Business Intelligence Project

Novel Corona Virus 2019

Submitted by

TEAM 1

Aarti Uniyal - 05204092018

Sreyashi Bhattacharjee - 05104092018

Chahat - 03904092018

Aakriti Mittal- 05804092018

Deepika - 05504092018

Under the Supervision of

Rishabh Kaushal

Assistant Professor

Department of Information Technology

Department of Information Technology

Indira Gandhi Delhi Technical University for Women

Kashmere Gate, Delhi - 110006

March 2020

Contents

1 Introduction **1**

2 Literature Survey **2**

3 Problem Statement **3**

3.1 Research questions 3

4 Dataset **5**

4.1 Data Description: 5

4.1.1 Number of data instances and type(s) of instances 5

4.1.2 Attribute count, type & one-line explanation of each attribute . . . 5

5 Data Preprocessing **7**

5.1 Data File Combining 8

5.2 Data Cleaning 8

6 Exploratory Data Analysis **9**

7 Data Visualization **16**

7.1 Bar Chart 16

7.2 Line Chart 17

7.3 Heatmap 18

7.4 Density Mapbox 19

8 Proposed Approach **20**

8.1 Disease Prediction 20

8.1.1 Prophet Algorithm 21

8.1.2 Linear Regression Algorithm 28

8.1.3 Support Vector Machine Model Regressor Algorithm 32

9 Model Evaluation **36**

9.1 Description 36

9.2 Evaluation on Prophet Algorithm 36

9.2.1 Confirmed Cases metrics 37

9.2.2 Death Cases metrics 38

9.2.3 Recovered Cases metrics 39

10 Data Clustering **40**

11 Result **42**

Chapter 1

Introduction

Novel Corona Virus 2019

Coronavirus disease (COVID-19) is a disease caused by a newly discovered coronavirus . It is highly infectious and can be easily spread through human touch or exchange of bodily fluid.
Most people infected with this virus will experience mild to moderate respiratory illness and no special medical treatment is required for recovery .

The COVID-19 virus spreads primarily through saliva or discharge from the nose when an infected person coughs or sneezes .
At this time, there are no specific vaccines or treatments is available for this virus.

Due to the lack of information governments and citizens around the world do not know how to prepare for this epidemic leading to stocking up of essential goods , medical supplies etc.

Chapter 2

Literature Survey

COVID-19 - Analysis , Viz, Prediction Comparisons : visualisation on different aspects
https://www.kaggle.com/imdevskp/covid%-19-analysis-viz-prediction-comparisons
Coronavirus analysis and predictions : regression analysis and visualisation
https://www.kaggle.com/andyyh/coronavirus-analysis-and-predictions Coronavirus growth in 2019/2020 - Updated 3-9-20 : visualisation , data cleaning and EDA
https://www.kaggle.com/gatunnopvp/coronavirus-growth-in-2019-2020-updated-3-9-20
CONVID19 Novel Coronavirus: EDA Forecasting Cases : used Prophet for forecasting time series data
https://www.kaggle.com/khoongweihao/covid19-novel-coronavirus-eda-forecasting-cases

Chapter 3

Problem Statement

As the deadly COVID-19 Virus spreads, it raises fear of a worldwide pandemic, researchers and corporates are using artificial intelligence and other technologies to predict where the virus might appear next — and even potentially sound the alarm before other new, potentially threatening viruses become public health crises.

Business applications related to the study are:

1. The goal of this task is to build a model that predicts the progression of this virus worldwide.
2. Due to the lack of information governments and citizens around the world do not know how to prepare for this epidemic leading to stocking up of essential goods , medical supplies etc.
3. A model that predicts how the virus could spread across different countries and regions may be able to help mitigation efforts (travel bans , supply chains etc.)

3.1 Research questions

We intend to find solution to the following research questions.

1. What is the number of increase / decrease in cases worldwide in the month of April/May, 2020?
2. How should the government make the decisions about the travel restrictions ?
3. How should the Government manage the supply chains of different commodities?
4. Can we slow the rate of new cases?

5. Will the virus slow down in warm weather?

6. Should the public panic and hoard commodities or will this last for a short period of time ?

7. The origin of coronovirus is wuhan , then why the Italy is the most affected country despite of being the second healthiest country?

8. For how long this disease will continue exist ,what is the deadbreak poin ?

9. Which country will see the decrease in cases 1st?

Chapter 4

Dataset

4.1 Data Description:

Source 1: covid_19_data.csv

Source 2: time_series_covid_19_confirmed.csv

Data collection started from 22nd January , 2020

4.1.1 Number of data instances and type(s) of instances

A dataset for all the countries around the world

Number of instances for first dataset: 8509 (as on 24th March , 2020)

Number of instances for second dataset: 503 (as on 24th March , 2020)

Label Information- No Label Present

4.1.2 Attribute count, type & one-line explanation of each attribute

Attribute count for first:8

Attributes:

- SNo : Serial Number - (numeric)
- Observation Date : Observation date in mm/dd/yyyy - (numeric)
- Province/State : Province or State - (categorical)
- Country/Region : Country or region - (categorical)
- Last Update : Last update date time in UTC - (numeric)

- Confirmed : Cumulative number of confirmed cases - (numeric)
- Deaths : Cumulative number of deaths cases - (numeric)
- Recovered : Cumulative number of recovered cases - (numeric)

Attribute count for second:4

Attributes:

- Province/State : Province or State - (categorical)
- Country/Region : Country or region - (categorical)
- Lat : Latitude of the country - (numeric)
- Long : Longitude of the country - (numeric)

Chapter 5

Data Preprocessing

For First Dataset:

The dataset was available in CSV format with ";" as its separator. The columns had the following datatypes:

Col1	Col2
Date	object
Province/State	object
Country	object
Last Update	datetime64
Confirmed	float64
Deaths	float64
Recovered	float64

For Second Dataset:

The dataset was available in CSV format with ";" as its separator. The columns had the following datatypes:

Col1	Col2
Province/State	object
Country	object
Lat	float64
Long	float64

5.1 Data File Combining

we are combining the data from source 1 and source 2 for visulaization of density mapbox all over the world .
ie we are using atributes like Lat , Long from source 2 and then merge it with source 1 dataset.

5.2 Data Cleaning

Parsing according to the last update datetime column
Using serial number as index
Renaming certain columns like 'ObservationDate':'Date', 'Country/Region':'Country' without making any additional copy of it we are the changing the names in the main dataset

Chapter 6

Exploratory Data Analysis

Information about the various data are mentioned below:

```
Rows :  8509

Columns :  7

Features :  ['Date', 'Province/State', 'Country', 'Last Update', 'Confirmed', 'Deaths', 'Recovered']

Missing Values :  3748

Unique values :
 Date              63
Province/State    289
Country           205
Last Update      1806
Confirmed        1061
Deaths            205
Recovered         648
dtype: int64
```

Figure 6.1: Information about the data

Basic Statistics Metrics about the various data are mentioned below:

Basic Statistics :

	Confirmed	Deaths	Recovered
count	8509.000000	8509.000000	8509.000000
mean	704.421201	25.542955	245.788342
std	5111.664699	252.402842	2774.093868
min	0.000000	0.000000	0.000000
25%	2.000000	0.000000	0.000000
50%	18.000000	0.000000	0.000000
75%	140.000000	1.000000	10.000000
max	69176.000000	6820.000000	60324.000000

Figure 6.2: Basic Statistics Metrics Confirmed, Deaths , Recovered

Earliest Cases of various attributes are mentioned below:

Earliest Cases :

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

Figure 6.3: Earliest Cases.

Latest Cases :

```
Latest Cases :
            Date Province/State        Country        Last Update  \
SNo
8505  03/24/2020  Wuhan Evacuee             US 2020-03-24 23:41:50
8506  03/24/2020        Wyoming             US 2020-03-24 23:41:50
8507  03/24/2020       Xinjiang  Mainland China 2020-03-24 23:41:50
8508  03/24/2020         Yunnan  Mainland China 2020-03-24 23:41:50
8509  03/24/2020       Zhejiang  Mainland China 2020-03-24 23:41:50

      Confirmed  Deaths  Recovered
SNo
8505        4.0     0.0        0.0
8506       29.0     0.0        0.0
8507       76.0     3.0       73.0
8508      176.0     2.0      172.0
8509     1240.0     1.0     1221.0
```

Figure 6.4: Latest Cases .

Maximum number of Confirmed,Deaths and Recovered Cases are mentioned below:

Maximum number of Confirmed,Deaths and Recovered Cases

	Date	Confirmed	Deaths	Recovered
0	03/24/2020	417966	18615	107705

Figure 6.5: Maximum number of cases in China followed by Italy and Iran

World View - Country wise are mentioned below:

World View - Country wise

	Country	Confirmed	Deaths	Recovered
0	Mainland China	3.60653e+06	122579	1.85828e+06
1	Italy	571924	46856	59688
2	Iran	278992	16870	88991
3	Spain	232288	12448	18492
4	US	223838	3127	1136
5	Germany	198142	675	5258
6	South Korea	190800	1670	21645
7	France	144802	4825	8050
8	Switzerland	58069	635	504

Figure 6.6: Maximum number of cases in China followed by Italy and Iran

Country Wise - Sorted(Alphabetically) order are mentioned below:

Country Wise - Sorted(Alphabetically) order

	Country	Confirmed	Deaths	Recovered
0	Azerbaijan	1.0	0.0	0.0
1	('St. Martin',)	2.0	0.0	0.0
2	Afghanistan	363.0	3.0	9.0
3	Albania	851.0	26.0	16.0
4	Algeria	1485.0	113.0	326.0
...	...	...	...	...
200	Venezuela	483.0	0.0	45.0
201	Vietnam	1696.0	0.0	602.0
202	Zambia	17.0	0.0	0.0
203	Zimbabwe	13.0	2.0	0.0
204	occupied Palestinian territory	25.0	0.0	0.0

205 rows × 4 columns

Figure 6.7: Country , Confirmed , Deaths , Recovered

Chapter 7

Data Visualization

7.1 Bar Chart

It is a chart that is used to represent categorical data with rectangular bars. The bars can be plotted either vertically or horizontally.
Therefore, the following chart show the comparisons among discrete categories.

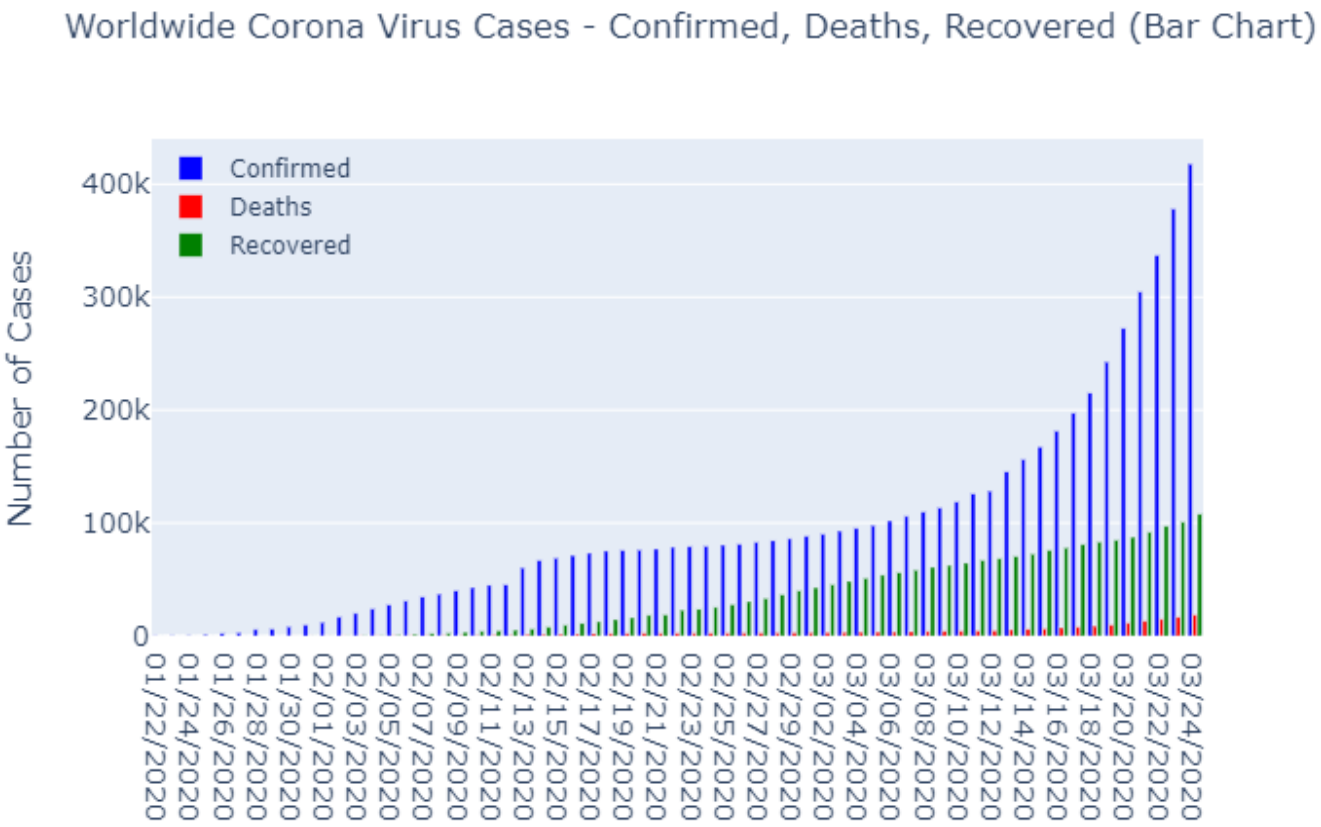


Figure 7.1: Worldwide Corona Virus Cases - Confirmed, Deaths, Recovered

7.2 Line Chart

Line Chart is used to connects a series of data points with a continuous line.

A graphical representation of Top 5 Countries with Corona Virus Cases.

This line chart is a way of visually representing an no of cases to the 5 countries Iran ,Italy ,China ,Spain, US.

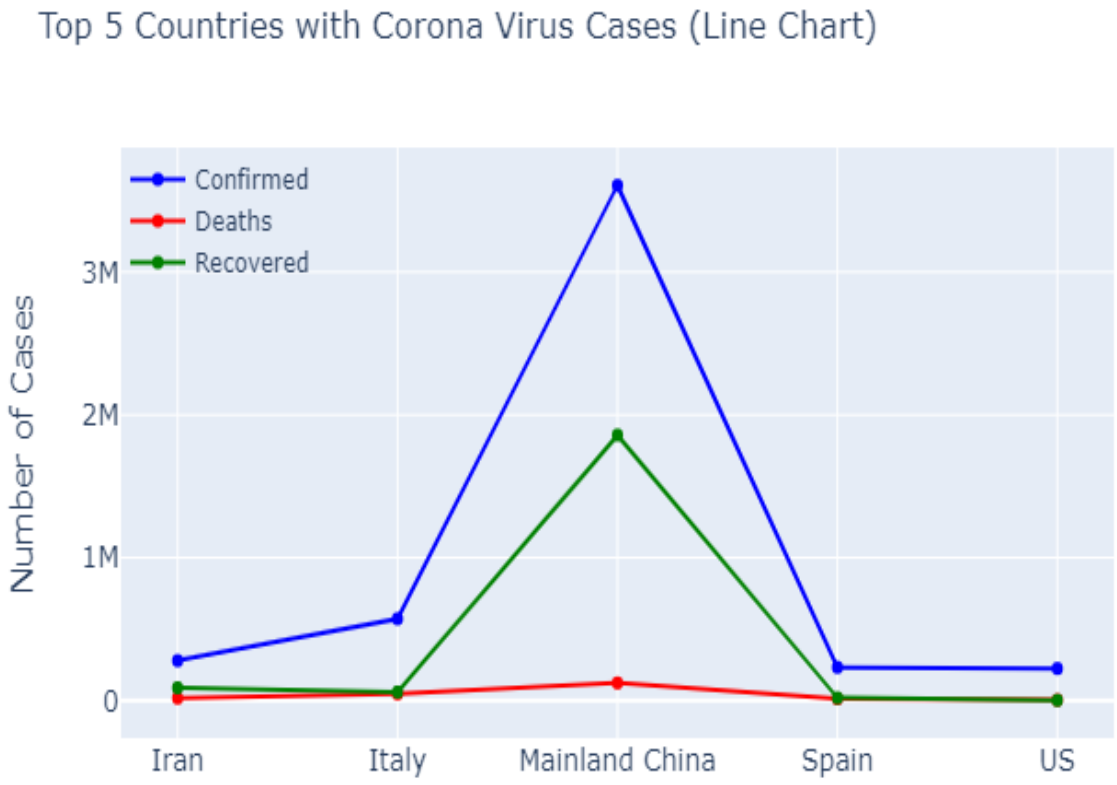


Figure 7.2: Top 5 Countries with Corona Virus Cases Line Chat

7.3 Heatmap

China number of cases per day

A heat map is a representation of a data matrix where values are represented as colors. Heatmaps can be used to display large amounts of points in a way that is visually engaging and encourages your audience to explore your map.

It showing No of people's cases per Date .It visualize hot spots within data sets, and to show patterns or correlations. Due to their compact nature, they are often used with large sets of data. as there are more dense no of cases which increased day by day

cases are still increasing . No notion of decrease till now.

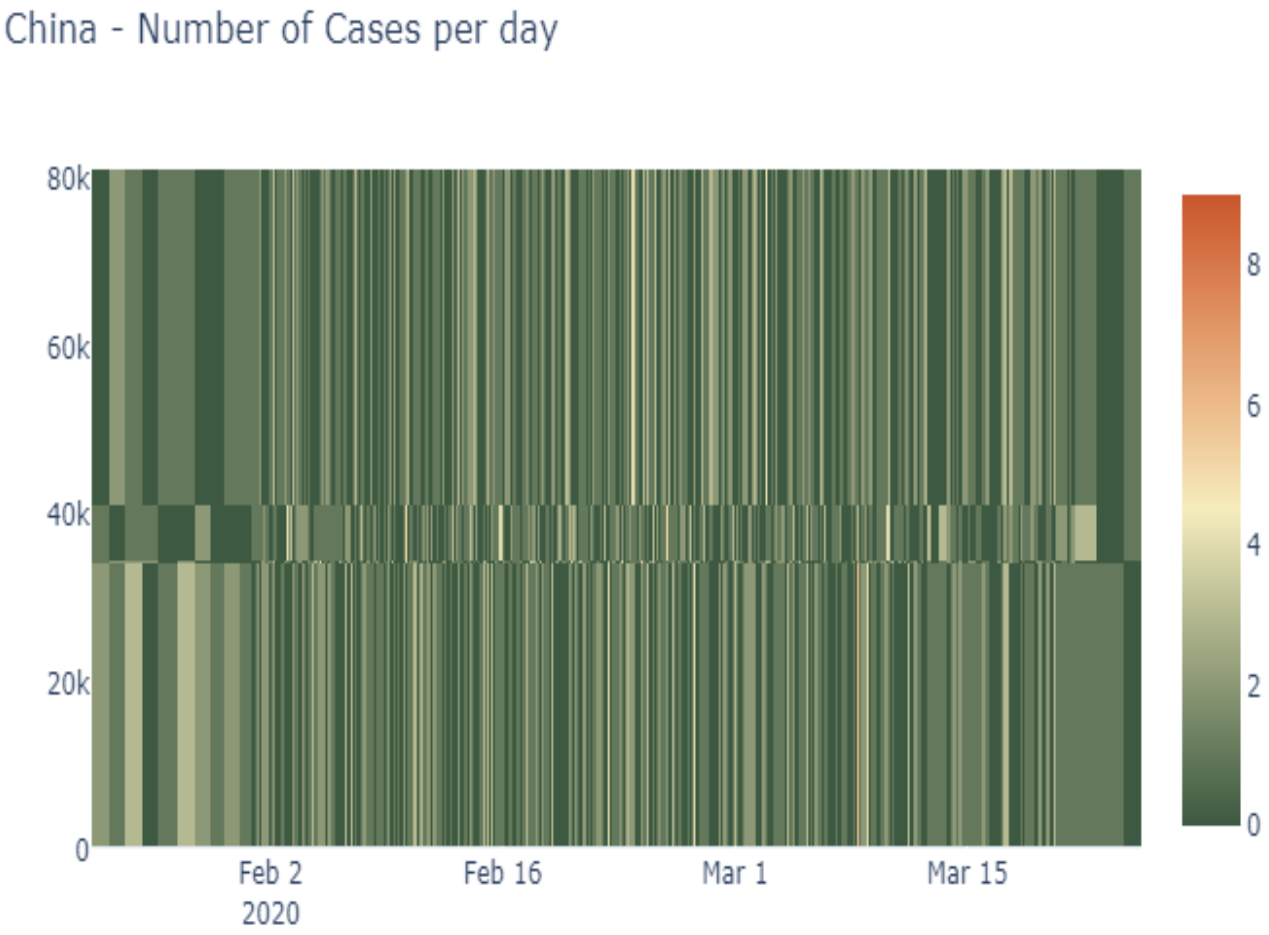


Figure 7.3: Density of no of cases at china

7.4 Density Mapbox

The project that depicts population density data for the entire world. To date, it visualize global 3D projections at this granular level in an interactive, browser-based map GHSL combines baseline data on human presence on the planet's surface using observations of buildings done via global Earth Observation systems (i.e, satellites).

visualizing the number of confirmed cases of corona virus increasing day by day .

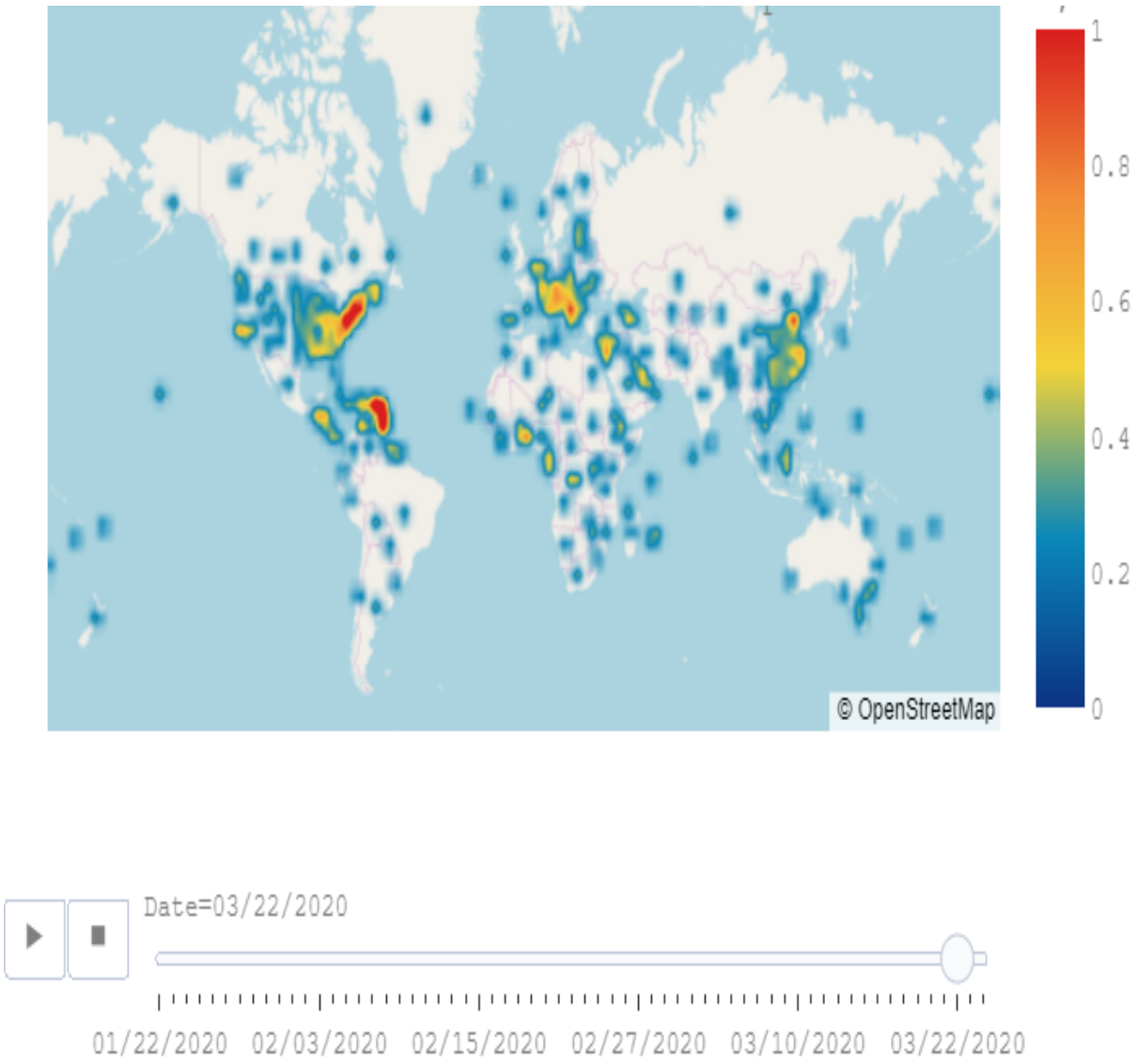
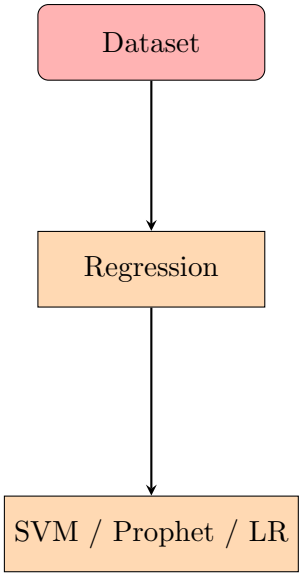


Figure 7.4: trend of spread and at what rate. Huge jump in March

Chapter 8

Proposed Approach

We have used the Regression and Time-Series Prediction approaches.



We have analysed the dataset and our goal is to find the number of cases which are real valued therefore we used regression models and the dataset is present in timeseries format so we have used timeseries prediction for (Confirmed, Death, Recovered) cases.

8.1 Disease Prediction

This Section predicts the number of Confirmed, Recovered, Deaths Cases in upcoming months by using different algorithms.
Input= No. of Confirmed, Recovered, Death (previous months) cases for particular date.
Output= No. of Confirmed, Recovered, Death (future months) cases for particular date.

8.1.1 Prophet Algorithm

Description

Prophet is an open source library published by facebook for forecasting time series data based on an additive model based on decomposible(trend ,seasonality , holiday) models.It provides fairly accurate predictions with simple parameters. Prophet uses time as a regressor and can fit several linear and non-linear components.

$y(t) = g(t) + s(t) + h(t) + \widehat{Et}$

where, g(t) is linear(default) or logistic growth.

s(t) is seasonality (based on fourier series).

h(t) is holiday effect.

$\widehat{Et}$ is error term.

Graphical Representation of Results

Confirmed Cases

This graph predicts the values for confirmed cases that there is a large increase in confirmed cases from april to july and then it will be satble after july.

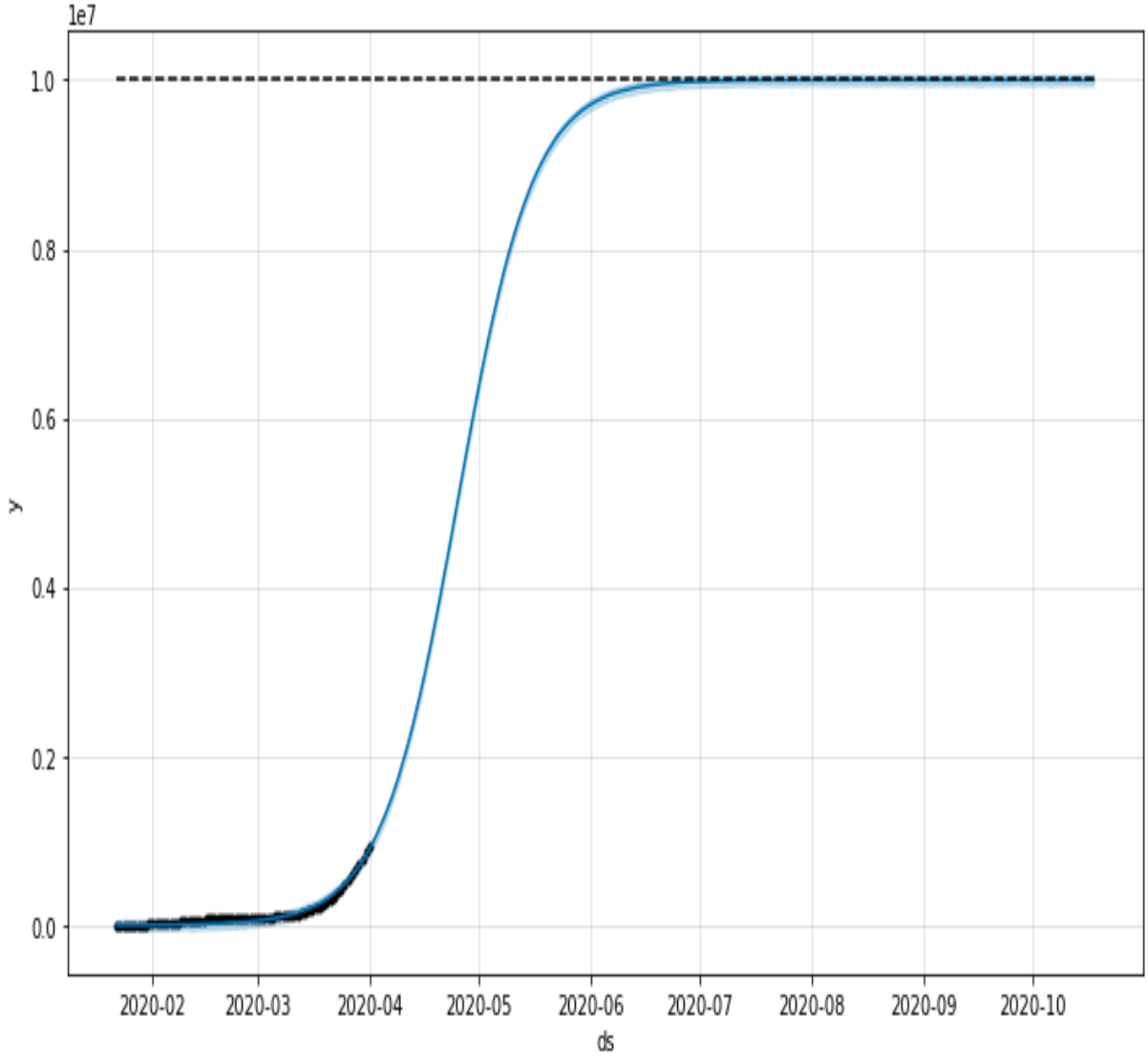


Figure 8.1: Confirmed Cases 1

These graphs shows the monthly and weekly trend of Confirmed Cases. In Weekly trend the least number of cases are on thursday.

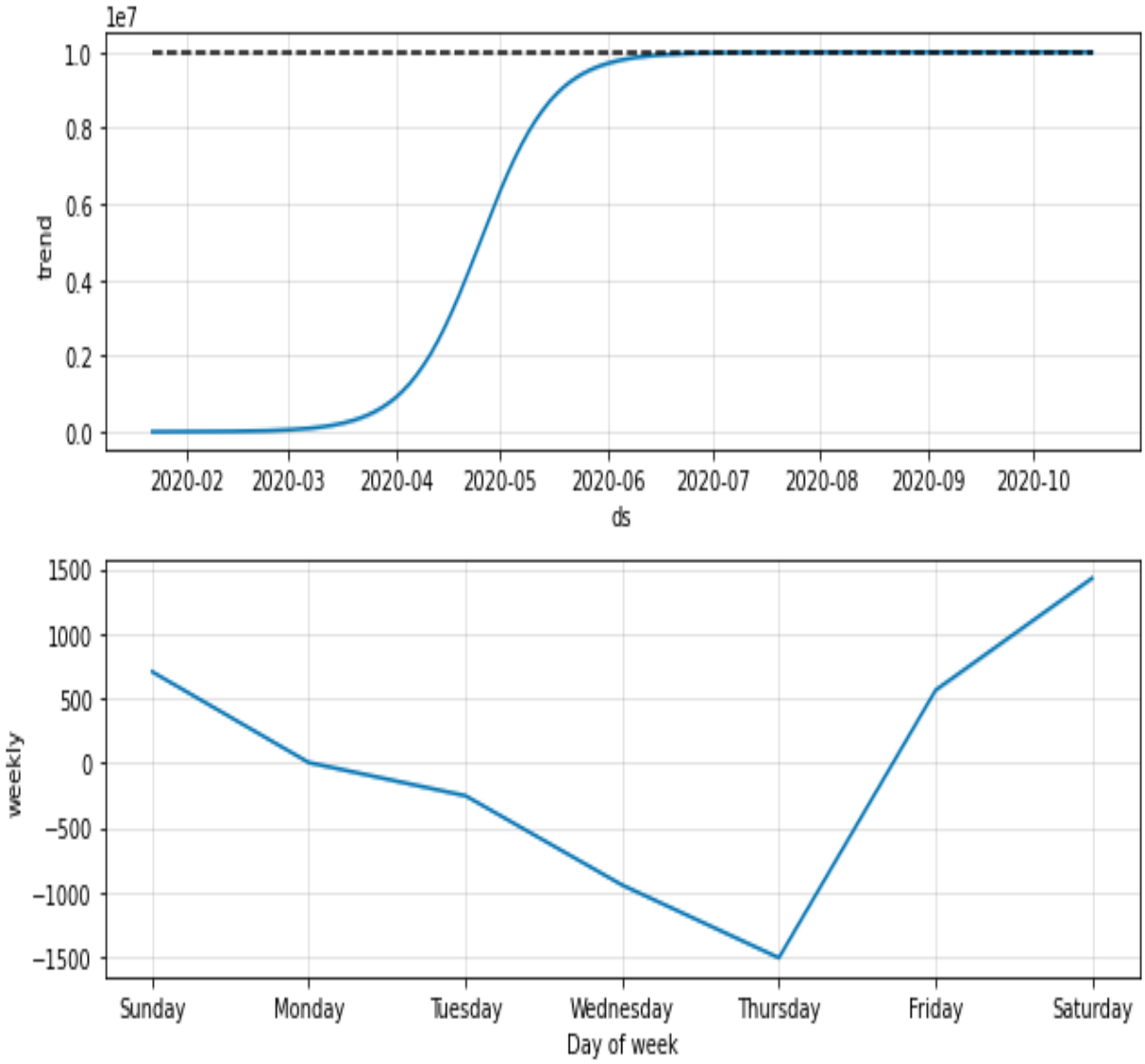


Figure 8.2: Confirmed Cases Trend

Death Cases

This graph predicts the values for death cases that there is a large increase in death cases from may to july and then it will be satble after july.

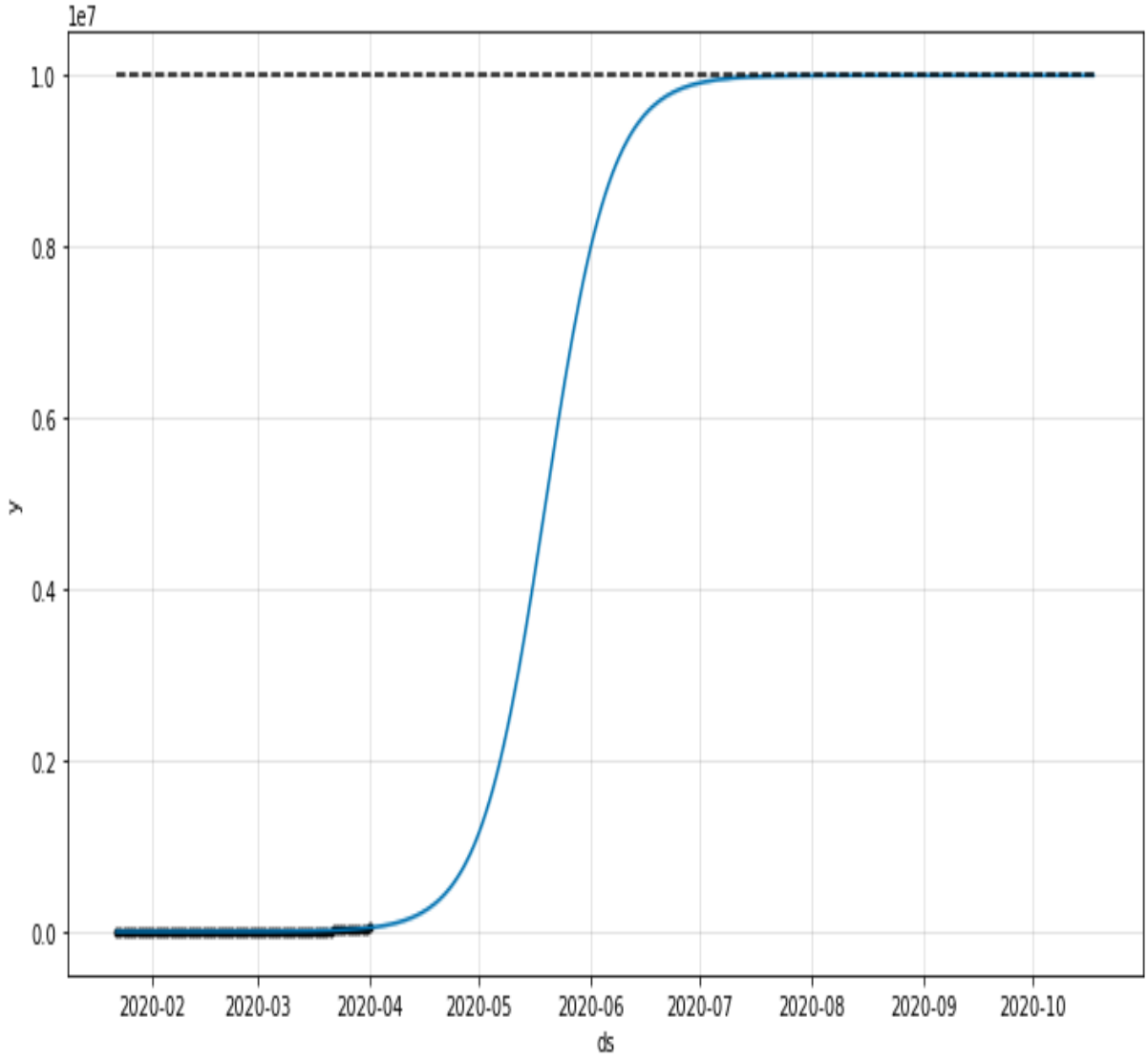


Figure 8.3: Death Cases 1

These graphs shows the monthly and weekly trend of Death Cases. In Weekly trend the least number of death cases are on thursday and then a large change of value on friday and saturday.

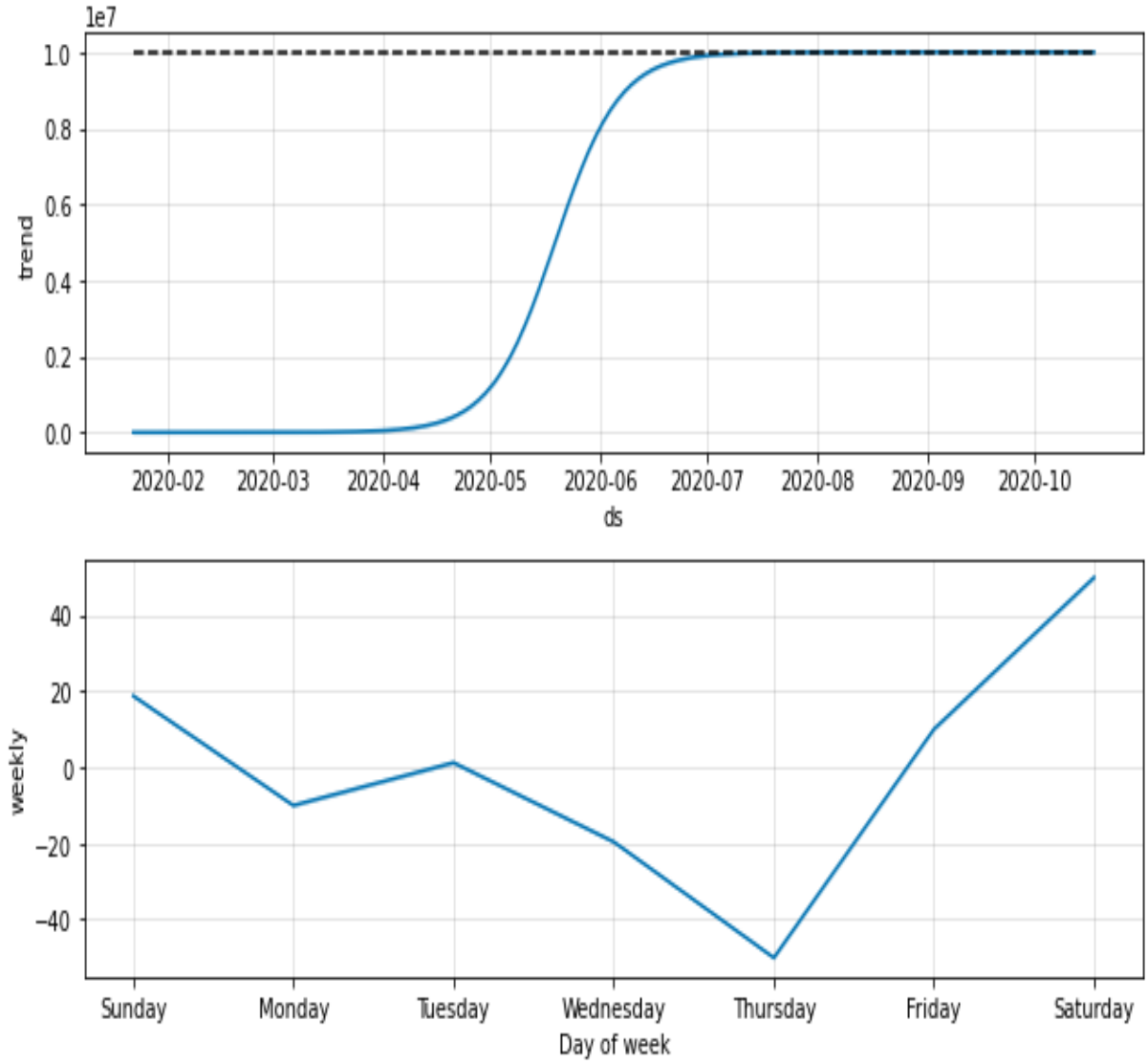


Figure 8.4: Deaths Cases Trend

Recovered Cases

This graph predicts the values for recovered cases that there is a large increase in recovered cases from may to september and then it will be stable after september.

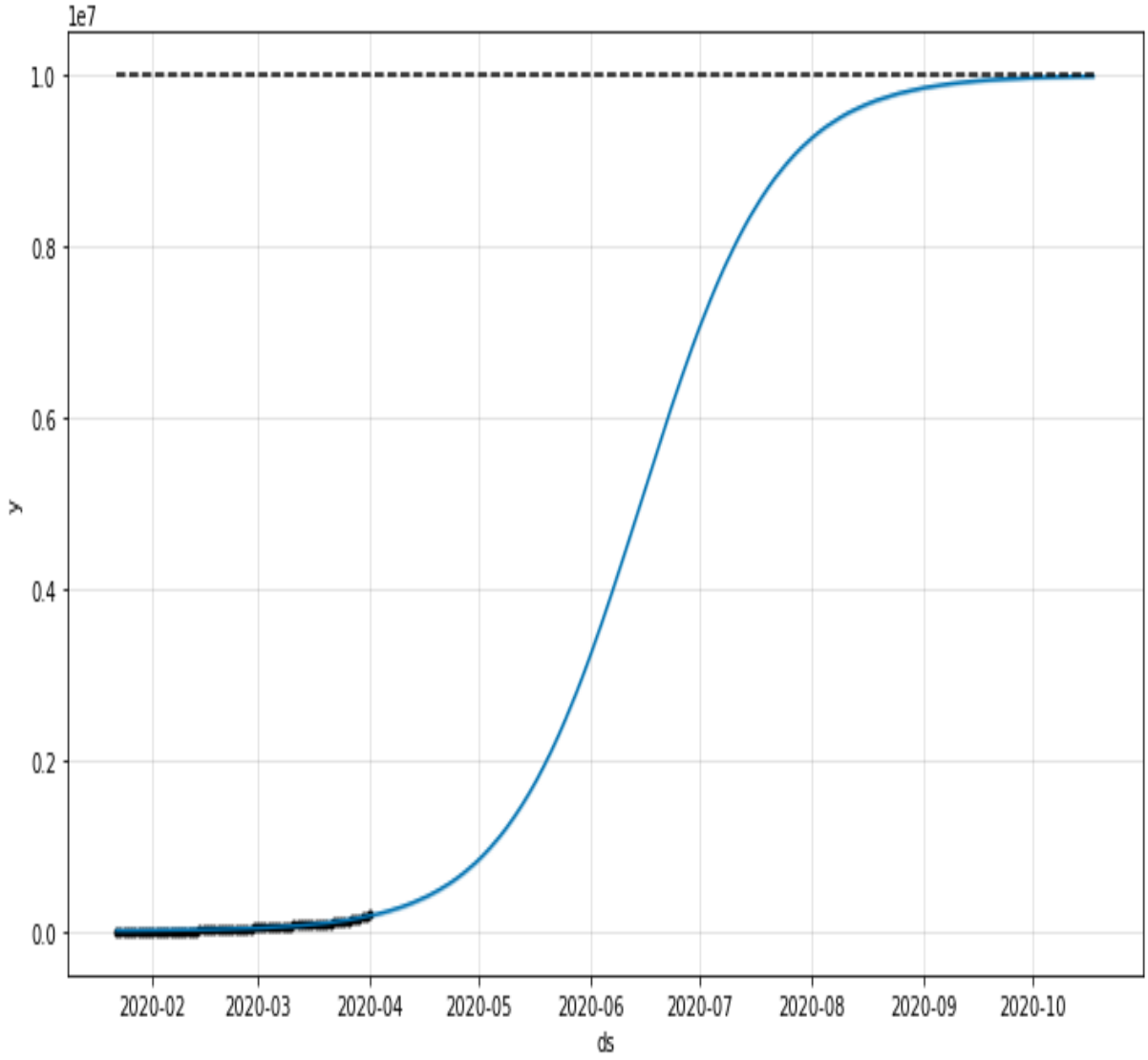


Figure 8.5: Recovery Cases 1

These graphs shows the monthly and weekly trend of Recovered Cases. In Weekly trend the least number of death cases are on thursday and then stable on friday and saturday.

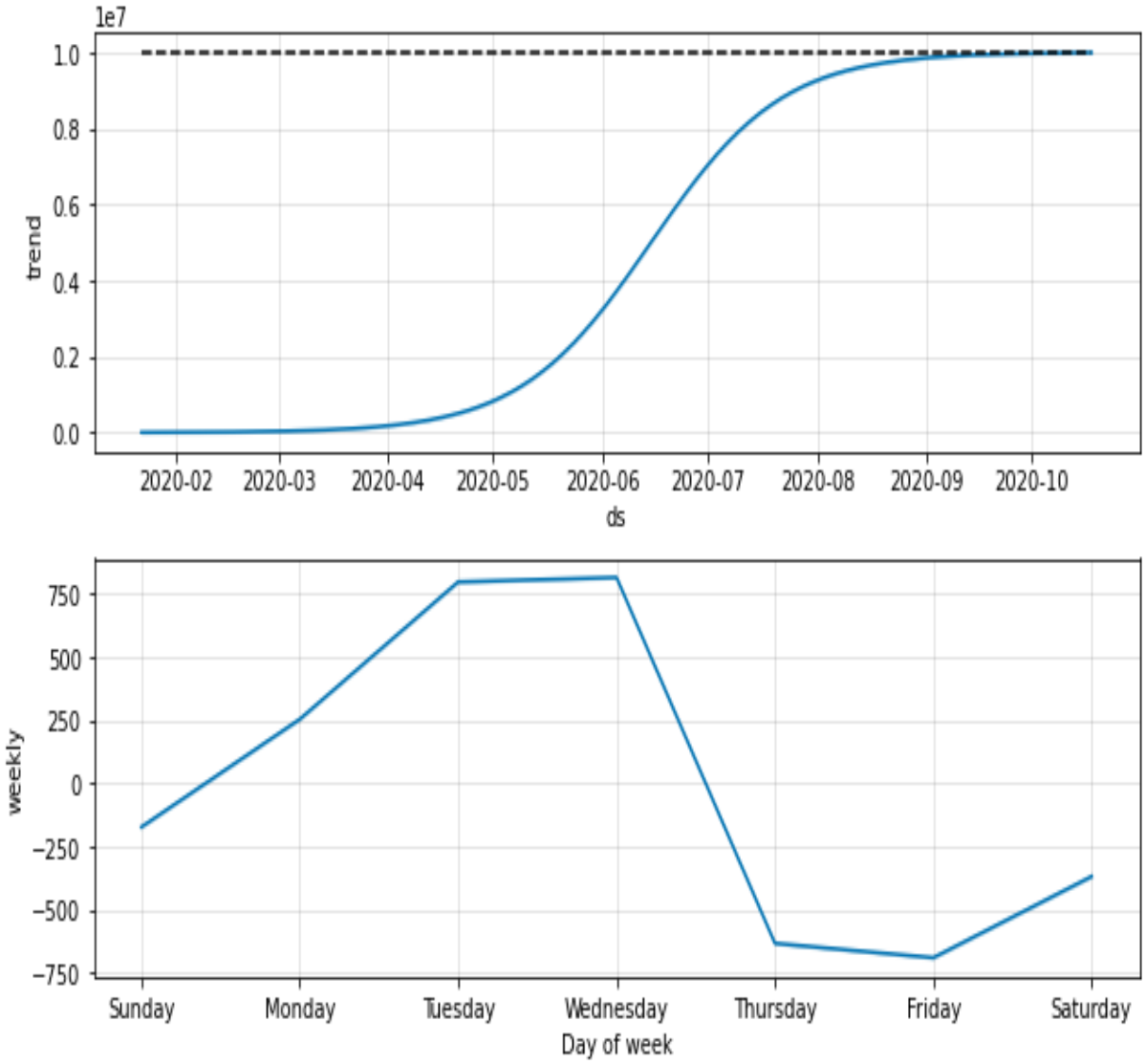


Figure 8.6: Recovery Cases Trend

8.1.2 Linear Regression Algorithm

Description

In statistics, linear regression is a linear approach to modeling the co-relation between dependent or independent variables).

The method in which only one explanatory variable is used is known as simple linear regression and For the method in which more than one explanatory variable, the method is known as multiple linear regression.

The representation is in form of linear equation that merges a set of input values (x) and the solution to which is the set of predicted output values (y).
Equation in the form Y = a + bX
where X is explanatory variable
Y is dependent variable.
The slope of the line is b, and a is the intercept (the value of y when x = 0).

Graphical Representation Of The Results

Confirmed Cases

This graph predicts the Confirmed Cases in given months and compares them with Actual confirmed data.This shows there is a gradual increase in no.of cases from mid march to april.

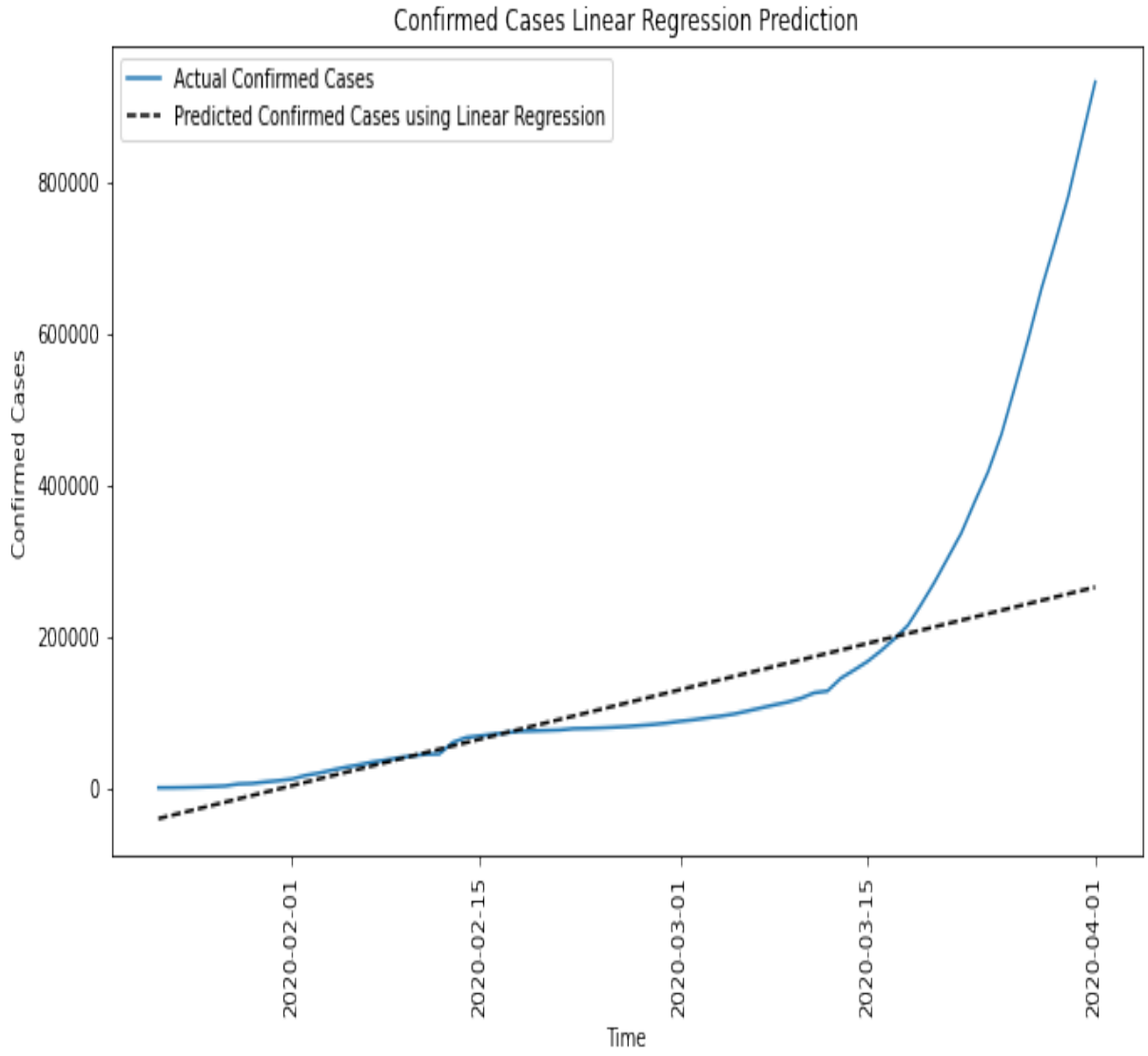


Figure 8.7: Confirmed Cases using Linear Regression

Death Cases

This graph predicts the Death Cases in given months and compares them with Actual Death data.This shows there is a gradual increase in no.of actual cases from mid march to april than predicted cases.

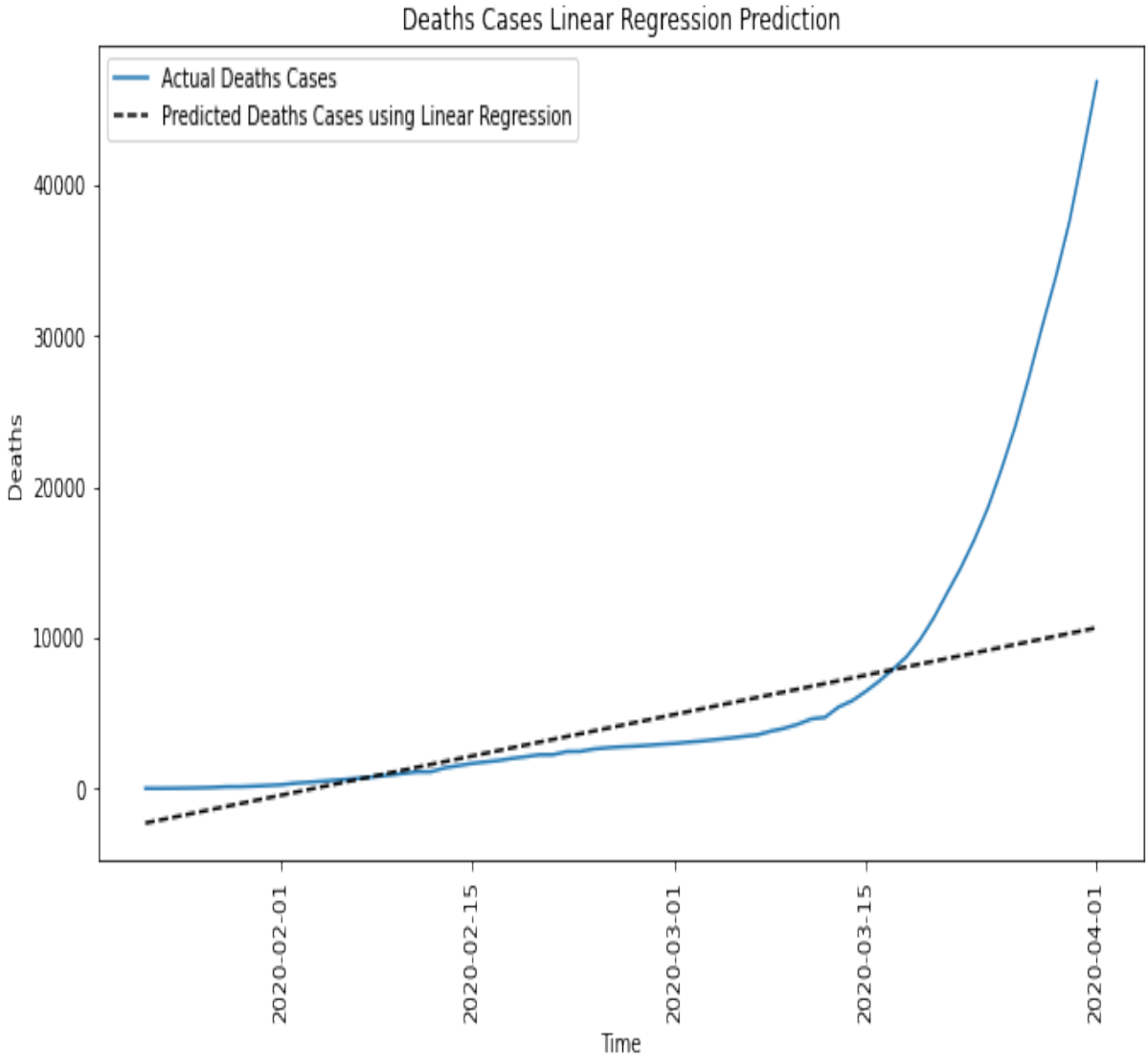


Figure 8.8: Death Cases using Linear Regression

Recovered Cases

This graph predicts the Recoverd Cases in given months and compares them with Actual Recoverd data.

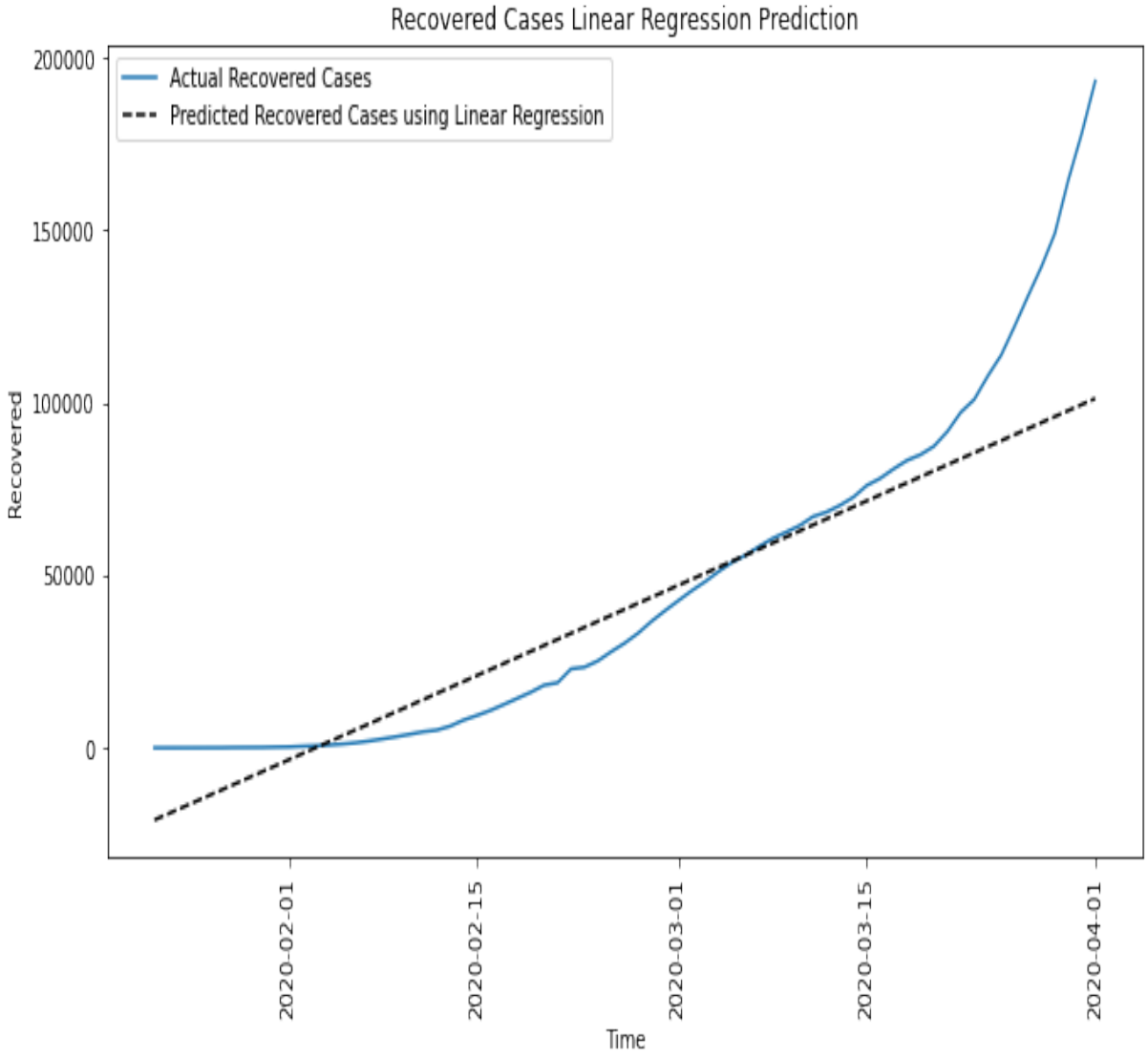


Figure 8.9: Recovery Cases using Linear Regression

8.1.3 Support Vector Machine Model Regressor Algorithm

SVM is used for classification as well as regression problem.

The main goal of SVM is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future.This best decision boundary is hyperplane.

There can be multiple lines /decision boundaries to segregate the classes in n-dimensional space,but we need to find out the best decision boundary that helps to classify the data points .This best boundary is known as the hyperplane of SVM.
Support vectors is the data point or vectors that are closest to hyperplane and which affect the position o the hyperplane.

Formula:

$$y = \sum_{i=1}^{N} (\alpha_i - \alpha_i^*) * (\phi(x_i), \phi(x)) + b$$

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

Graphical Representation Of The Results

Confirmed Cases

This graph shows that there is a slight change in predicted value and actual value of confirmed cases.

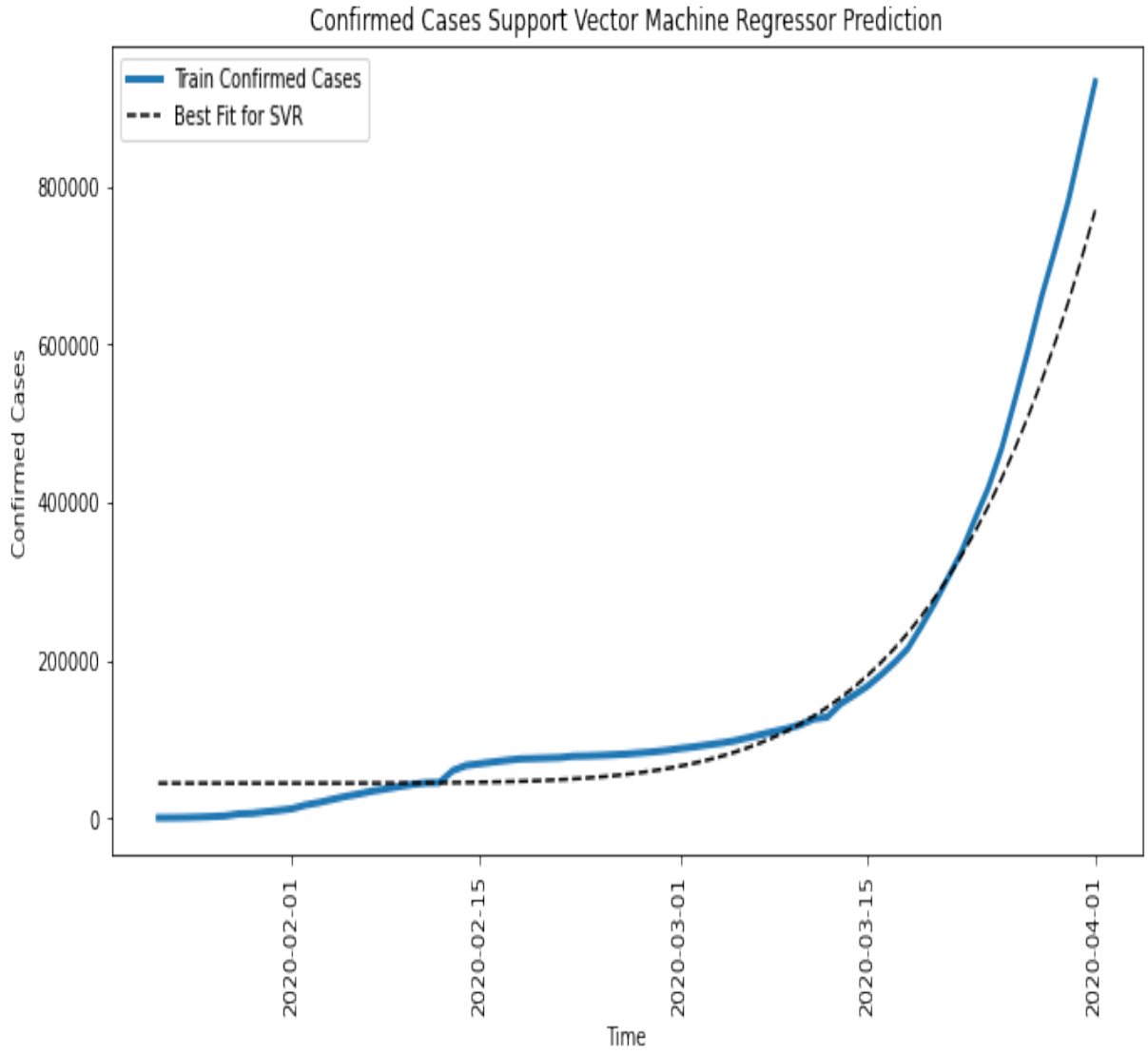


Figure 8.10: Confirmed Cases using Support Vector Machine Model

Death Cases

This graph shows that there is a slight change in predicted value and actual value of Death cases.

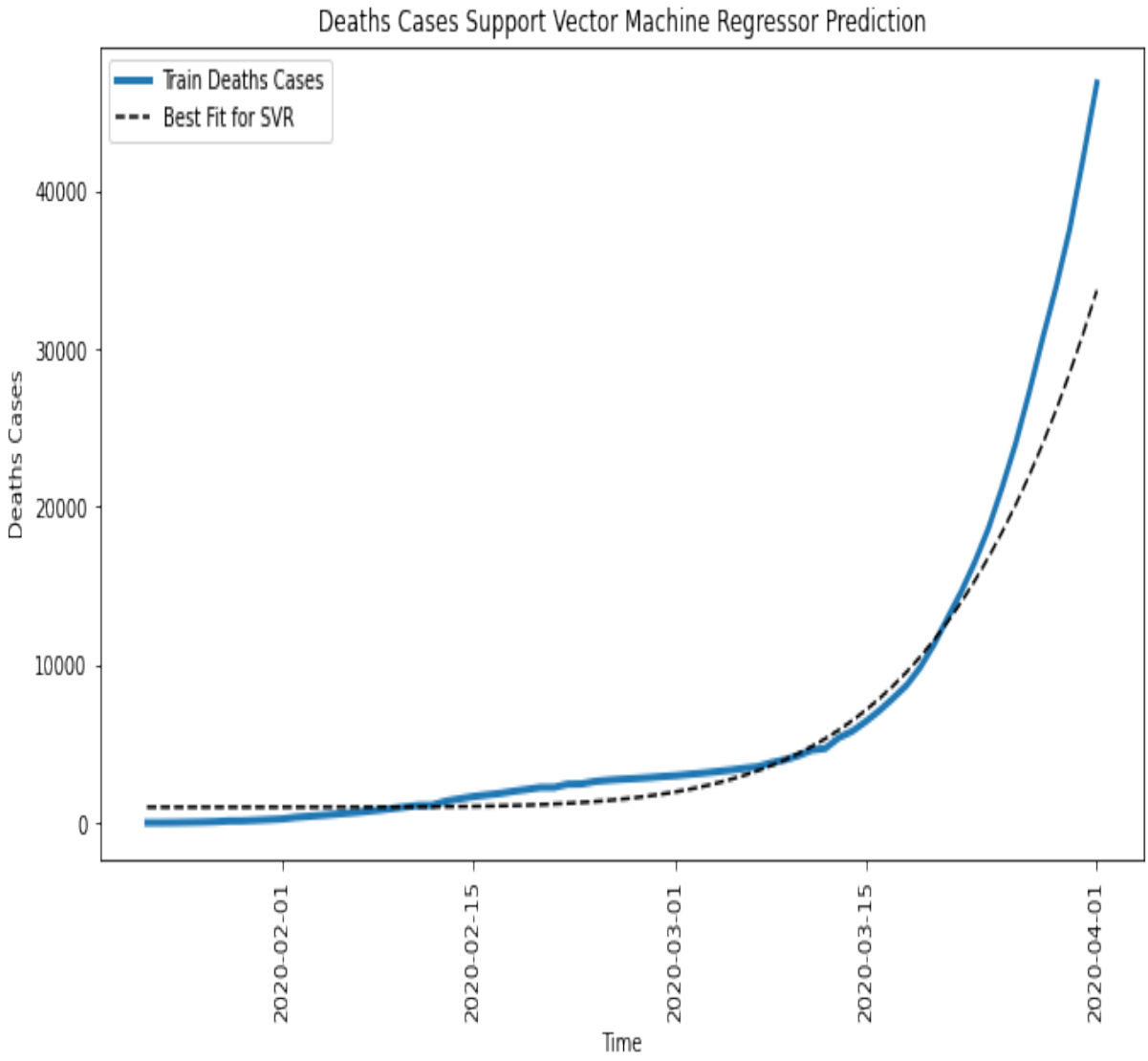


Figure 8.11: Death Cases using Support Vector Machine Model

Recovered Cases

This graph shows that there is a slight change in predicted value and actual value of recovered cases.

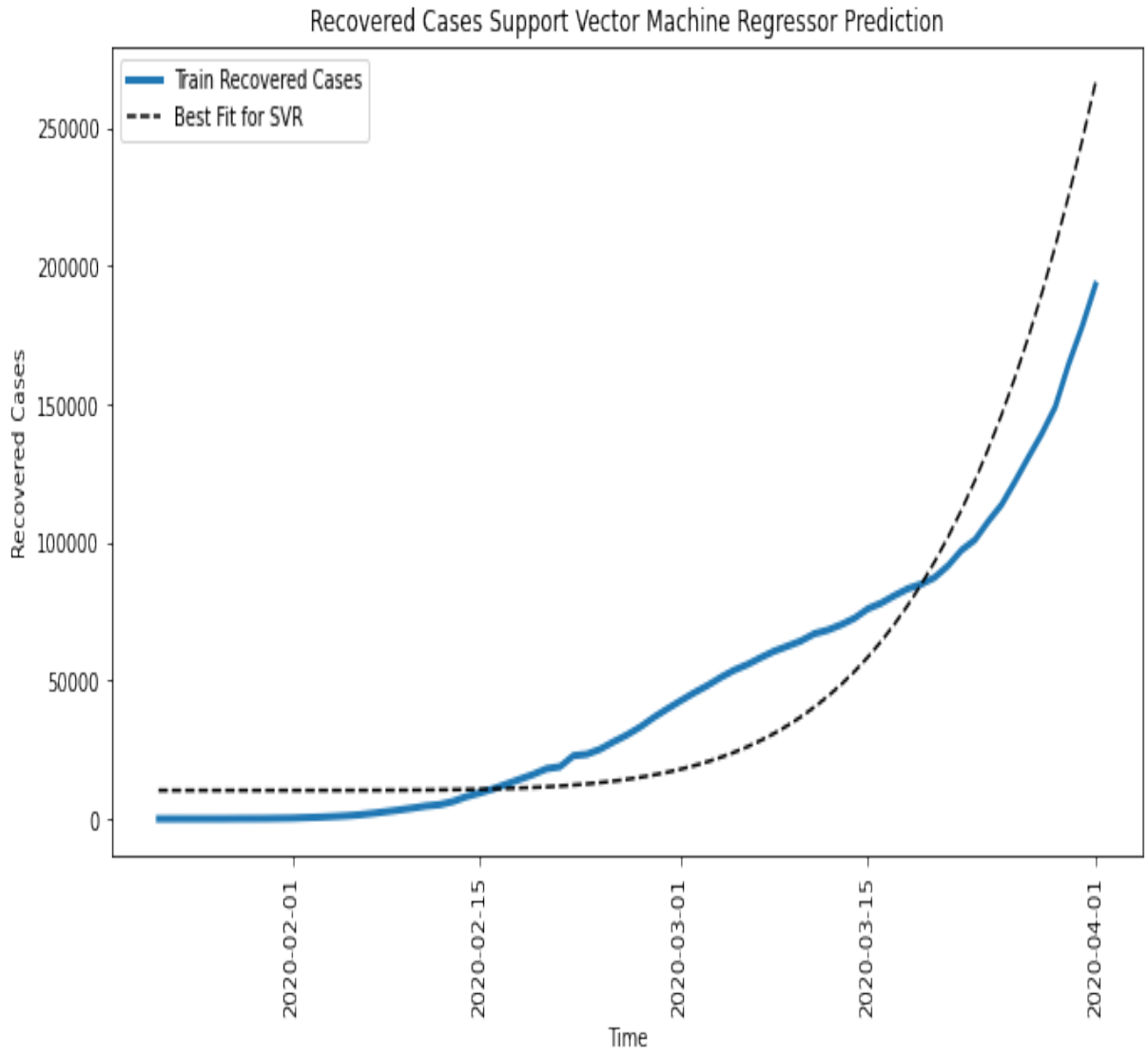


Figure 8.12: Recovery Cases using Support Vector Machine Model

Chapter 9

Model Evaluation

9.1 Description

Handling issues by evaluating the performance of a machine learning model, is an component of any data science project.It is very helpful in estimating the generalization accuracy of a model on future (unseen/out-of-sample) data.

Methods are divided into 2 categories:

1. Holdout
2. Cross-validation.

A test set is used by both the methods to evaluate model performance.

In order to prevent our model from remembering the whole training set, and it predict the correct label for any point in the training set. This is known as overfitting.

9.2 Evaluation on Prophet Algorithm

Prophet Algorithm evaluates performance automatically and it states the flag issues that requires manual intervention.

Setting a base line with some simple forecasting methods like seasonal naive,drift etc ,can be one of the simplest and the easiest evaluation methods ,whereas using a more complex model in order to know whether additional performance can be gained or not ,it is sometimes good to compare simplistic and advance forecasting methods.

Some of the evaluation metrics that can be used in cross validation are MAE(Mean Absolute Errors) MAPE(Mean Absolute Percentage Errors) ,RMSE(Root Mean Square Errors).

9.2.1 Confirmed Cases metrics

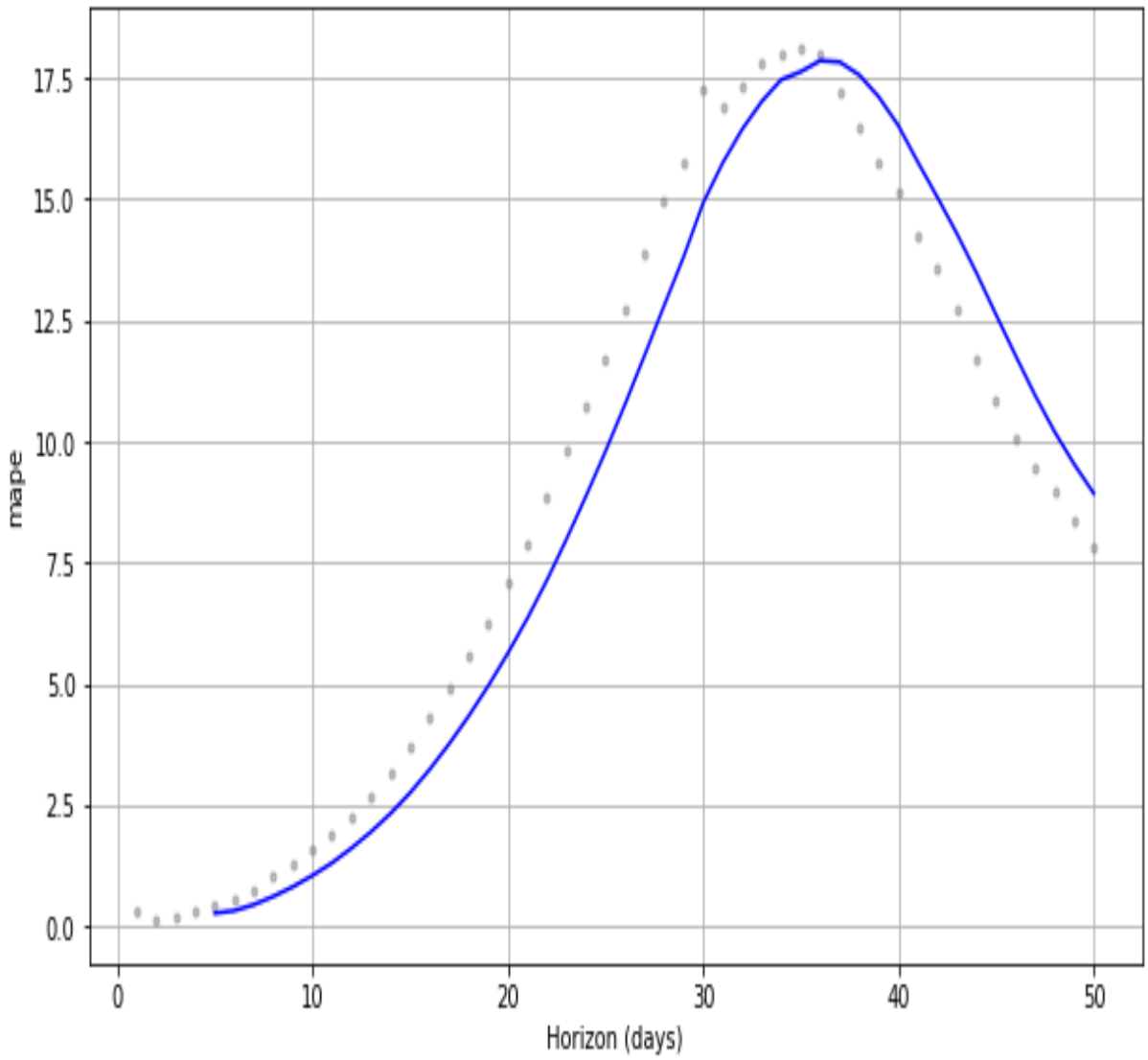


Figure 9.1: Confirmed Cases Evaluation

9.2.2 Death Cases metrics

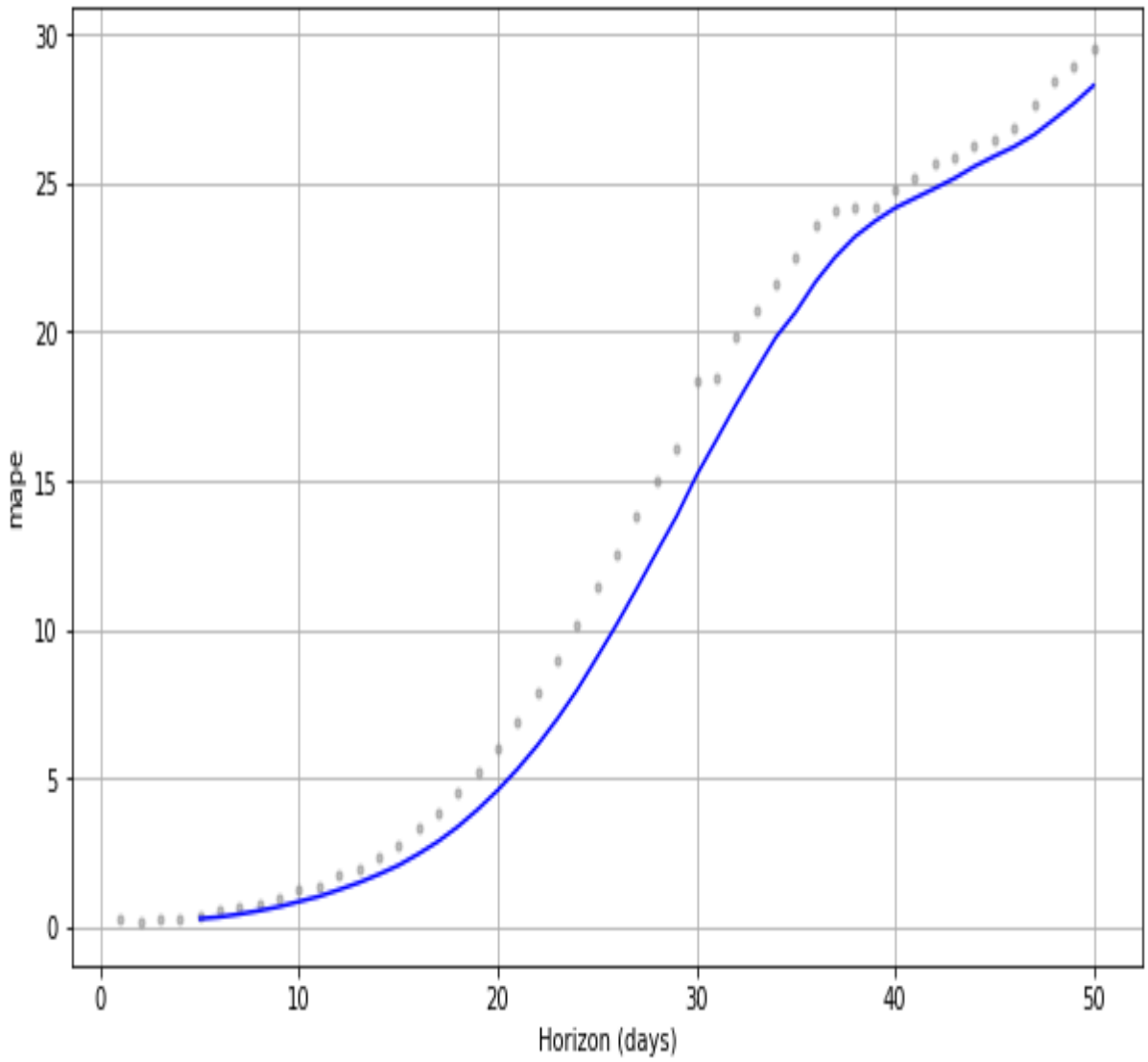


Figure 9.2: Deaths Cases Evaluation

9.2.3 Recovered Cases metrics

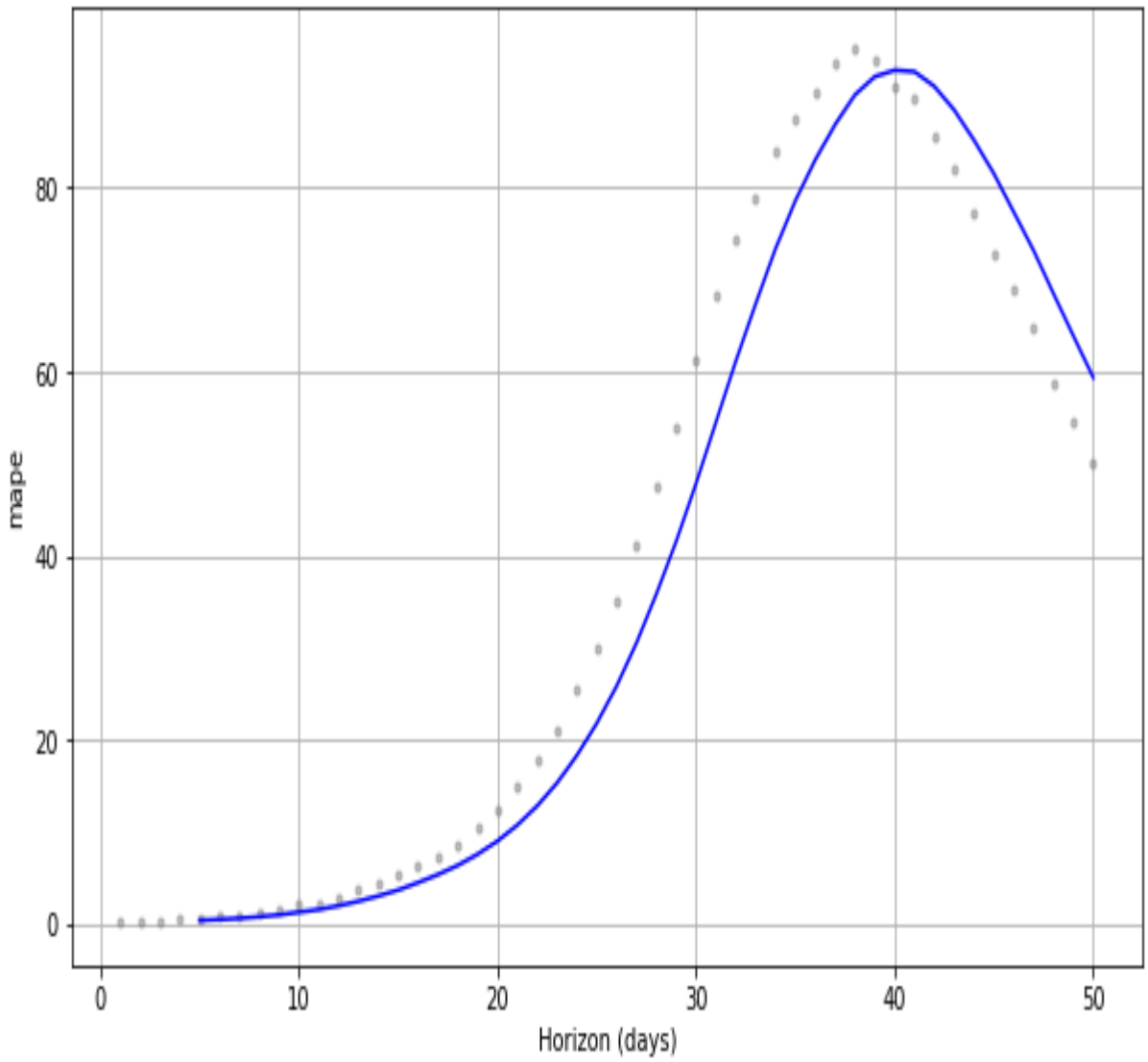


Figure 9.3: Recovery Cases Evaluation

Chapter 10

Data Clustering

In machine learning,a technique or method of unsupervised learning that involves the grouping of data points is known as Clustering.
This clustering algorithm is used to divide each data points into a group of data points having similar properties or features.
Data points of different groups have dissimilar properties or features. This techniques is also used for statistical data analysis,used in various fields .The fact that we can see ,in what groups our data points fall on applying clustering algorithm ,this clustering analysis helps us in gaining valuable insights from our data,in the field of data science.
Hierarchical clustering algorithm is divided into two categories:Top-down and Bottom-up
The bottom-up hierarchical clustering also known as HAC [Hierarchical Agglomerative Clustering]treats each data point as single cluster at the outset and then successively merges the pairs of clusters by the time all the clusters have been combined into a single cluster containing all data points.
A tree is used to represent that hierarchy of clusters,where the root of the tree is the unique . This tree is known as Dendogram.

Cluster gathering all the samples and leaves are the clusters with only one sample.

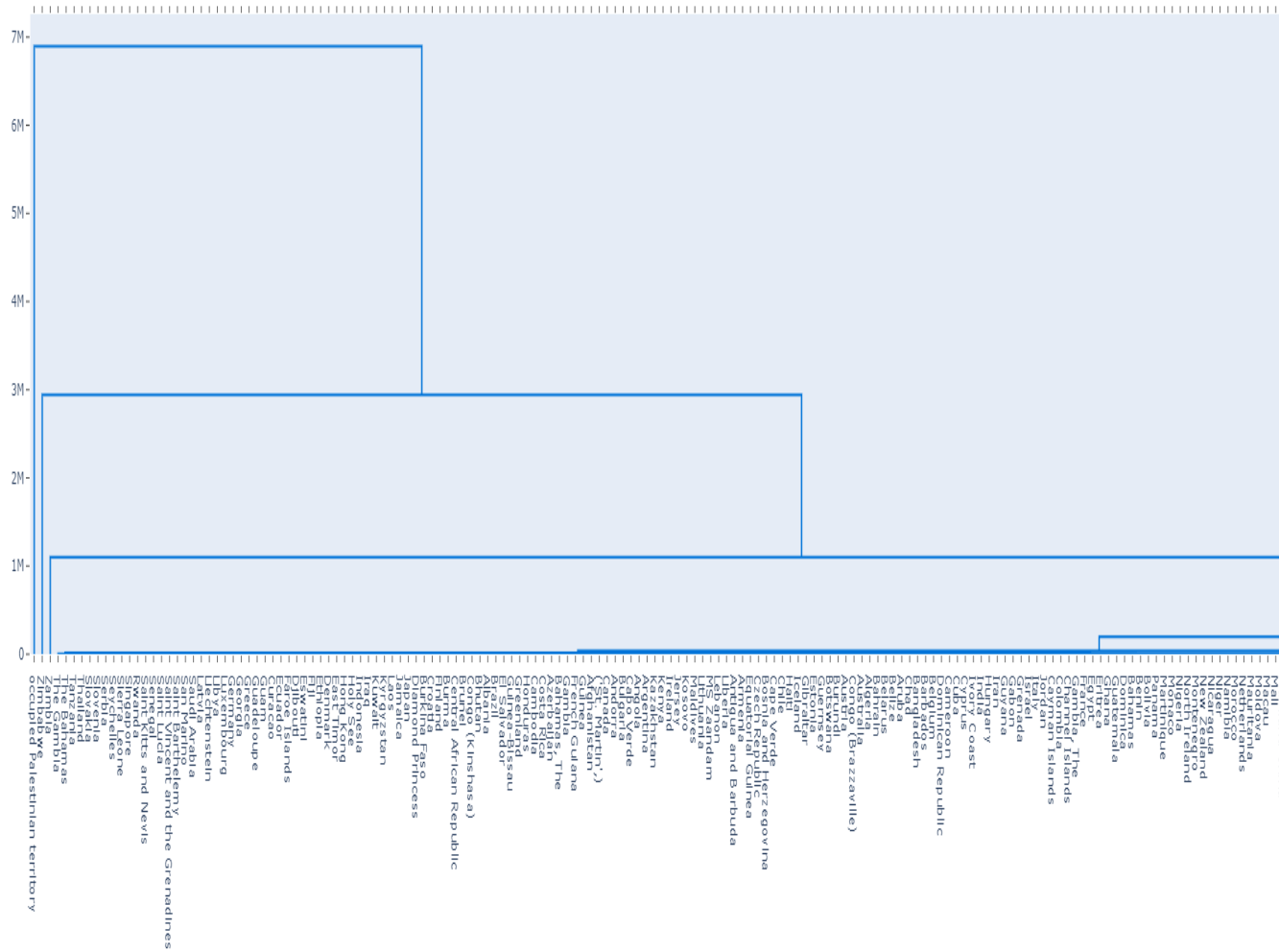


Figure 10.1: Clustering

Chapter 11

Result

The unsupervised learning model can be used to determine state wise Inter Cluster Movement to better understand the scenario geographically.

Through this journey of this project we came to the conclusion that the cases will rise gradually and plateau in the month of July. The deaths will plateau around July - August and the recovery rate will increase in July - August and plateau in the months of October-November .

These observations help us as certain that the governments can plan to restart their executive day to day work around July - August and the countries that have similar patterns ,as observed from the clusters , can work together into devising a common plan of action.

The countries can also learn from the other clusters and their journey to make a better informed decision

The accuracy of the supervised learning models can be further improved by including Temperature and Precipitation for observation sites as features for prediction. These values are not present in the dataset used and will have to be compiled from different sources to be added to the dataset.

More Regression algorithms can be used and compared for a more comprehensive approach.