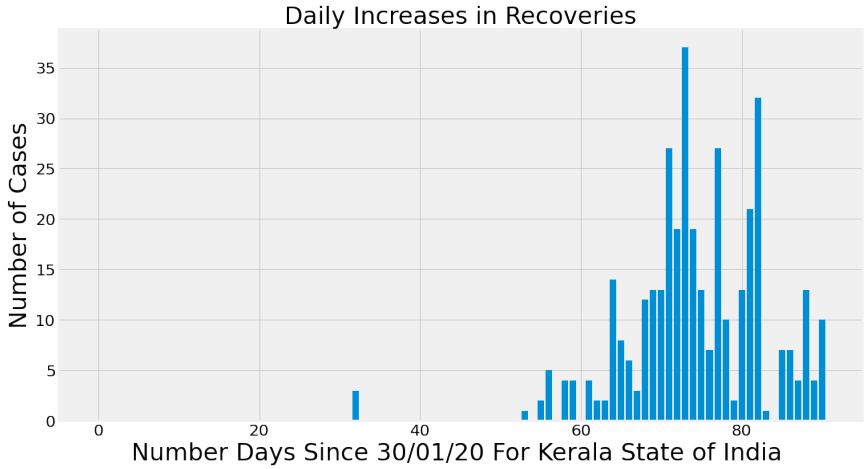


Figure 6.11: Bargraph of increasing cured cases

- **Analysis of Increase rate of death cases in kerala**

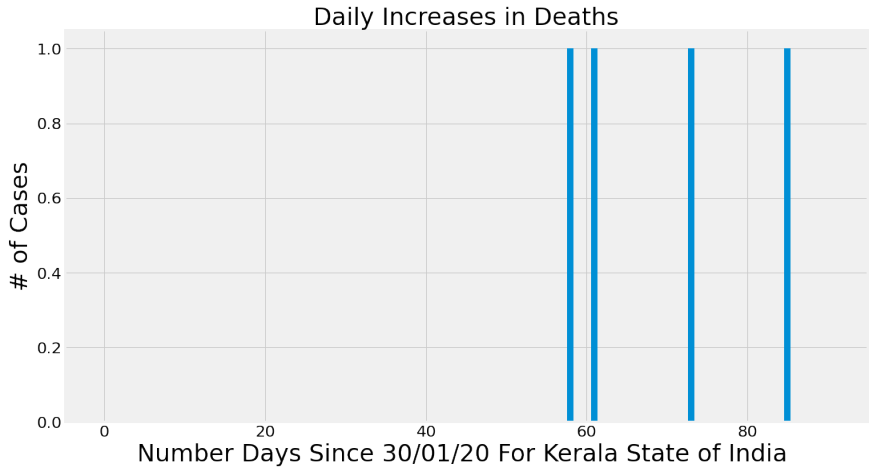


Figure 6.12: Bargraph of increasing death cases

- **plot Increase rate of cured cases in Kerala**

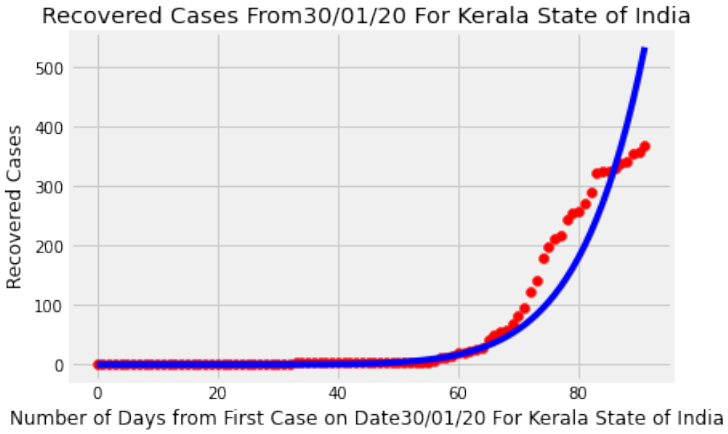


Figure 6.13: plot of increasing cured cases

- **plot Increase rate of death cases in Kerala**

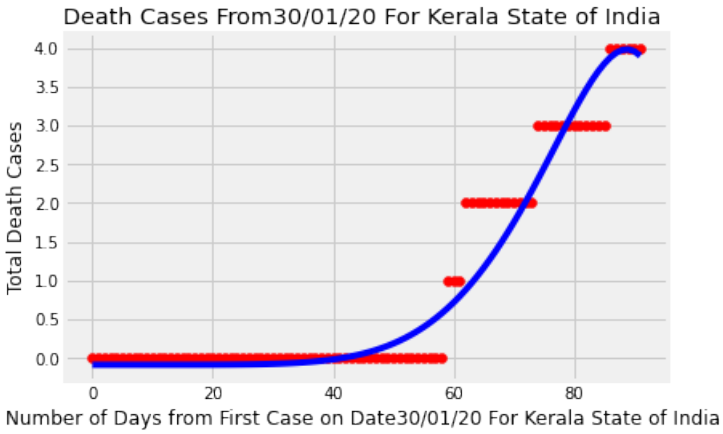


Figure 6.14: plot of increasing death cases

- **plot Increase rate of confirmed cases in Kerala**

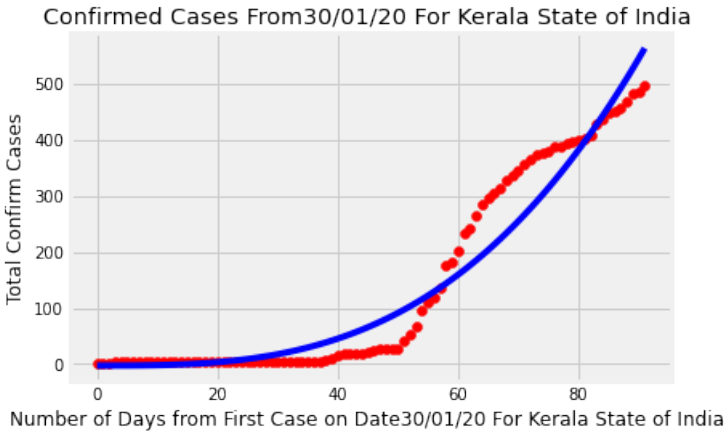


Figure 6.15: plot of increasing confirmed cases

6.3.3 Convergence Analysis

- We found Convergence at day = 94 from starting date.

  We found highest active cases = 212.31452477218295 at day 74

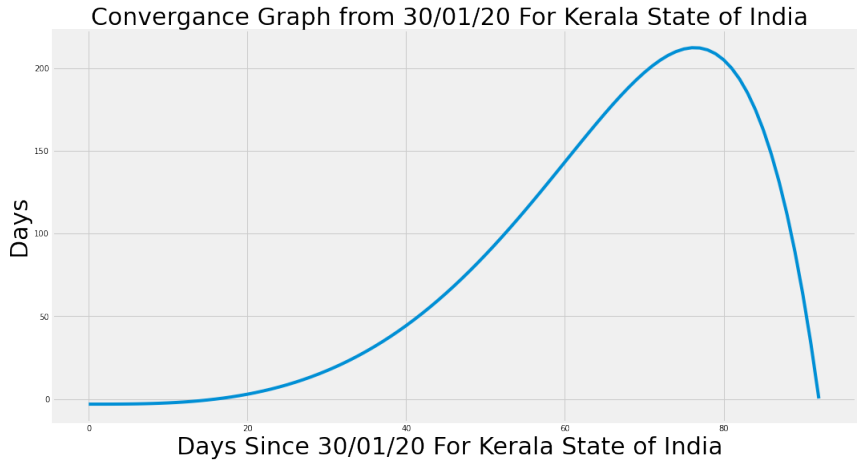


Figure 6.16: convergence analysis of Kerala

6.4 World COVID dataset

- **Data representing COVID-19 cases throughout the world**

	Date	Country	Confirmed	Recovered	Deaths
0	2020-01-22	Afghanistan	0	0	0
1	2020-01-22	Albania	0	0	0
2	2020-01-22	Algeria	0	0	0
3	2020-01-22	Andorra	0	0	0
4	2020-01-22	Angola	0	0	0
5	2020-01-22	Antigua and Barbuda	0	0	0
6	2020-01-22	Argentina	0	0	0
7	2020-01-22	Armenia	0	0	0
8	2020-01-22	Australia	0	0	0
9	2020-01-22	Austria	0	0	0

Figure 6.17: country-wise data

- **Comparison of confirmed cases for multiple countries**

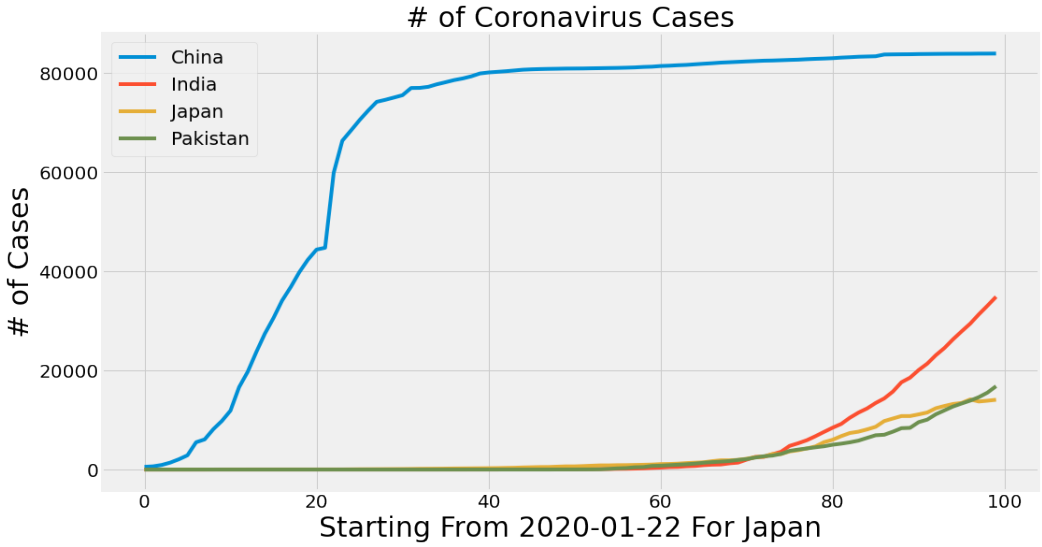


Figure 6.18: Country-wise data comparision

6.5 Analysis on Japan's COVID-19 dataset

6.5.1 Data-set of Japan

- **As first case found on 22/01/2020**

	Date	Country	Confirmed	Recovered	Deaths
87	2020-01-22	Japan	2	0	0
274	2020-01-23	Japan	2	0	0
461	2020-01-24	Japan	2	0	0
648	2020-01-25	Japan	2	0	0
835	2020-01-26	Japan	4	1	0
1022	2020-01-27	Japan	4	1	0
1209	2020-01-28	Japan	7	1	0
1396	2020-01-29	Japan	7	1	0
1583	2020-01-30	Japan	11	1	0
1770	2020-01-31	Japan	15	1	0

Figure 6.19: Delhi state data

6.5.2 Analysis of Increase rate of Japan cases

- **Analysis of Increase rate of Confirmed cases in Delhi**

  We found maximum Increase at Day 98

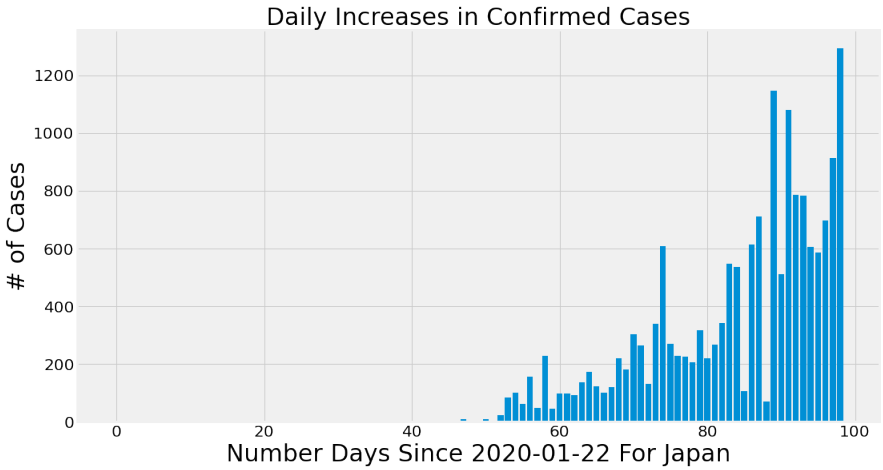


Figure 6.20: Bargraph of increasing confirmed cases

- **Analysis of Increase rate of cured cases in Japan**

  We found maximum Increase at Day 98

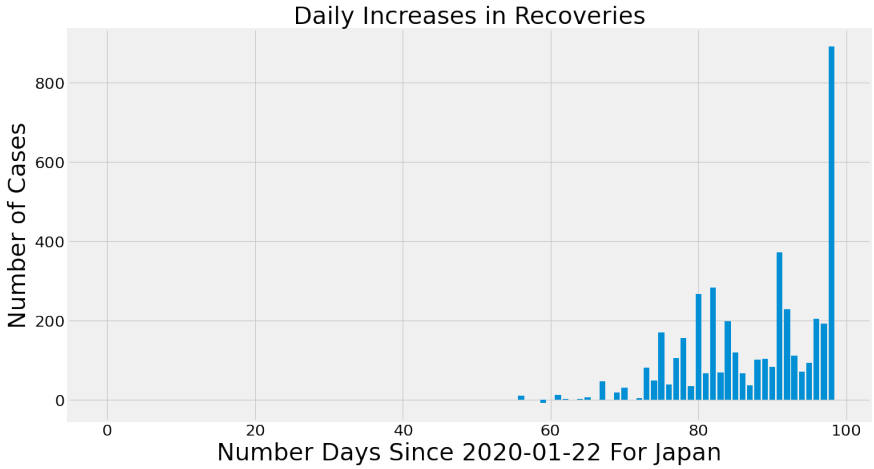


Figure 6.21: Bargraph of increasing cured cases

- **Analysis of Increase rate of Death cases in Japan**

  We found maximum Increase at Day 98

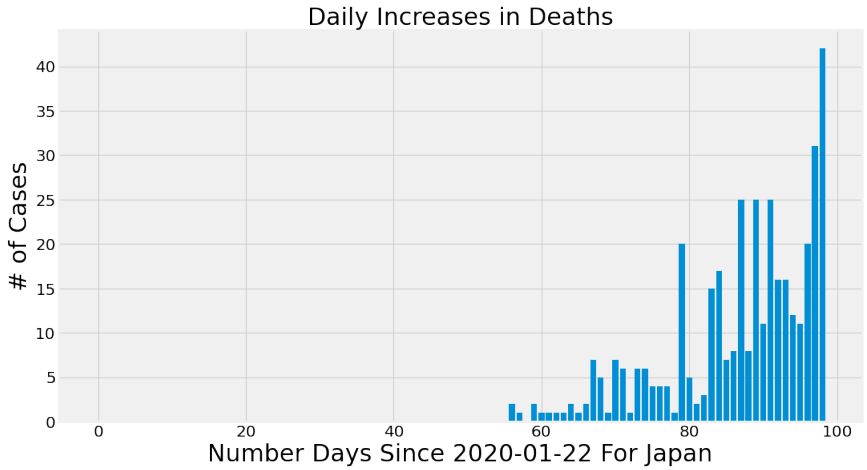


Figure 6.22: Bargraph of increasing death cases

- **plot Increase rate of confirmed cases in Japan**

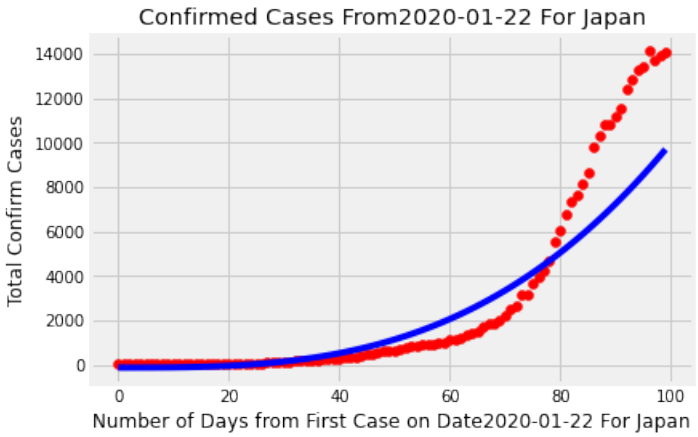


Figure 6.23: plot of increasing cured cases

- **plot Increase rate of cured cases in Japan**

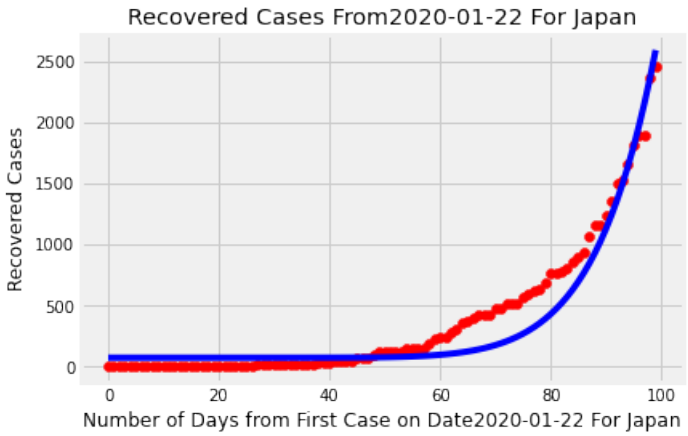


Figure 6.24: plot of increasing cured cases

- **plot Increase rate of death cases in Japan**

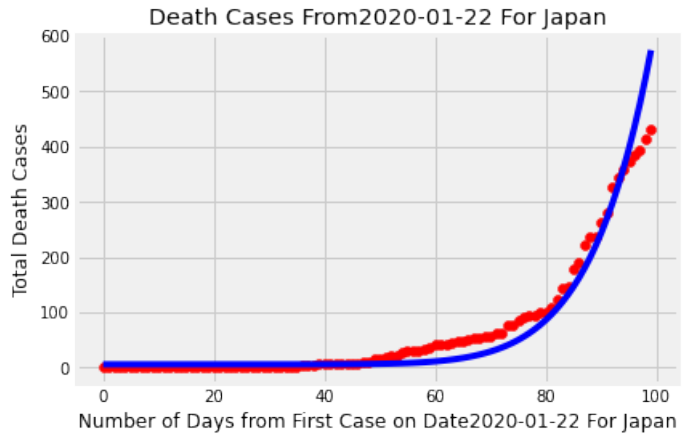


Figure 6.25: plot of increasing death cases

6.5.3 Convergence Analysis

- We found Maximum cases = 8245.250176951544 at day 109

  Cases = 0 at day = 128

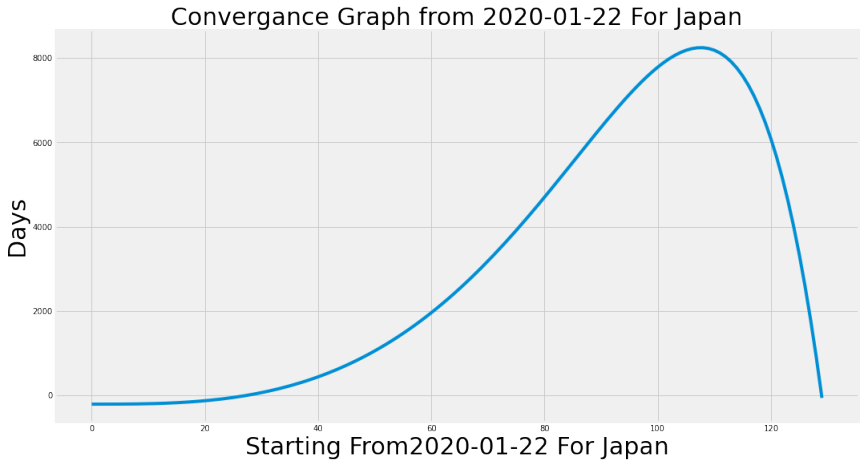


Figure 6.26: convergence analysis of Japan

6.6 Analysis on Pakistan's COVID-19 dataset

6.6.1 Data-set of Pakistan

- **The data-set is shown as:**

	Date	Country	Confirmed	Recovered	Deaths
16959	2020-04-21	Pakistan	9565	2073	201
17146	2020-04-22	Pakistan	10076	2156	212
17333	2020-04-23	Pakistan	11155	2527	237
17520	2020-04-24	Pakistan	11940	2755	253
17707	2020-04-25	Pakistan	12723	2866	269
17894	2020-04-26	Pakistan	13328	2936	281
18081	2020-04-27	Pakistan	13915	3029	292
18268	2020-04-28	Pakistan	14612	3233	312
18455	2020-04-29	Pakistan	15525	3425	343
18642	2020-04-30	Pakistan	16817	4315	385

Figure 6.27: Pakistan state data

6.6.2 Analysis of Increase rate of Pakistan cases

- **Analysis of Increase rate of Confirmed cases in Pakistan**

  We found maximum Increase at Day 98

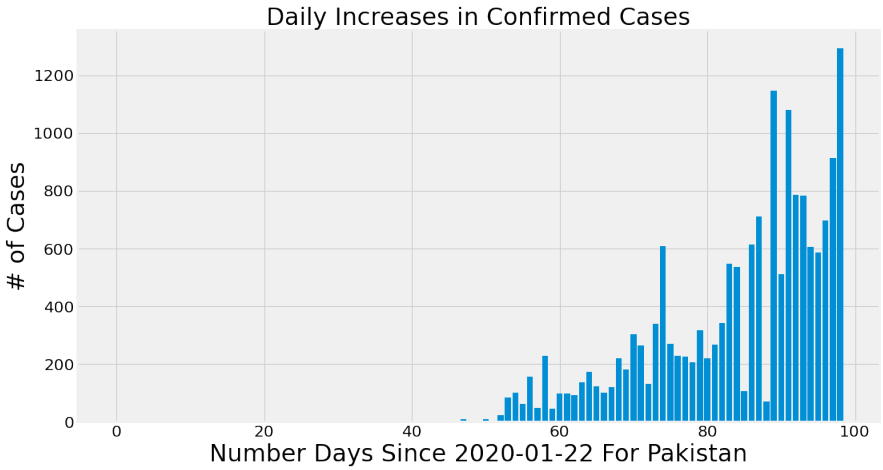


Figure 6.28: Bargraph of increasing confirmed cases

- **Analysis of Increase rate of cured cases in Pakistan**

  We found maximum Increase at Day 98

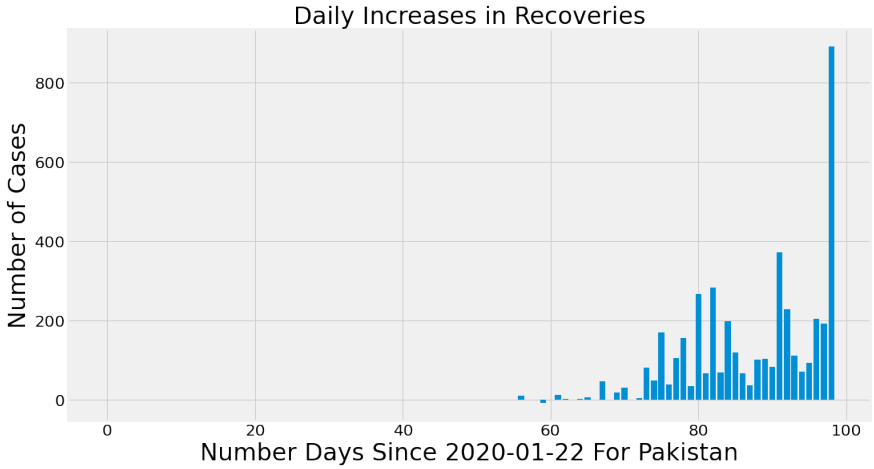


Figure 6.29: Bargraph of increasing cured cases

- **Analysis of Increase rate of death cases in Pakistan**

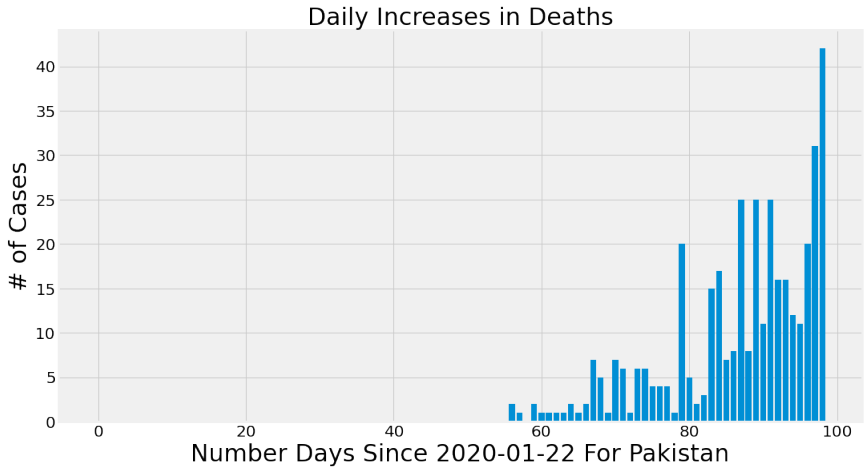


Figure 6.30: Bargraph of increasing death cases

- **plot Increase rate of cured cases in Pakistan**

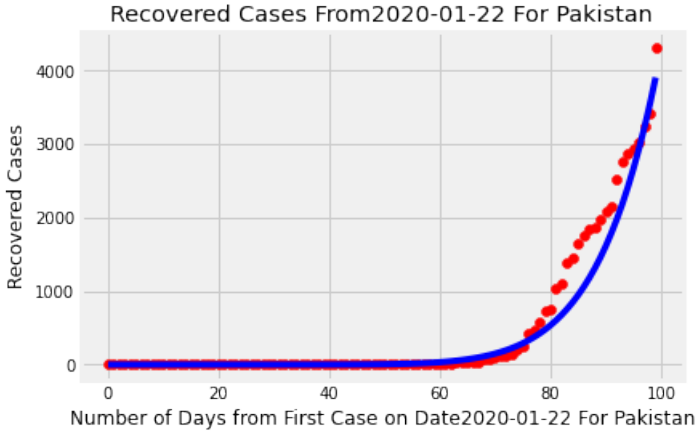


Figure 6.31: plot of increasing cured cases

- **plot Increase rate of death cases in Pakistan**

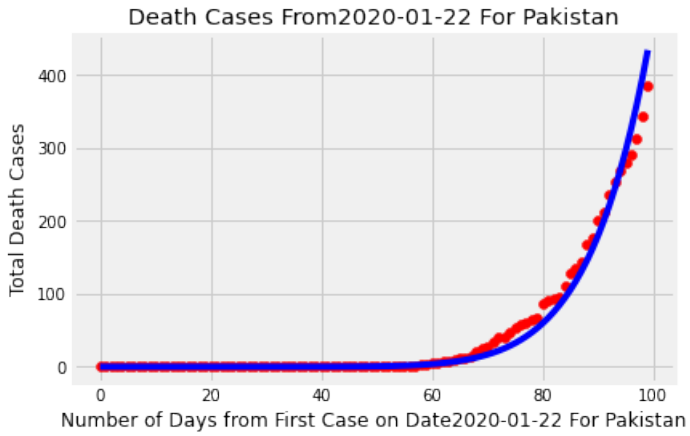


Figure 6.32: plot of increasing death cases

- **plot Increase rate of confirmed cases in Pakistan**

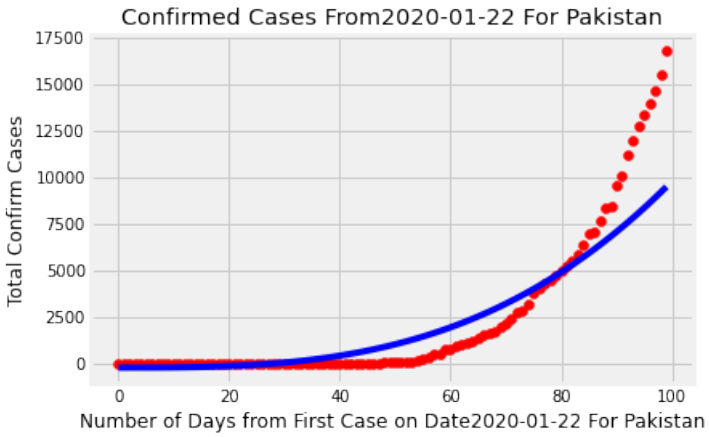


Figure 6.33: plot of increasing confirmed cases

6.6.3 Convergence Analysis

- We found Maximum Active Cases = 6025.460875940959 at day = 99

  We found Minimum Active Cases = 0 at day = 99 at day 116

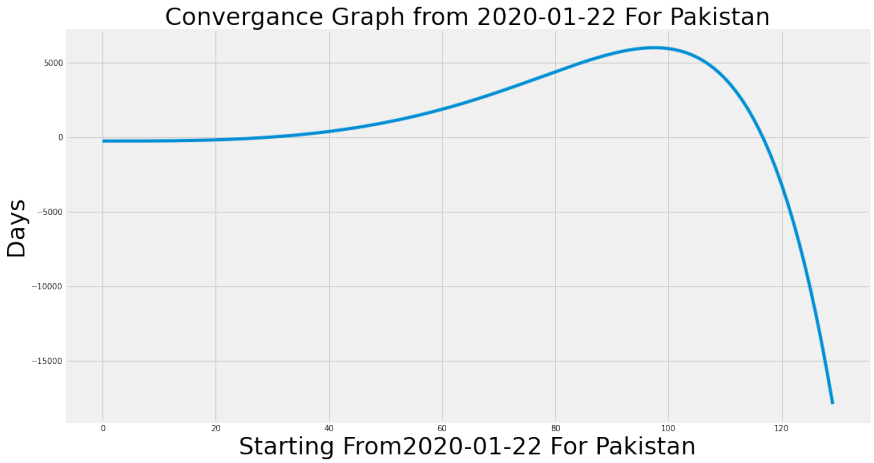


Figure 6.34: convergence analysis of Pakistan

Chapter 7

CONCLUSIONS

The proposed model designed to work in such a way it can able to predict and do future forecasting of the outbreak. The result obtained for the prediction part gives the prediction of spread of the COVID-19 depends on current scenario. The conclusion of the project that it can be able to accurately predict the future outbreak conditions. It can be used in different medical fields. In can also be used to estimate the economic crisis and stability.

Bibliography

[1] Kaggle dataset: https://www.kaggle.com/tags/covid19

[2] kaggle Indian dataset : https://www.kaggle.com/sudalairajkumar/covid19-in-india

[3] COVID-19 information : https://www.mygov.in/covid-19