Term Paper Report

On

Analysis and Visualisation On Covid-19 Outbreak and On The Medical Situation In India

Submitted to:
Amity University Uttar Pradesh

In partial fulfilment of the requirements for the award of the degree of
Bachelor of Technology
in
Computer Science and Engineering

by
Shivang Gupta
(A2305218642)
Under the guidance of
Ms. Smriti Sehgal

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
AMITY SCHOOL OF ENGINEERING AND TECHNOLOGY

AMITY UNIVERSITY UTTAR PRADESH
JUNE 2020

DECLARATION

I, Shivang Gupta student of B. Tech Computer Science and Engineering, hereby declare that the report entitled "**Analysis and Visualisation on Covid-19 outbreak and on the medical situation in India**" which is submitted by me to the Department of Computer Science and Engineering, Amity School of Engineering and Technology, Amity University, Noida, Uttar Pradesh in partial fulfilment of requirement for the award of the degree of Bachelor of Technology in Computer Science and Engineering, has not been previously formed the basis for the award of any degree, diploma or other similar title or recognition. The author attests that permission has been obtained for the use of any copyrighted material appearing in the report, other than brief excerpts requiring only proper acknowledgement in scholarly writing, and that all such use is acknowledged.

Submitted by:
Shivang Gupta
A2305218642
2018-2022
5CSE-10Y

CERTIFICATE

On the basis of declaration submitted by **Shivang Gupta** student of B.Tech CSE, I hereby certify that the project titled "***Analysis and Visualisation on Covid-19 outbreak and on the medical situation in India***" which is submitted to Department of Computer Science and Engineering, Amity School of Engineering and Technology, Amity University Uttar Pradesh, Noida, in partial fulfilment of the requirement for the award of the degree of Bachelor of Technology in Computer Science and Engineering, is an original contribution with existing knowledge and faithful record of work carried out by them under my guidance and supervision. To the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Ms. Smriti Sehgal
Department of computer science and engineering
ASET, Noida

ACKNOWLEDGEMENT

I would like to thank everyone who has helped me to accomplish my report. I sincerely thank all my respected teachers, who have helped me with their valuable and appropriate suggestions and supported me throughout the development of my report. I am highly thankful to my project guide Ms. Smriti Sehgal for providing her help and assistance at every stage of the study and research done for the report.

(Signature of the Student)
Date:

Shivang Gupta
(A2305218642)
5CSE-10Y

Index

1. Abstract

2. Introduction

3. Machine Learning Model Developed

4. Case Studies of Countries

5. Prediction of the machine learning model on real time data

6. Further Opportunities and Areas of research

7. Conclusion

8. References

Analysis and Visualisation on Covid-19 outbreak and on the medical situation in India

1. Abstract

The coronavirus outbreak started from the Hubei province of Wuhan, China, when they reported a sudden rise in the detection of pneumonia of an uncertain cause on 31st of December 2019.The outbreak of this Severe Respiratory Syndrome Coronavirus 2 continues to grow and has already spanned more than 5.5 million cases worldwide. What will be the impact of the COVID-19 pandemic? Machine Learning makes it possible to study and extract certain physical and clinical features and diagnose the disease in a better way. The need to analyse and visualise the current data of the infected, current cases and deaths are of an urgent need so that we are able to determine how much more the virus may spread, increase in the number of cases and if the government should increase the lockdown and other prevention methods. This report uses simple but highly powerful algorithms on the assumption that the data provided is genuine and the pattern of disease in the future will be constant. Also, all geological changes like temperature, wind direction, and other phenomenon are not taken into consideration. The paper has described three different models for predicting the future number of confirmed cases using SVM (Support Vector Machine), Polynomial regression and Bayesian Ridge. It also comments on the medical situation in India and what will be the trend of the same.

2. Introduction

2.1 Overview of the pandemic

The coronavirus infection has stunned the continents with its rapid multiplication, survival in extreme conditions and a fatal impact on the lives of billions of people. It has affected all phases of life, being it social during the lockdown or economical, hitting all businesses and industries majorly, making them lose billions of dollars in a span of few months. As of this writing, there are more than 5.5 million cases with around 350k deaths already recorded. There are several major differences between a normal SARS virus and a COVID-19 virus some of them being, that the COVID-19 virus spreads much more rapidly on human interaction and approximately twenty percent of the cases go undetected and the symptoms take 2-14 days to show their effects. Also, as per the reports of the WHO (World health Organisation), this virus is more fatal to people who are suffering from a disease already like HIV/AIDS, cancer, thyroid and diabetes. It would more likely affect people with a weaker immune system.

With no new discoveries of any curative vaccine and no reduction in the effect of the virus, the only way to curb and reduce the outspread is to practise self-isolation, to cut off the current pattern of community spread .In pandemic situation like this, where the number of cases are rising day by day, and data of the same is being generated in bulk, a powerful mathematical estimation and future forecast can be modelled using the current network of those affected, their personal details, their community movement and their receiving clinical aid which can be retrieved from the pharmaceutical and public health data available with the government. Apart from using data generated from public health sectors, other forms of data like messages on social media, articles, online blogs and awareness campaigns can also be used to extract valuable information regarding the current trend in community spread. Also travel history, tracking interaction with people can also be a source of information. All this data inculcated and tied could be used to develop a robust model which could be used to predict future spread rates and give valuable prediction for prevention and necessary steps and decisions that the government could take to minimise the outbreak and its effect.

2.2 Proposed model for forecasting and Visualisation

All data scientist around the world are spending day and night to come up with machine learning models which could help us in this situation and analyse the data flowing all over the internet and the web. A large amount of data has been made open to the public to aid in making strategies in management, business and to curb its effect. Although this initiative is highly appreciable, yet there is a need to make and develop models for better comprehension and knowledge.

In the proposed model we first graph the number of active cases country wise so as to better visualise the data and get a deeper understanding on the rapid spread of the virus. With the current graphs and data, we then go on to make country wise graph. Using three of the most successful and accurate algorithms we then determine and predict the number of cases in future as mentioned above.

3. Machine Learning Model Developed

3.1 Importing Libraries

1. ***"NumPy"*** **-** The core use of "numpy" library in python is data cleansing and manipulation. It is used to handle multi-dimensional data by supporting n-dimensional arrays and mathematical operations used on the data. It has all sorts of capabilities as to manipulate shape, sort, select, Fourier transformations and quick computations of algebra and statistics. It also has provisions for supporting C/C++.

2. ***"Pandas"*****-** Pandas, again is a library used to manipulate and analyse the data by making use of data-frames. It is necessary to before preparing the dataset to work on for training and other for testing. The pandas library is very robust and makes working on time-series and multi-dimensional data easier by enhanced reshaping, merge and join between various data sets, and handling NAN values in the dataset. It also helps in aligning the data, provides multiple indexing options and also offers various filtering techniques.

3. ***"Scikit-learn"*****-** or sklearn, is a library which is the most popularly used in machine learning models. It provides error-free compilation with the above-mentioned libraries and supports a large number of algorithms like – Decision-Tree, Random-forests, SVM, polynomial regression and many more. It also provides facilities for classification, regression, clustering, reduction in dimensionality and model selection and pre-processing.

4. ***"Xgboost"*****-** Xgboost is a powerful algorithm which takes an iterative approach in training and tuning the data .In situations where more than one model has to be trained, xg-boost trains all the models in succession so that which each new model being trained, it automatically corrects and accommodates the errors in the previous models. It improves the overall prediction and classification of data being made iteratively.

5. ***"Matplotlib"*****-** Matplotlib is used to for creative data-visualisation, used for making 2-D, 3-D, plots of the dataset filtered and give a clear perspective of the outliers and data .It is used to make a large number of statistical figures, graphs and provides utilities like box-plot/whisker-plot, grids, bar-graphs, pie charts, contour plot, wave frame, and surface plot etc. It also helps us to create and improve our model's accuracy by constructing the validation curve. It also supports intricate mathematical functions used by numpy and incorporates them well.

6. ***"Plotly"*** **–** It is an open-source plotting library that has the ability to work on forty different types of character types and create scientific, geographic, and 3-D plots. It enables us to integrate html to visualise data in beautiful web-based graphs and themes.

3.2 Data preparation and cleansing

1.Extracttion of number of affected (dead, recovered, latest) people

The data is imported in the jupyter notebook using pandas library. The confirmed, recovered, deaths and the latest data is imported into the file in the form of variables storing values in arrays. The head function is used to see the tabular form of data extracted from the source.

For better accuracy and prediction, we convert the dates which are in integer form to the date data-type. We then select all the dates of the outbreak from the data imported.

Two other parameters the mortality_rate and the recovery_rate is calculated for better visualisation. For making and understanding the spread of virus in each country, we create variables with the name of the country and assign them values from the data we imported as following:

Mortality-rate = $deaths/confirmed$

Recovery-rate = $recovered/confirmed$

1. *latest_data*:

```
latest_data.head()
```

	FIPS	Admin2	Province_State	Country_Region	Last_Update	Lat	Long_	Confirmed	Deaths	Recovered	Active	Combined_Key
0	45001.0	Abbeville	South Carolina	US	2020-05-22 02:36:51	34.223334	-82.461707	36	0	0	36	Abbeville, South Carolina, US
1	22001.0	Acadia	Louisiana	US	2020-05-22 02:36:51	30.295065	-92.414197	269	15	0	254	Acadia, Louisiana, US
2	51001.0	Accomack	Virginia	US	2020-05-22 02:36:51	37.767072	-75.632346	709	11	0	698	Accomack, Virginia, US
3	16001.0	Ada	Idaho	US	2020-05-22 02:36:51	43.452658	-116.241552	792	23	0	769	Ada, Idaho, US
4	19001.0	Adair	Iowa	US	2020-05-22 02:36:51	41.330756	-94.471059	6	0	0	6	Adair, Iowa, US

2. *confirmed_df:*

```
confirmed_df.head()
```

	Province/State	Country/Region	Lat	Long	1/22/20	1/23/20	1/24/20	1/25/20	1/26/20	1/27/20	...	5/17/20	5/18/20	5/19/20	5/20/20	5/21/20	5/22
0	NaN	Afghanistan	33.0000	65.0000	0	0	0	0	0	0	...	6664	7072	7653	8145	8676	9
1	NaN	Albania	41.1533	20.1683	0	0	0	0	0	0	...	946	948	949	964	969	
2	NaN	Algeria	28.0339	1.6596	0	0	0	0	0	0	...	7019	7201	7377	7542	7728	7
3	NaN	Andorra	42.5063	1.5218	0	0	0	0	0	0	...	761	761	761	762	762	
4	NaN	Angola	-11.2027	17.8739	0	0	0	0	0	0	...	48	50	52	52	58	

5 rows × 130 columns

3. *deaths_df:*

```
deaths_df.head()
```

	Province/State	Country/Region	Lat	Long	1/22/20	1/23/20	1/24/20	1/25/20	1/26/20	1/27/20	...	5/17/20	5/18/20	5/19/20	5/20/20	5/21/20	5/22
0	NaN	Afghanistan	33.0000	65.0000	0	0	0	0	0	0	...	169	173	178	187	193	2
1	NaN	Albania	41.1533	20.1683	0	0	0	0	0	0	...	31	31	31	31	31	
2	NaN	Algeria	28.0339	1.6596	0	0	0	0	0	0	...	548	555	561	568	575	5
3	NaN	Andorra	42.5063	1.5218	0	0	0	0	0	0	...	51	51	51	51	51	
4	NaN	Angola	-11.2027	17.8739	0	0	0	0	0	0	...	2	3	3	3	3	

5 rows × 130 columns

4. *recoveries-df*

```
recoveries_df.head()
```

	Province/State	Country/Region	Lat	Long	1/22/20	1/23/20	1/24/20	1/25/20	1/26/20	1/27/20	...	5/17/20	5/18/20	5/19/20	5/20/20	5/21/20	5/22
0	NaN	Afghanistan	33.0000	65.0000	0	0	0	0	0	0	...	778	801	850	930	938	9
1	NaN	Albania	41.1533	20.1683	0	0	0	0	0	0	...	715	727	742	758	771	
2	NaN	Algeria	28.0339	1.6596	0	0	0	0	0	0	...	3507	3625	3746	3968	4062	4
3	NaN	Andorra	42.5063	1.5218	0	0	0	0	0	0	...	617	624	628	639	639	
4	NaN	Angola	-11.2027	17.8739	0	0	0	0	0	0	...	17	17	17	17	17	

2. Extraction of the dates from the data source

The other information we need in the form of the data is the dates and the number of patients on each day, and also the daily increase in the number of cases so that our algorithms can predict the future increase in the number of cases

For this data to be extracted, we need to fetch dates from the confirmed_df using the loc[] selector and store them in a variable "confirmed". Similarly, we can create a date wise increase in the number of deaths, recoveries and the latest data.

3. Creating county wise case studies

We assign each country variables to store the number of infected people residing there. From the same information we can use "matplot" library to create graphs and plots of the infected people there.

3.3 Machine learning algorithms used to predict future cases

1. ***Support-Vector-Machine (SVM)* –** It is a supervised ML algorithm used both to classify and predict the outcome of the data using regression. They have the power to handle various continuous and categorical character type variables. The algorithm is simply a representation of categories in a "*hyperplane in multi-dimensional space".* The goal of the algorithm is to separate data classes to find a "*maximum marginal hyperplane".*
   Support vector are the closest data points to the plane in the multi-dimensional space. The hyper plane is usually defined with the help of support vectors.
   Hyperplane is the plane which divides the classes and set of data points.
   Margin is the distance or the space between two lines on the nearest data sets of dissimilar classes.

   SVM Kernels - The kernel in the SVM is used to change the source data space into the required form, usually form a low dimensional data space to a high-dimensional data space, or transforms non distinguishable problems into distinguishable ones.

Types of SVM kernels:

1.1 Linear Kernel: States that "The product between any two vectors is the sum of the multiplication of each pair of input values"

$$K(x, x_i) = sum(x * x_i)$$

1.2. Polynomial Kernel: Polynomial Kernel is a modification to the linear Kernel which has the ability to differentiate between curves and non-linear input space. For a polynomial of degree *d*:

$$k(X, Xi) = 1 + sum(X * X_i)\text{^}d$$

1.3. Radial Basis Function (RBF) Kernel: The most widely used kernel in

SVM which maps the data space into infinite dimensional space. Here gamma is a hyper-parameter, whose value can be changed for tuning the machine learning model and its value ranged from 0-1.

$$K(x, xi) = exp(-gamma * sum(x - xi\text{^}2))$$

2. ***Polynomial Regression:*** As the name suggests, polynomial regression is used to create and fir the data points on the curve or a polynomial of degree n.

$$y = a + b1x + b2x\text{^}2 + \ldots + bnx\text{^}n$$

Polynomial regression is just a modification to the linear regression where a large number of regression lines are used to increase the accuracy of the predicted data. Used to fit intricate non-linear functions and datasets.

3. Bayesian Ridge: In simple linear regression we try to fit and accommodate points on a single line using data estimates. However in Bayesian ridge, linear regression is done using probability distributions. The output or the predicted value is drawn from the probability normal distribution, which is defined by the mean and the variance.

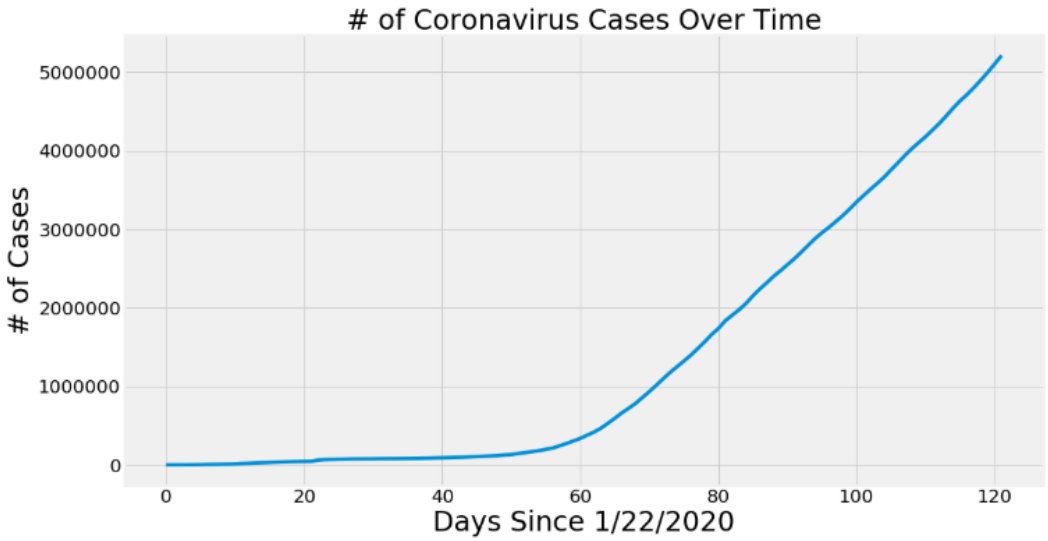
$$Posterior = \frac{Likelihood * Prior}{Normalization}$$

Since all the data points in machine learning are characterised by vectors, the mean is given by the "*transpose of the weight matrix into the predictor matrix" The* main objective of the algorithm is to determine the posterior for the model parameters.

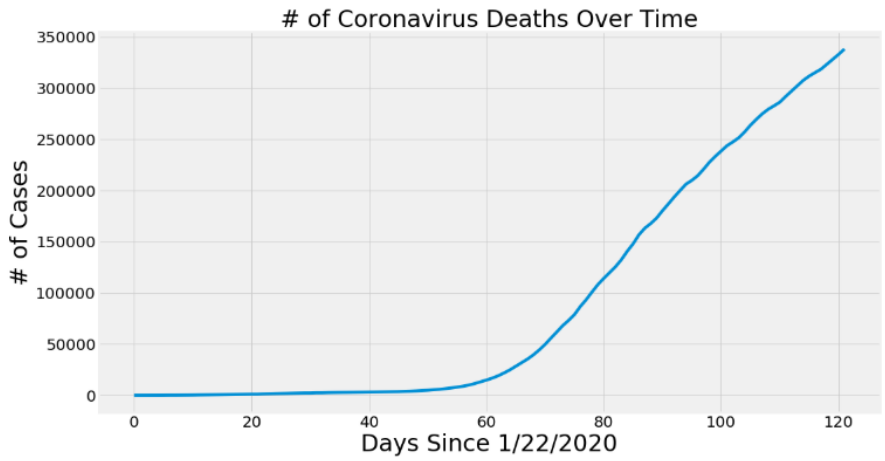
$$P(\beta|y,X) = \frac{P(y|\beta,X) * P(\beta|X)}{P(y|X)}$$

3.4 Graphing the number of confirmed cases, active cases, deaths, recoveries, mortality rate, and recovery rate

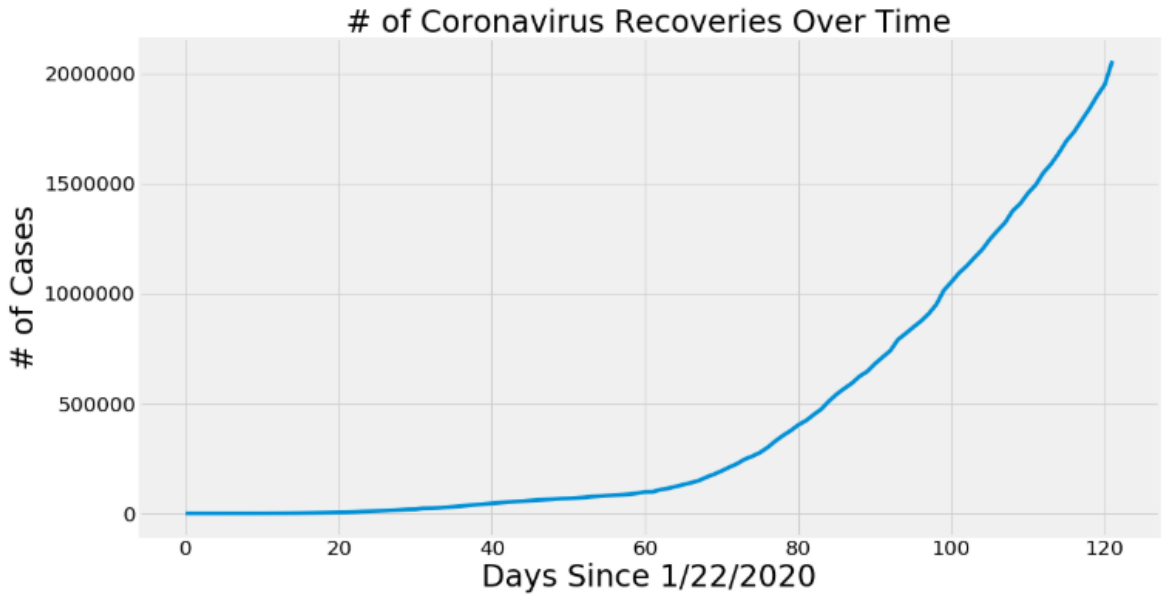
1. ***Worldwide -Cases:***



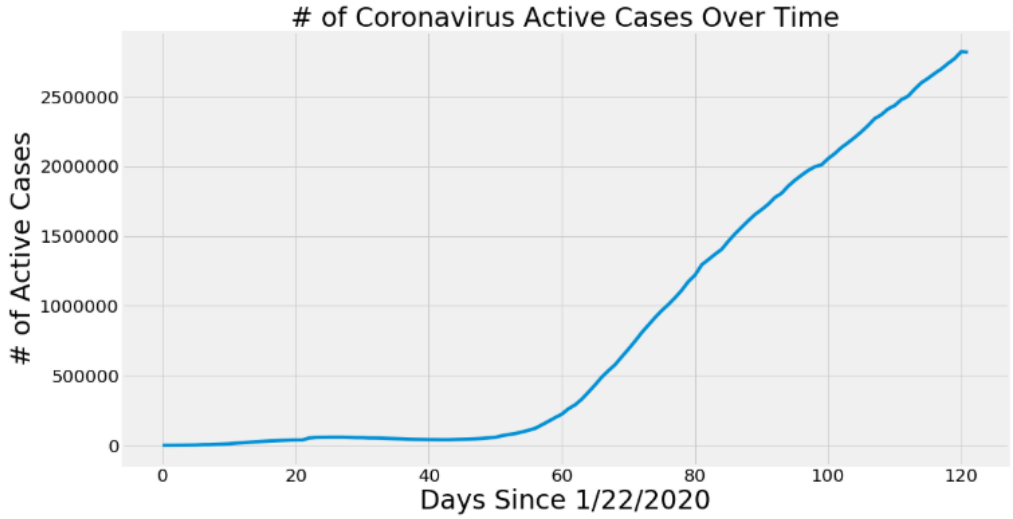
2. ***Death-cases:***



3. *Recovery-Rate*



4. *Active-Cases*



3.5 Prediction estimated by the algorithms used on the test data (30% data source)

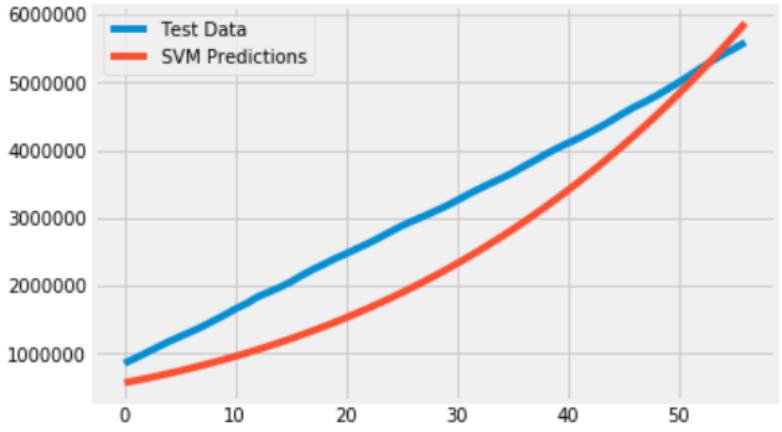
The data set is divided into 2 parts: one a training data based on the model is trained and other a testing data based on which the model's accuracy is determined. The following are the outputs based on the most optimal algorithms used:

1. SVM:

Mean-Absolute-Error: 646256.1342828622
Mean-Squared-Error: 499160815454.5278

	Date	SVM Predicted # of Confirmed Cases Worldwide
0	05/23/2020	5056087.0
1	05/24/2020	5222926.0
2	05/25/2020	5393884.0
3	05/26/2020	5569028.0
4	05/27/2020	5748427.0
5	05/28/2020	5932148.0
6	05/29/2020	6120261.0
7	05/30/2020	6312835.0
8	05/31/2020	6509940.0
9	06/01/2020	6711646.0

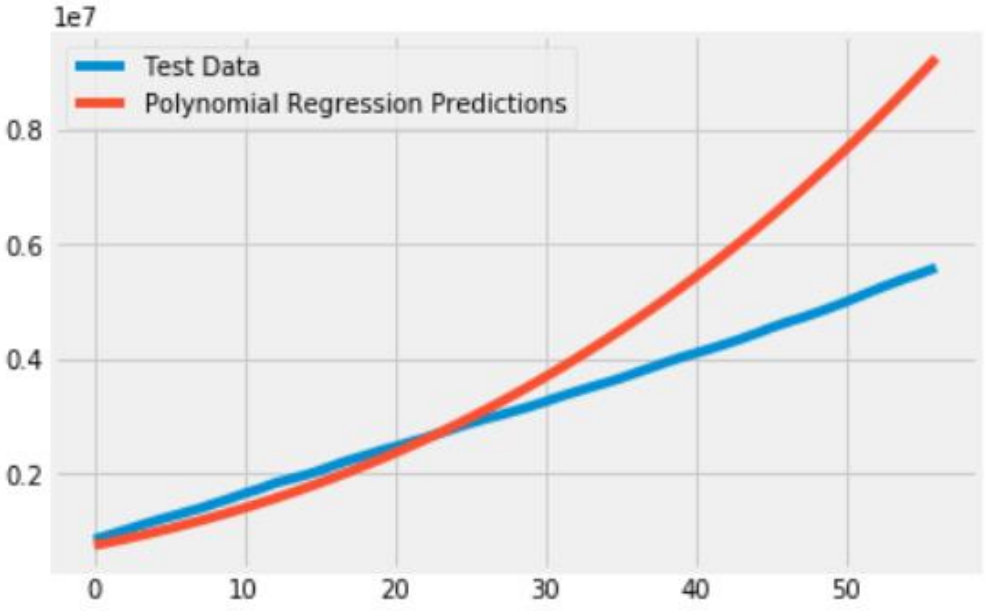
```
MAE: 646256.1342828622
MSE: 499160815454.5278
```



2. *Polynomial Regression:*

Mean-Absolute-Error: 953699.56687965
Mean-Squared-Error: 2020306494191.2744

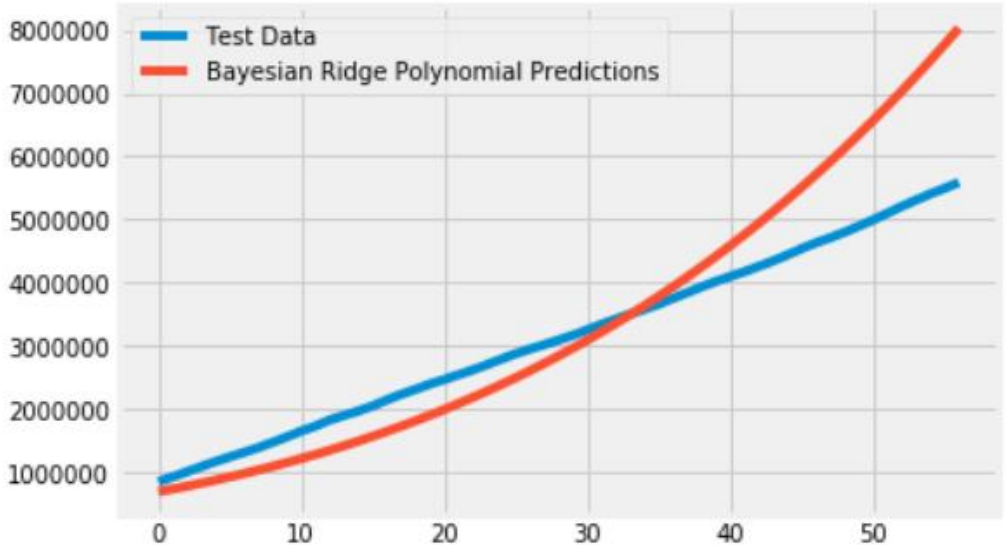
	Date	Polynomial Predicted # of Confirmed Cases Worldwide
0	05/23/2020	7489341.0
1	05/24/2020	7725018.0
2	05/25/2020	7965603.0
3	05/26/2020	8211147.0
4	05/27/2020	8461699.0
5	05/28/2020	8717310.0
6	05/29/2020	8978030.0
7	05/30/2020	9243911.0
8	05/31/2020	9515001.0
9	06/01/2020	9791352.0



3. Bayesian Ridge:

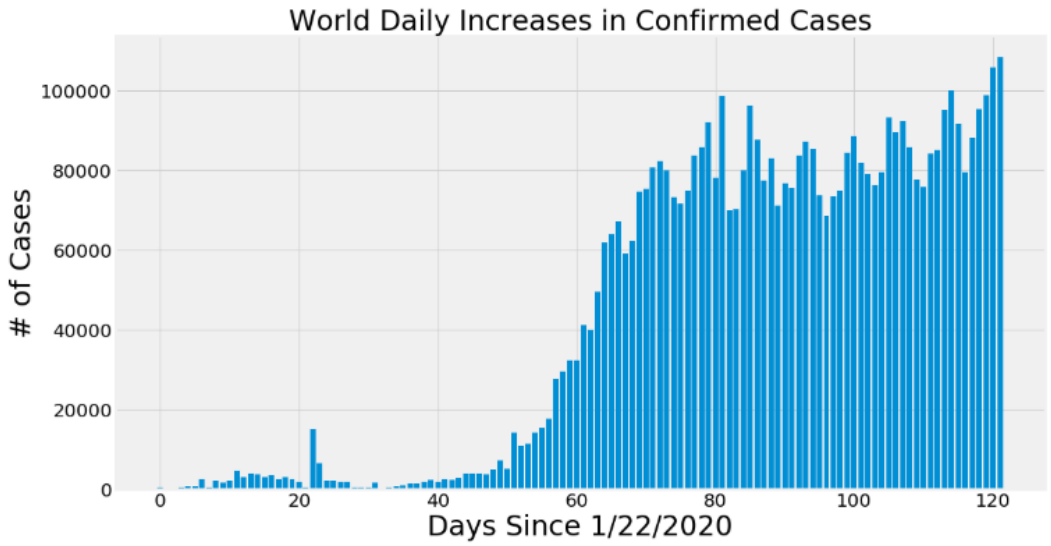
Mean-Absolute-Error: 638308.4556198396
Mean-Squared-Error: 754418716142.2955

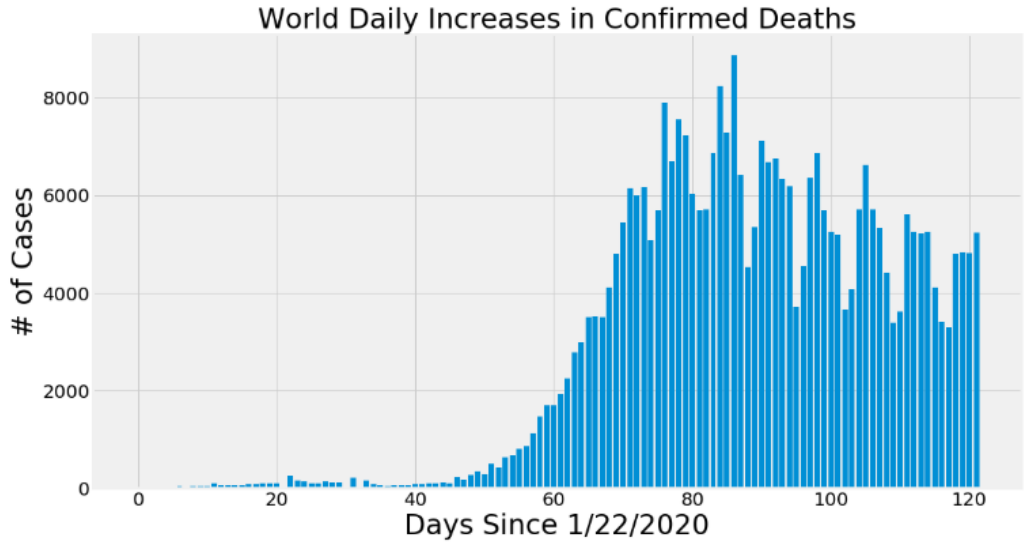
	Date	Bayesian Ridge Predicted # of Confirmed Cases Worldwide
0	05/23/2020	6281316.0
1	05/24/2020	6489825.0
2	05/25/2020	6703482.0
3	05/26/2020	6922372.0
4	05/27/2020	7146579.0
5	05/28/2020	7376189.0
6	05/29/2020	7611287.0
7	05/30/2020	7851961.0
8	05/31/2020	8098299.0
9	06/01/2020	8350387.0

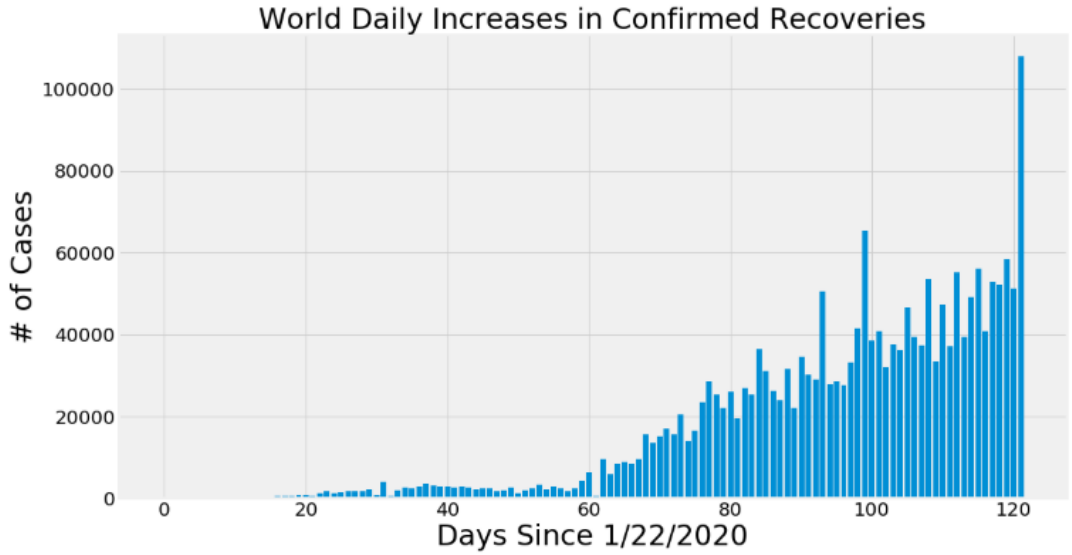


3.6 World Daily increase of those affected

The number of the affected people increasing each day can be plotted using the plot () function using the data which we created. This would provide an attractive and an easy visualisation in the trend of how fast or rapidly is the virus affecting people. We have plotted daily increase in the number of worldwide cases, deaths and recoveries. This graphical representation also helps us to determine the preventions to be taken by maintaining social distancing.



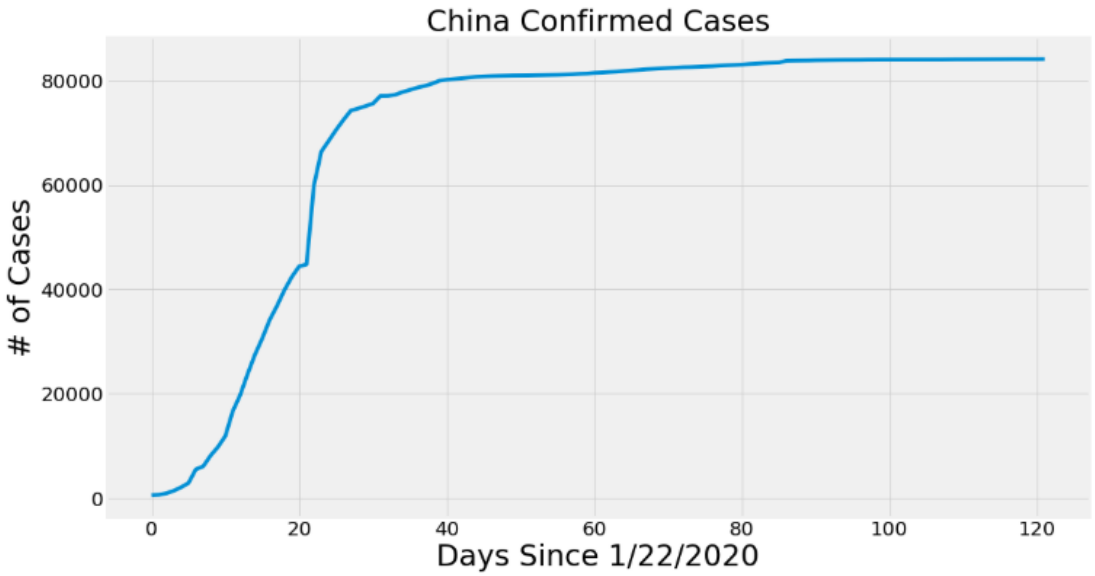


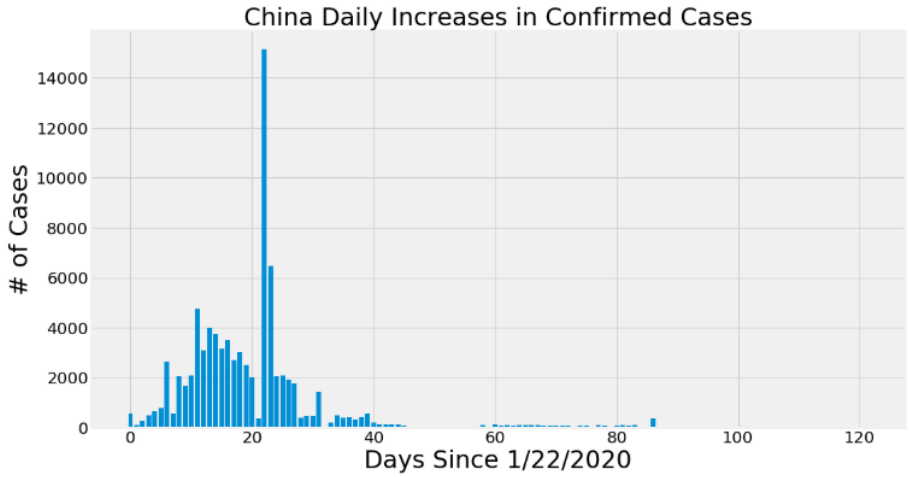


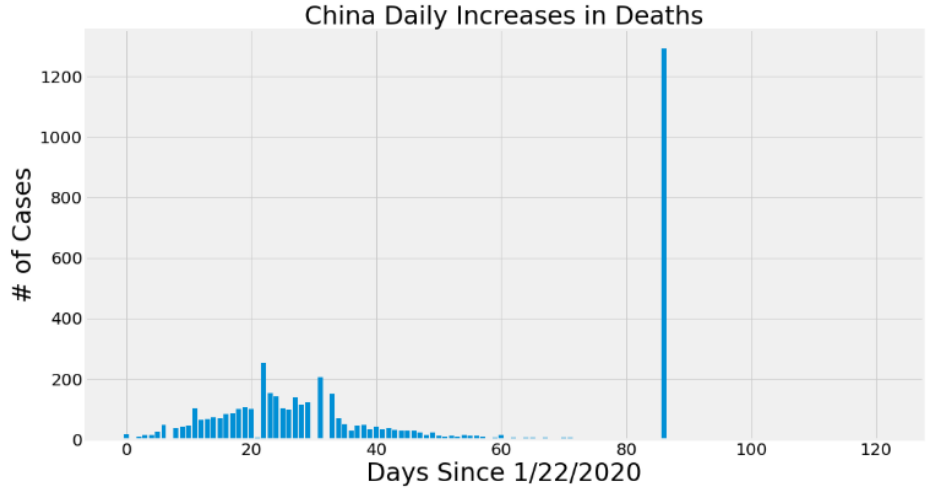
4. Case Studies of countries

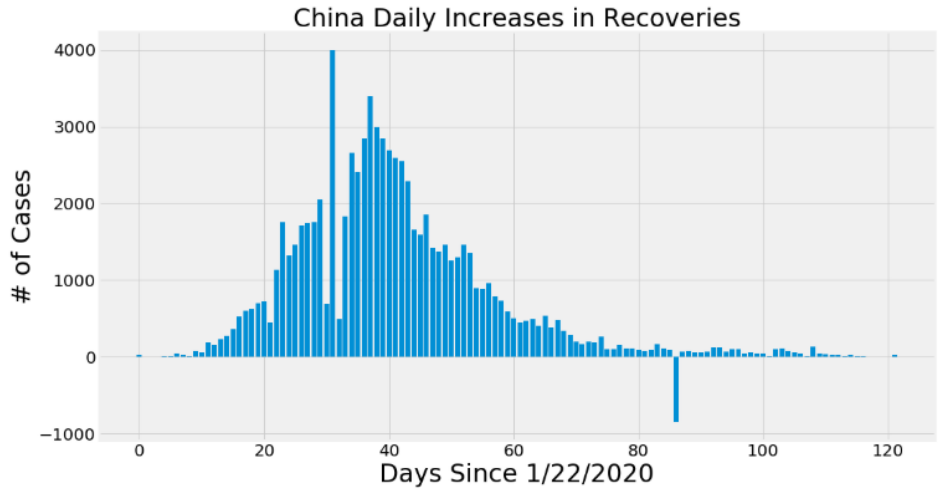
Country wise data is filtered from the data source so as to analyse the data on a grpahical bases.We study the effect of the virus in countries like China, United-States, Italy and India.China is the main epicenter from where virus spread across the entire world, Italy being the worst hit country from this virus and United states and India being majorly affected from the same .We have plotted confirmed cases, increase in confirmed, recoveries and deaths.

1. China:

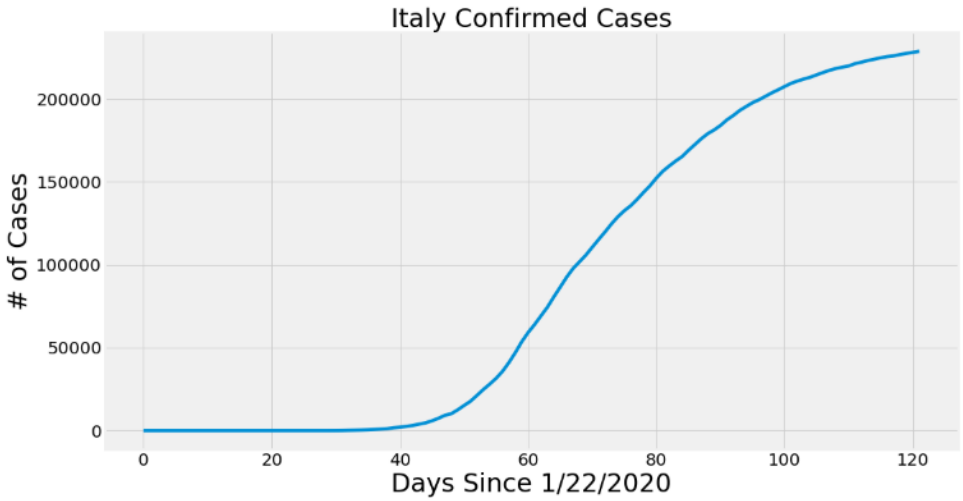


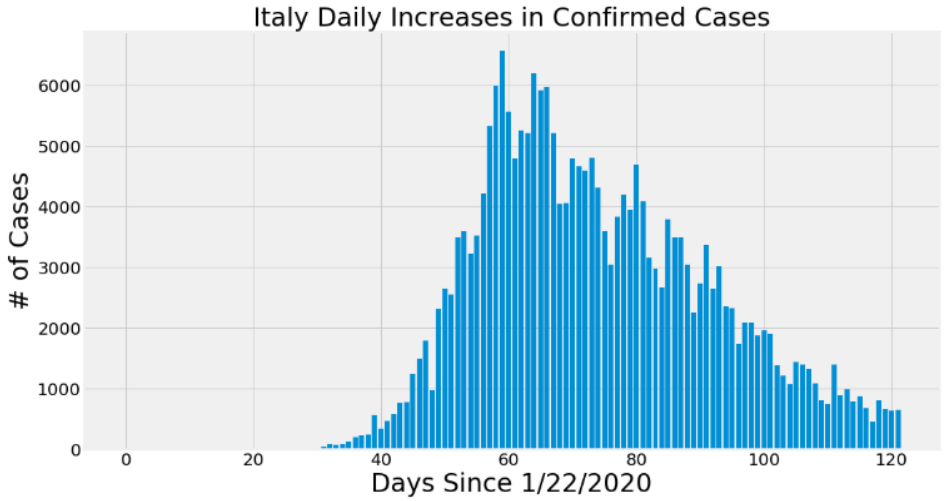

China Daily Increases in Confirmed Cases
# of Cases
Days Since 1/22/2020

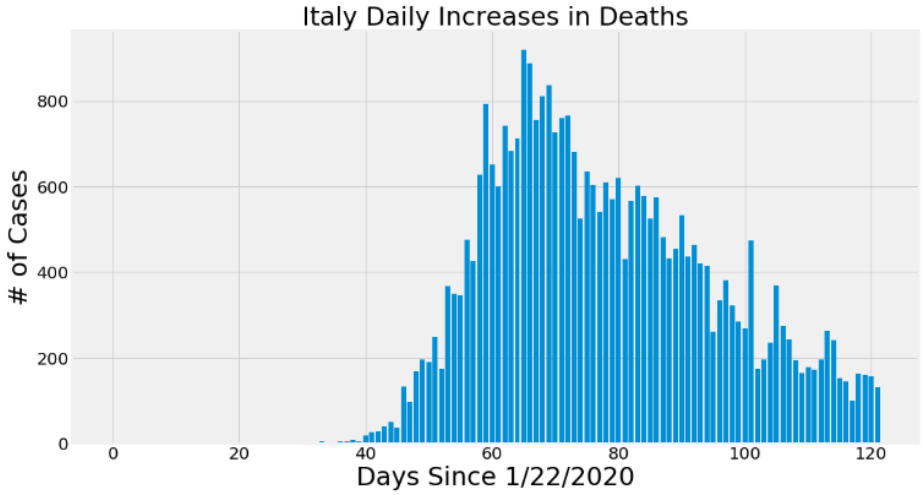

China Daily Increases in Deaths
# of Cases
Days Since 1/22/2020

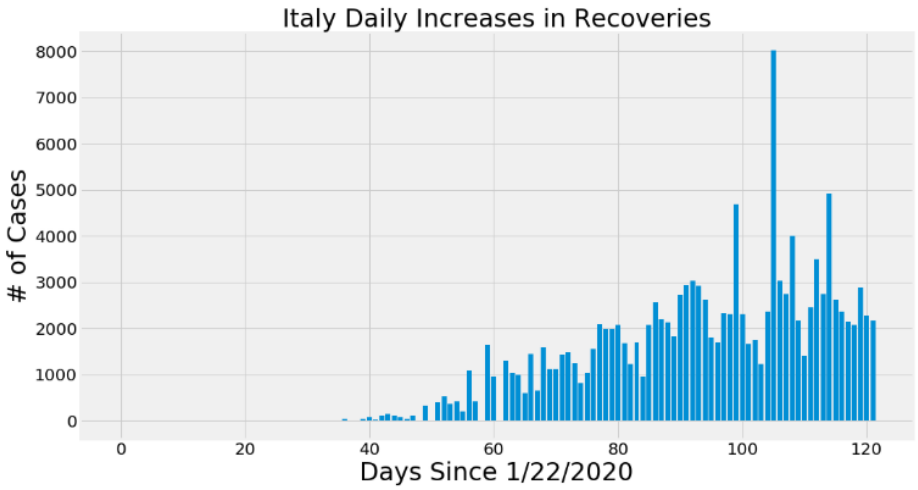

China Daily Increases in Recoveries
# of Cases
Days Since 1/22/2020

2. *Italy:*

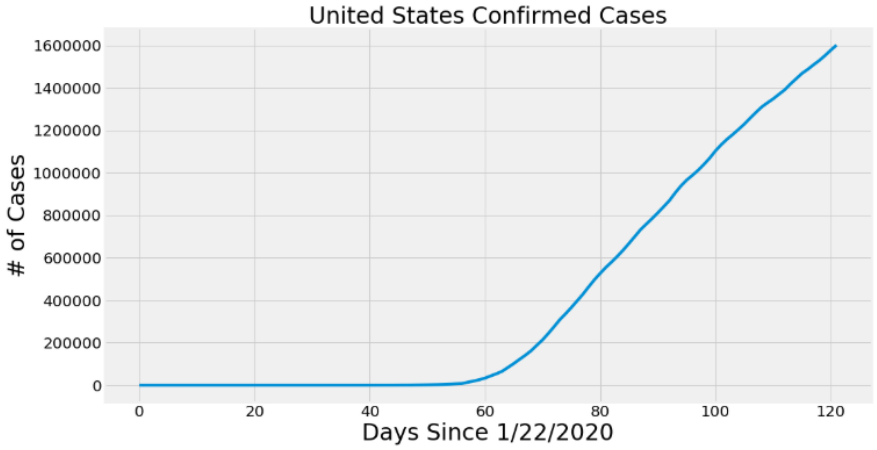

Italy Confirmed Cases
# of Cases
0
50000
100000
150000
200000
0
20
40
60
80
100
120
Days Since 1/22/2020

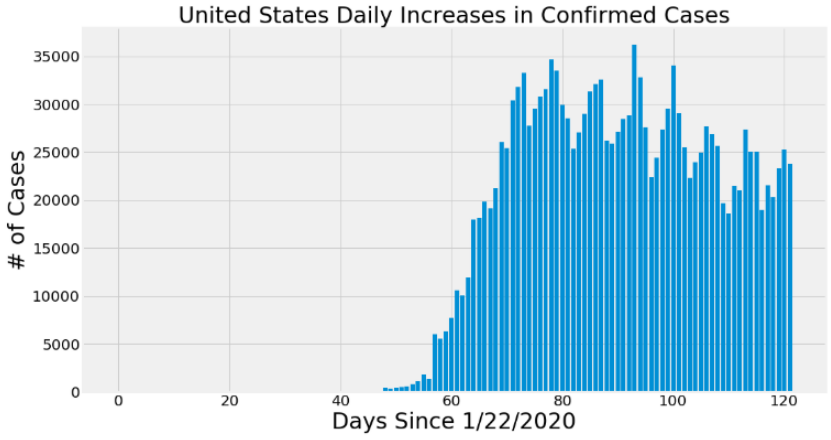

Italy Daily Increases in Confirmed Cases
# of Cases
0
1000
2000
3000
4000
5000
6000
0
20
40
60
80
100
120
Days Since 1/22/2020

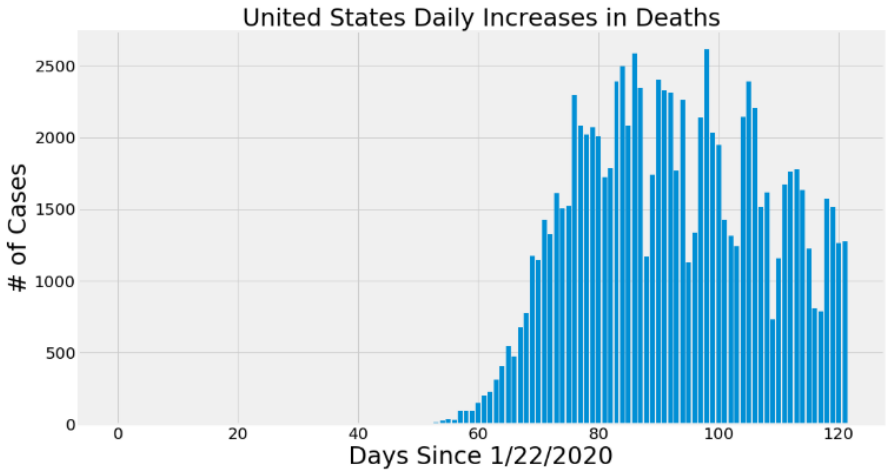

Italy Daily Increases in Deaths
# of Cases
0
200
400
600
800
0
20
40
60
80
100
120
Days Since 1/22/2020

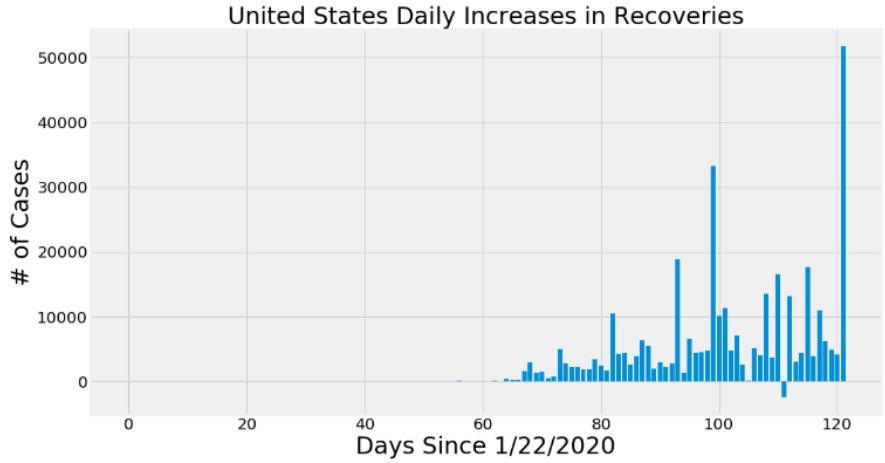


3. United States

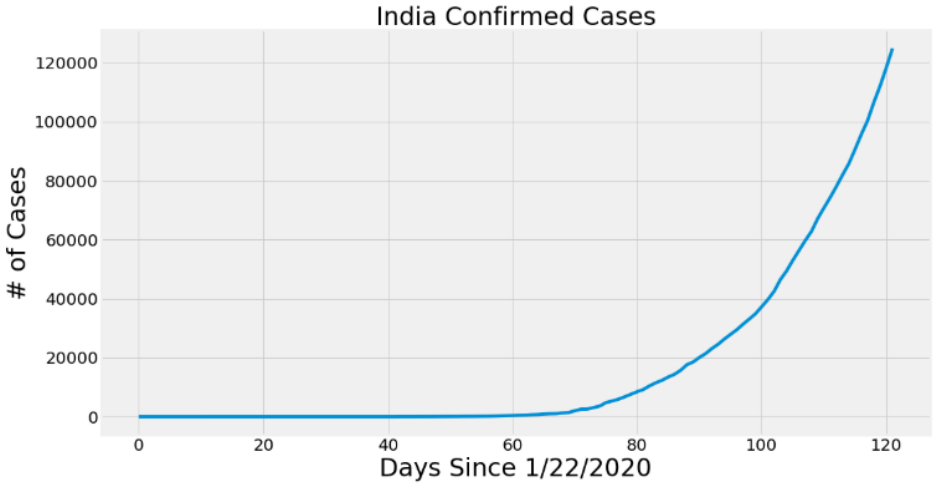


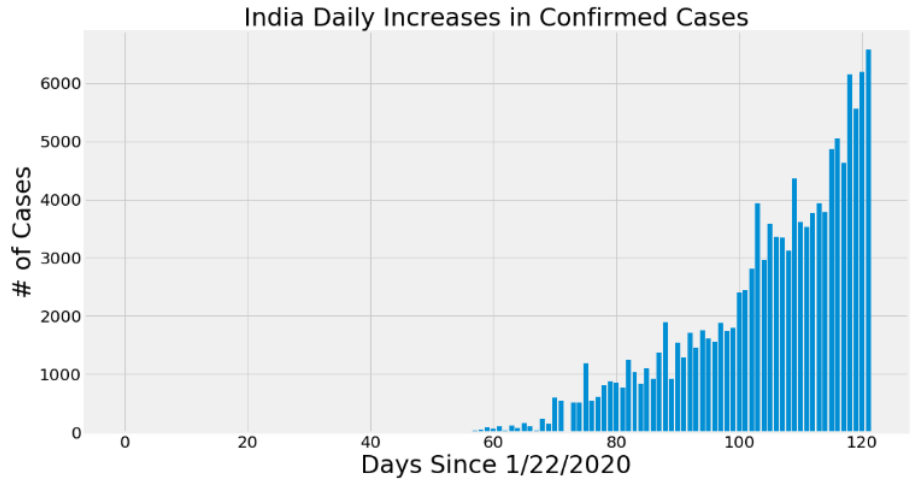


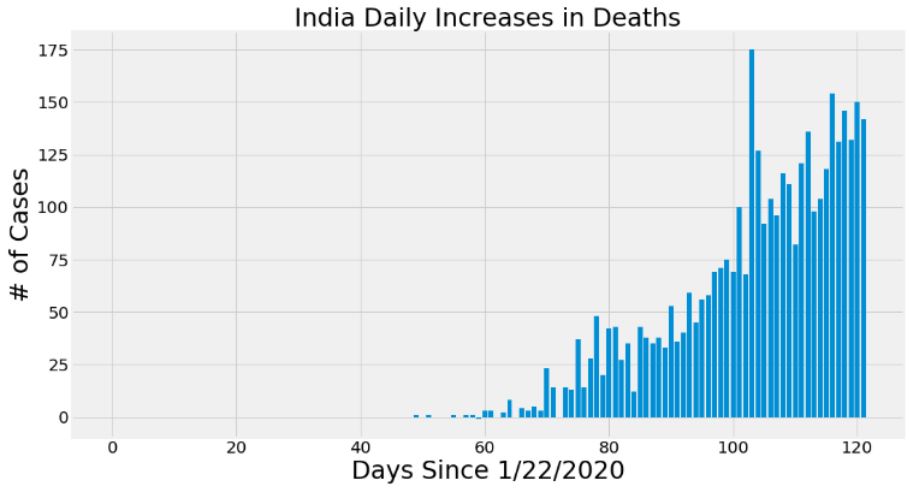

United States Daily Increases in Deaths
# of Cases
Days Since 1/22/2020

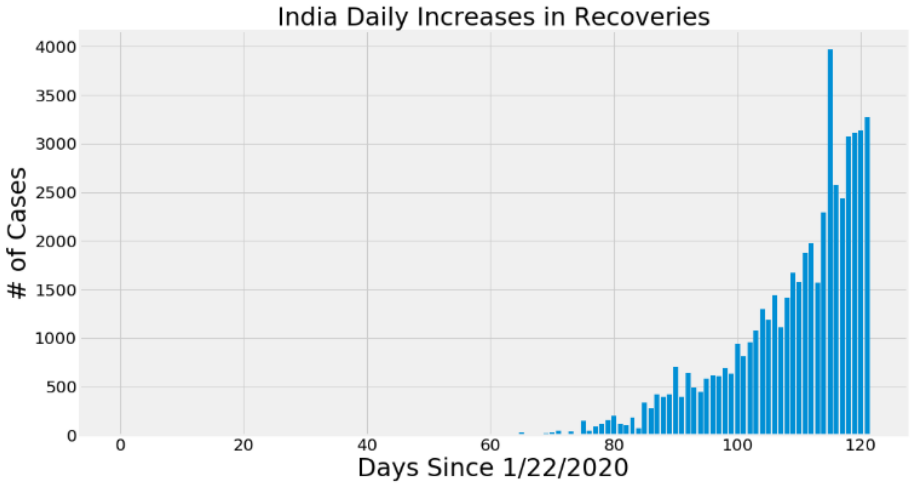

United States Daily Increases in Recoveries
# of Cases
Days Since 1/22/2020

4. India


India Confirmed Cases
# of Cases
Days Since 1/22/2020


India Daily Increases in Confirmed Cases
# of Cases
0
1000
2000
3000
4000
5000
6000
0
20
40
60
80
100
120
Days Since 1/22/2020


India Daily Increases in Deaths
# of Cases
0
25
50
75
100
125
150
175
0
20
40
60
80
100
120
Days Since 1/22/2020


India Daily Increases in Recoveries
# of Cases
0
500
1000
1500
2000
2500
3000
3500
4000
0
20
40
60
80
100
120
Days Since 1/22/2020

Most of the graphs as seen imply some common characterstics:

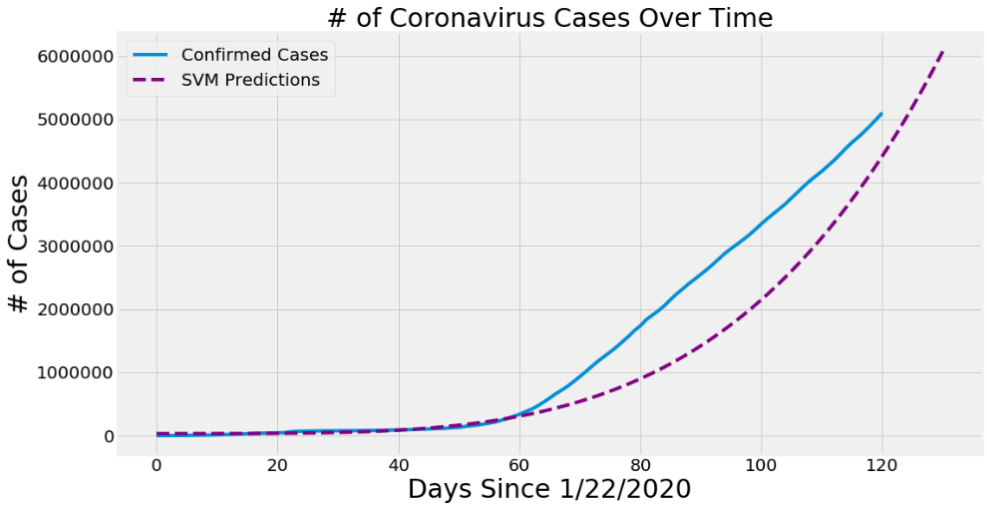
1. Increasing trend in the detection, deaths and recovered cases
2. An exponential growth in the number of cases within 3-4 days of community spread

5. Prediction of the machine learning model on real time data

Our machine learning model is now tested against real time data and is able to predict the cases with an accuracy of 83.2%.

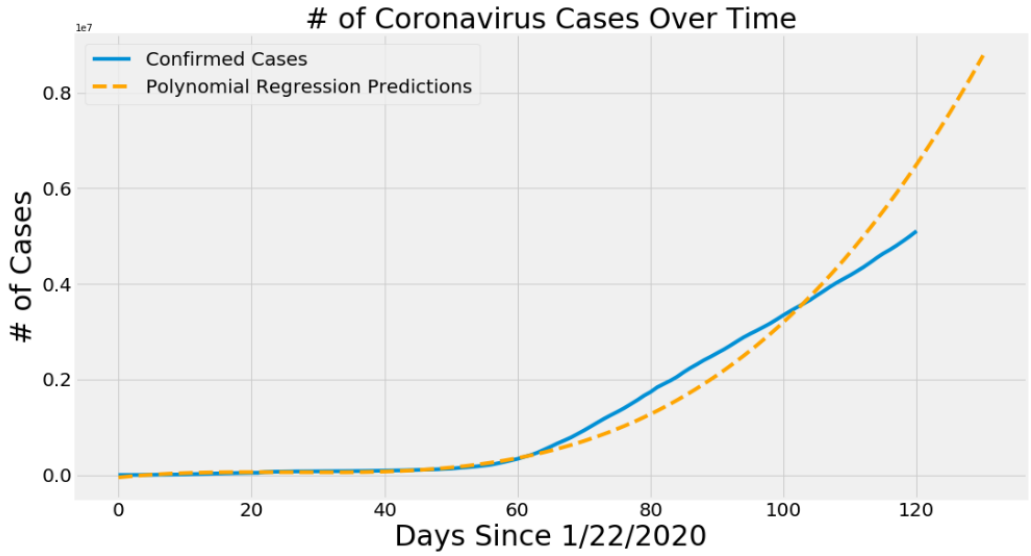
Using

1. SVM:



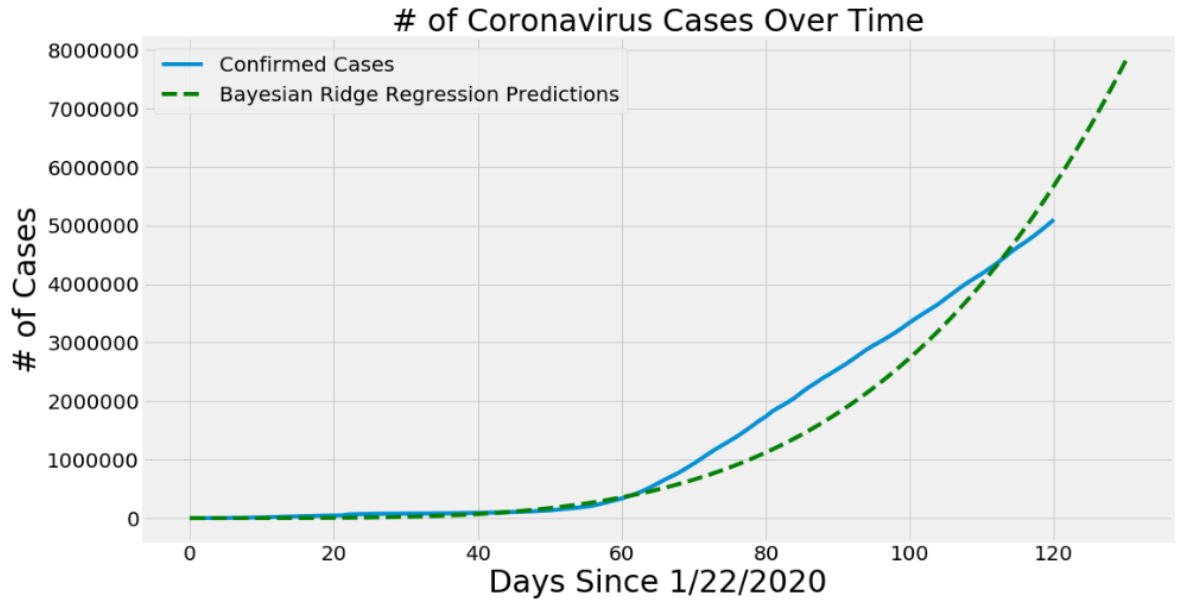
	Date	SVM Predicted # of Confirmed Cases Worldwide
0	05/22/2020	4565117.0
1	05/23/2020	4716807.0
2	05/24/2020	4872272.0
3	05/25/2020	5031576.0
4	05/26/2020	5194782.0
5	05/27/2020	5361951.0
6	05/28/2020	5533149.0
7	05/29/2020	5708439.0
8	05/30/2020	5887885.0
9	05/31/2020	6071554.0

2. *Polynomial Regression*



	Date	Polynomial Predicted # of Confirmed Cases Worldwide
0	05/22/2020	6706804.0
1	05/23/2020	6918813.0
2	05/24/2020	7135266.0
3	05/25/2020	7356210.0
4	05/26/2020	7581690.0
5	05/27/2020	7811753.0
6	05/28/2020	8046444.0
7	05/29/2020	8285809.0
8	05/30/2020	8529895.0
9	05/31/2020	8778747.0

3. Bayesian Linear Regression (Ridge)



	Date	Bayesian Ridge Predicted # of Confirmed Cases Worldwide
0	05/22/2020	5871432.0
1	05/23/2020	6067931.0
2	05/24/2020	6269321.0
3	05/25/2020	6475684.0
4	05/26/2020	6687099.0
5	05/27/2020	6903649.0
6	05/28/2020	7125417.0
7	05/29/2020	7352485.0
8	05/30/2020	7584938.0
9	05/31/2020	7822859.0

6. Further Opportunities and Areas of research

Machine learning has made it possible to study and predict the future spread of the disease with a decent precision and accuracy. Apart from the modelled proposed, there are other areas, where research could be carried out and other parameters could be taken into consideration to increase the accuracy of the model.

1. Incorporation of other parameters: such as density of people in a country, climatic conditions, wind speed, and also taking into account the travel history and area of contact of the people infected by the virus.
2. Understanding the genome sequence of the coronavirus may be helpful in coming up with compounds and elements which could be used to create a curable vaccine to fight the disease
3. The content specific section of various blogs, articles and news reports presented in the social media
4. Understanding the effect of this pandemic on government plans, allocation of funds, how it has affected industries and its implication on all the sections of the society

7. Conclusion

In this report, we have seen how machine learning, data pre-processing and intrinsic mathematical functions can be used to forecast the growth of the pandemic in the coming months proactively. There is also an understanding that, for proper tuning of the model, and to get optimal results many algorithms must be tested and the best one should be selected.

We could also have plotted graphs of deaths vs recoveries, and mortality_rate and recovery_rate of the same.

8. REFERENCES

Most of the data has been taken from the GitHub repository of John-Hopkins-University

Link:https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data

The Indian Health Ministry website

Link: https://www.mohfw.gov.in/