Analysis and Prediction of Covid-19

Abrar Athar Hashmi and Abdul Wahed

Department of Computer Science and Engineering,

CBIT, Hyderabad, India

hashmiabrar14@gmail.com abdul.wahed11314@gmail.com

Abstract. Coronavirus is a family of viruses that can cause illness, which can vary from common cold and cough to sometimes more severe disease. Middle East Respiratory Syndrome (MERS-CoV) and Severe Acute Respiratory Syndrome (SARS-CoV) were such severe cases the world already has faced. SARS-CoV-2 is the new virus of the coronavirus family, which was first discovered in 2019, which has not been identified in humans before. It is a contiguous virus which started from Wuhan in December 2019 and was later declared as Pandemic by WHO due to a drastic spread throughout the world. Currently, this leads to a total of 1 million+ Deaths across the globe. Pandemic is spreading all over the world; it has become essential to understand its spread. At the same time, COVID-19 has spread in more than 179 countries. There is no doubt that this epidemic has become a disaster for all mankind. No country, no one can survive alone. This paper is an effort to analyze the cumulative data of confirmed, deaths, and recovered cases over time. Intuitive visualizations and inferences are made. Later a time series forecasting is applied to predict confirmed cases in India.

Keywords: Covid-19, ARIMA, K-Means Clustering

1 Introduction

The 200 countries span a diverse range of social, economic, health, and weather conditions. Because of the COVID19 pandemic, all of these countries have already experienced some COVID19 cases. What's important is to develop data mining tools that can help the medical community find answers to high priority scientific questions. The COVID-19 dataset represents the most detailed machine-readable coronavirus collection available for data mining to date. It has allowed the worldwide AI research community an opportunity to find answers, and share insights across this content in support of the ongoing COVID-19 response efforts worldwide. There is a huge urgency for these approaches because of the rapid increase in coronavirus, making it difficult for the medical community to keep up.

Every Pandemic has three stages[1]:

Stage 1: Beginning of Local Transmission

Stage 2: Countries impacted with local transmission

Stage 3: Significant Transmission across the World

2 Related Work

This paper is an effort to analyze the cumulative data of confirmed, deaths, and recovered cases over time. In this paper, the main focus is to analyze the spreading trend of this virus all over the world.

The Dataset consists of time-series data from 22 JAN 2020 to Till date (Updated on daily Basis)[2]

Table 1. List of parameters analyzed from the World Covid-19 Data

Province/State
Country/Region
Date
Confirmed
Deaths
Recovered
Active

Dataset consists of time-series data from 22 JAN 2020 to Till date (Updated on daily Basis).

It consists of the following data-

Table 2. Sample Data

	Province/State	Country/Region	Date	Confirmed	Deaths	Recovered	Active
129		India	2020-01-22	0	0	0	0
383		India	2020-01-23	0	0	0	0
637		India	2020-01-24	0	0	0	0
891		India	2020-01-25	0	0	0	0
1145		India	2020-01-26	0	0	0	0
...	...	...	...	...	...	...	...
72265		India	2020-11-01	8229313	122607	7544798	561908
72519		India	2020-11-02	8267623	123097	7603121	541405
72773		India	2020-11-03	8313876	123611	7656478	533787
73027		India	2020-11-04	8364086	124315	7711809	527962
73281		India	2020-11-05	8411724	124985	7765966	520773

289 rows × 7 columns

3 Proposed System

There are three steps:

1. Preprocessing of Data

2. Visualization of Data

3. Training and Testing

3.1. Preprocessing of Data

3.1.1 Data Cleaning

Noisy, irrelevant data is identified and removed. We also understand through visualization which factors are more important.

3.1.2 Identifying Missing Values

There can be missing values. We will investigate each column with total missing values. We will not be replacing it with the mean or median value since the dataset is big enough to perform analysis.

3.2 Visualization Data

Various technologies are used to obtain inferences from the data through intuitive patterns.

3.2.1 Pandas

It is among the fundamental high-level building blocks for doing practical, real world data analysis in Python[3].

Pandas is well suited for many different kinds of data:

- Tabular data with heterogeneously-data types, as in an SQL table or Excel spreadsheet
- Ordered and unordered time series data.
- Random matrix data (homogeneously typed or heterogeneous) with row and column labels.
- Any other form of observational / statistical data sets. The data doesn't need not be labeled to be placed into a pandas data structure

3.2.2 Matplotlib

Matplotlib is a plotting package that provides MATLAB-like plotting functionality. Matplotlib utilities lie under the pyplot submodule, and are usually imported under the plt. Matplotlib is designed to be as usable as MATLAB, with the ability to use Python

3.2.3 Plotly

Plotly Express is a new high-level Python visualization library which is an interactive, open-source plotting library that exposes a simple syntax for complex charts. You can make richly interactive plots in just a single function call, including faceting, maps, animations that can be displayed in Jupyter notebooks

3.2.4 K-Means Clustering

K-Means Clustering with Elbow Method- A fundamental step for any unsupervised algorithm is to determine the optimal number of clusters into which the data may be clustered. The is one of the most popular methods to determine this optimal value of k[4].

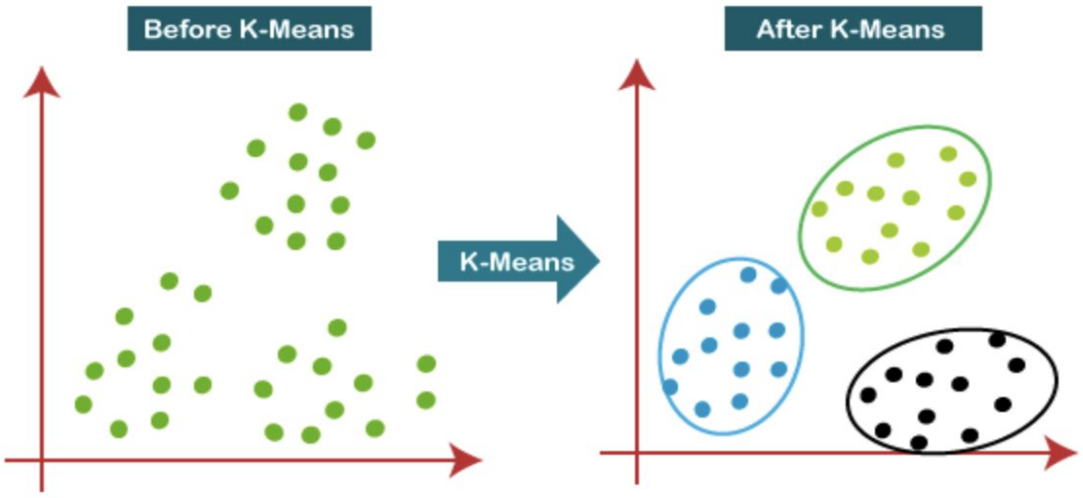

Fig.1. K-Means Clustering Explained

3.3 Training and Testing

3.3.1 Validation

Machine learning, especially supervised learning techniques such as classification and regression require training data to build a model. Training data consists of labelled data, i.e. datasets that are complete with the target value together with input feature vectors. A good classification or regression model can be built if a significant amount of training data is supplied during the training process. This is followed by the validation process where test data is fed into the trained model to evaluate its predictive accuracy. It is important to test the model properly with enough test data so that the model would yield accurate predictions in the production environment.

Unfortunately, scarcity of data often prompts machine learning practitioners to split the dataset in hand into two subsets, namely training data and test data. These subsets emerge from splitting the original dataset according to a certain ratio such as 80:20 or 60:40, with the bigger proportion making up the training data subset. Training and validating a model using a single train-test split (a.k.a. holdout method) would not yield significant predictive accuracy due to bias. Bias in this case means that in a single train-test split, data points could be clustered in such a way that one cluster gets stuck in the training set and another cluster gets stuck in the test set. Such a situation leads to bias in the train-test split, thus adversely affecting the predictive accuracy of a model[7].

3.3.2 ARIMA Model

ARIMA is an acronym which stands for AutoRegressive Integrated Moving Average. It is a class of models that explains a given time series based on its own past values. It is a class of statistical models for analyzing and forecasting time series data[5].

An Auto Regressive (AR only) model is one where Yt depends only on its own lags. A nonseasonal ARIMA model is classified as an "ARIMA(p,d,q)" model, where[6]:

- p is the number of autoregressive terms
- d is the number of nonseasonal differences needed for stationarity
- q is the number of lagged forecast errors in the prediction equation

If d=0: $y_t = Y_t$ (1)

If d=1: $y_t = Y_t - Y_{t-1}$ (2)

If d=2: $y_t = (Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2}) = Y_t - 2Y_{t-1} + Y_{t-2}$ (3)

A pure **Auto Regressive (AR only)** model is one where Yt depends only on its own lags. That is, Yt is a function of the 'lags of Yt'.

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + .. + \beta_p Y_{t-p} + \epsilon_1$$

(4)

Likewise a pure **Moving Average (MA only) model** is one where Yt depends only on the lagged forecast errors.

$$Y_t = \alpha + \epsilon_t + \phi_1 \epsilon_{t-1} + \phi_2 \epsilon_{t-2} + .. + \phi_q \epsilon_{t-q}$$

(5)

An ARIMA model is one where the time series is differenced at least once to make it stationary and you combine the AR and the MA terms. So the equation becomes:

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + .. + \beta_p Y_{t-p} \epsilon_t + \phi_1 \epsilon_{t-1} + \phi_2 \epsilon_{t-2} + .. + \phi_q \epsilon_{t-q}$$

(6)

3.3.3 Root Mean Squared Error

RMSE is a measure of how spread out the prediction errors are[10].

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

(7)

3.3.4 Mean Absolute Error

It is the absolute difference between the actual values and the values that are predicted[10].

$$\mathrm{MAE} = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

(8)

4 Results and Discussions

4.1 Data Visualization

4.1.1 Visualizing core trends in the World

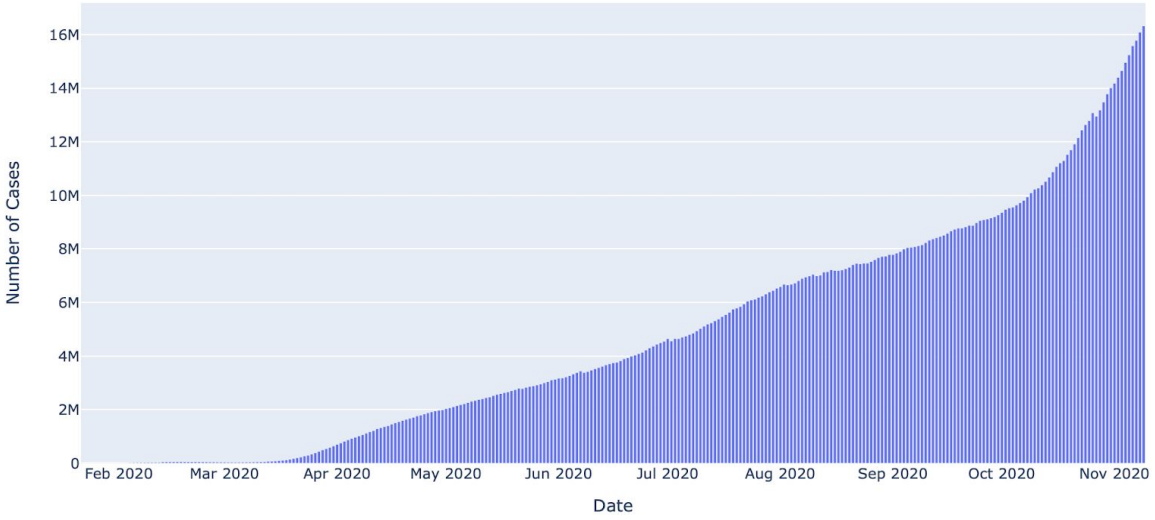

Fig.2. Number of Active Cases

Number of Active Cases = Number of Confirmed Cases - Number of Recovered Cases - Number of Death Cases

Increase in the number of Active Cases is probably an indication that Recovered or Death cases are dropping in comparison to the number of Confirmed Cases drastically. We will look for the conclusive evidence for the same ahead.

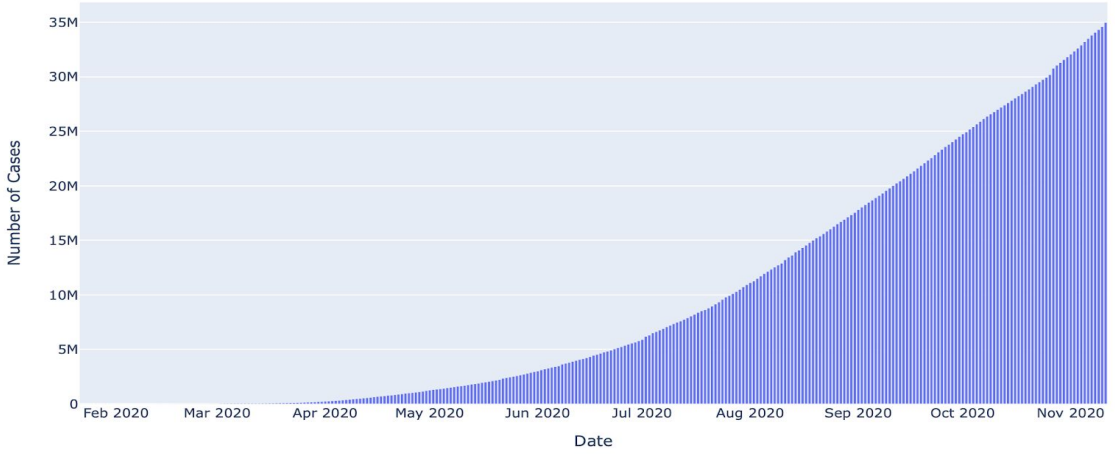

Fig.3. Number of Closed Cases

Number of Closed Cases = Number of Recovered Cases + Number of Death Cases

Increase in number of Closed classes implies that either more patients are getting recovered from the disease or more people are dying because of COVID-19

Weekly Growth-

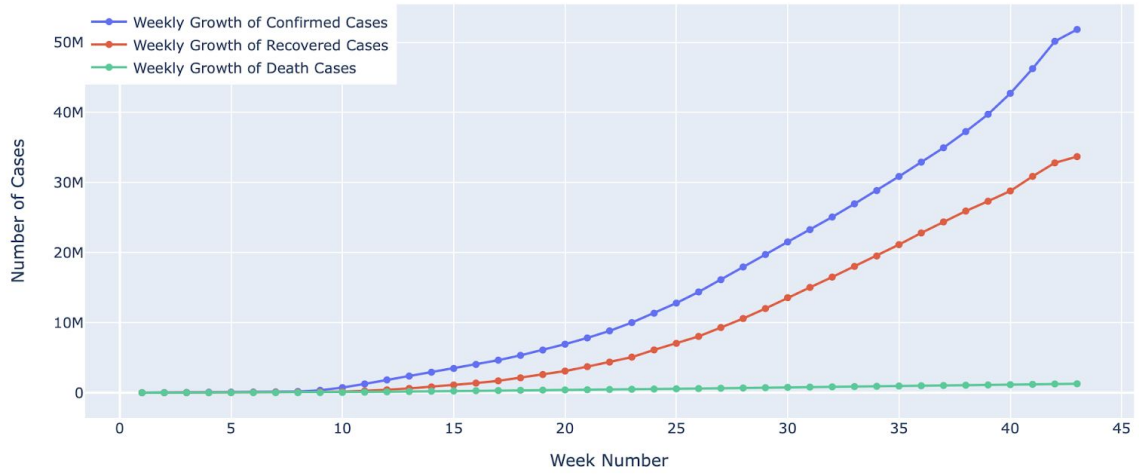

Fig.4. Weekly growth of different types of Cases in India

Number of Confirmed Cases = Active Cases + Number of Recovered Cases + Number of Death Cases

The death toll was low until the 12th week. Number of Death cases were consistently dropping from 14th week, upto 19th week. After which it's again showing a spike. We might somehow be able to reduce the Death Numbers or maybe control it, but new infections are increasing with considerable speed recording 1 million+ cases since the 20th week which is a huge number.

43rd Week has recorded yet another peak in the number of Confirmed Cases (50Million+) which means that the infection rate is increasing with every passing week[8].

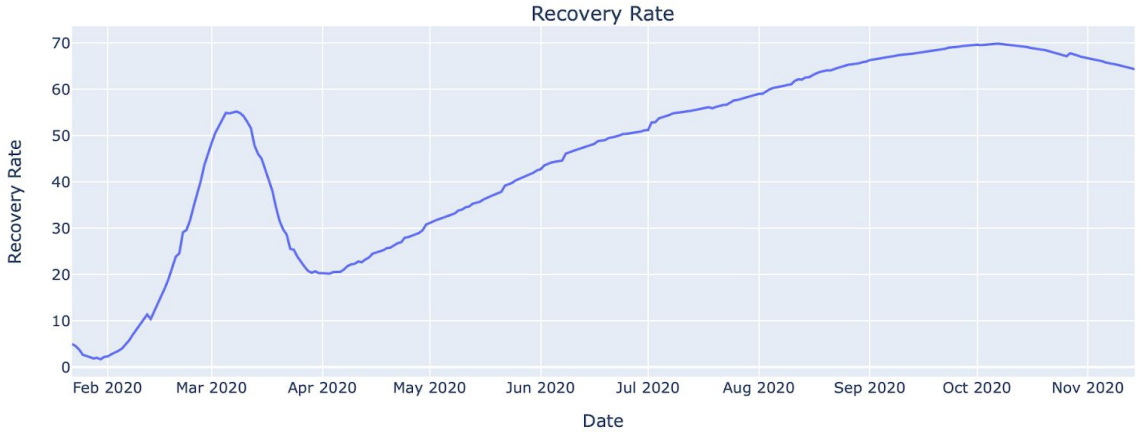

Fig.5. Recovery Rate

Recovery Rate %= (Number of Recovered Cases / Number of Confirmed Cases) x 100

```
Average Recovery Rate 46.42208098320608
```

Recovery Rate has started to pick up again which is a good sign, another supportive reason to why the number of Closed Cases are increasing.

Fig.6. Mortality Rate

Mortality rate = (Number of Death Cases / Number of Confirmed Cases) x 100

```
Average Mortality Rate is 4.211506436734152
```

Mortality rate which climbed up to a record high of 7 in May has shown a considerable decline for a pretty long time, which is a positive sign.

4.1.2 Growth Factor for Active and Closed Cases

Growth factor is the factor by which a quantity multiplies itself over time.

The formula used is:

Formula: Every day's new (Active and Closed Cases) / new (Active and Closed Cases) on the previous day.

(i) A growth factor above 1 indicates an increase in corresponding cases. A growth factor of above 1 but trending downwards is a positive sign.

(ii) A growth factor constant at 1 indicates there is no change in the trend.

(iii) A growth factor of below 1 indicates a really positive sign implying more patients are getting recovered or dying.

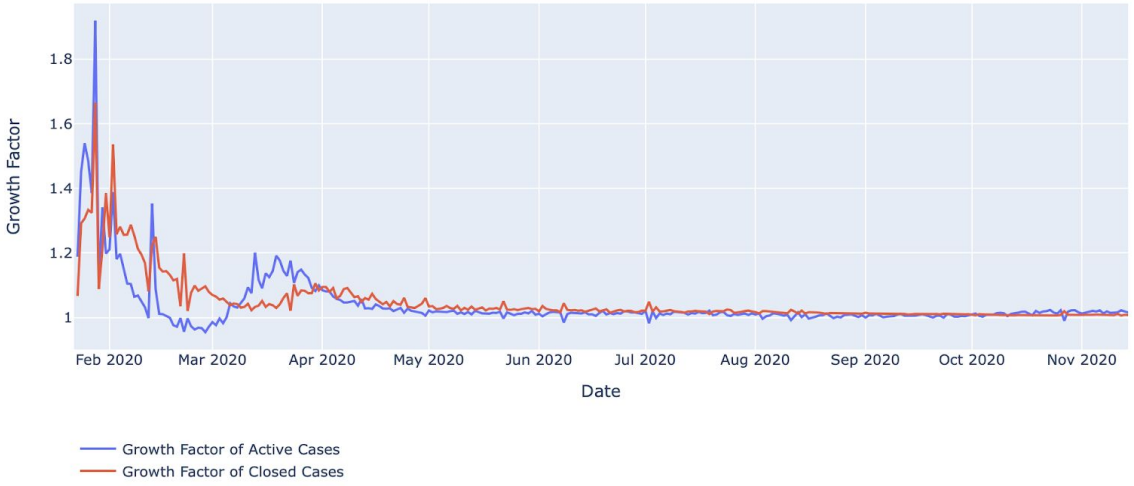

Fig.7. Date Wise Growth Factor of Active and Closed Cases

The growth Factor is always above 1 is a clear indication of Exponential increase in all forms of cases.

4.1.3 Country Wise Analysis

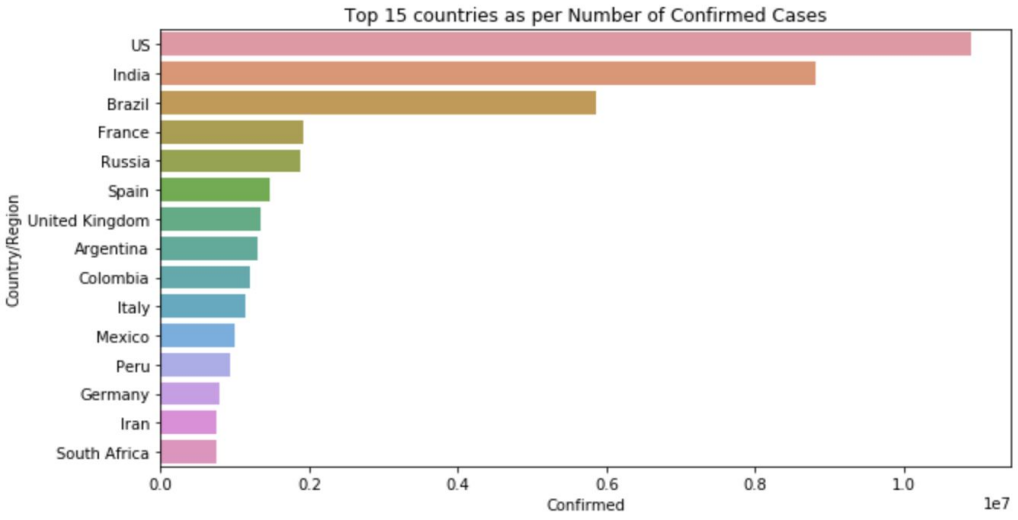

Fig.8. Country-wise Confirmed Cases

The US has the highest number of Confirmed Cases followed by India.

Now let's look at countries without any recoveries having a high mortality rate.

Country/Region	Confirmed	Deaths	Recovered	Mortality Rate
MS Zaandam	9	2	0	22.2222
Sweden	177355	6164	0	3.47552
Belgium	531280	14303	0	2.69218
Serbia	81086	989	0	1.21969

Fig.9. Countries with Zero Recoveries

Sweden currently has the maximum number of Confirmed Cases, with no Recovered patient being recorded. It also has a high mortality rate compared to the overall mortality rate of the World(4.2).

Lets look at Countries with more than 100 Confirmed Cases, no Deaths with considerably high Recovery Rate.

Country/Region	Confirmed	Recovered	Deaths	Recovery
Seychelles	160	157	0	98.125
Cambodia	302	289	0	95.6954
Bhutan	375	353	0	94.1333
Eritrea	493	444	0	90.0609
Mongolia	428	328	0	76.6355

Fig.10 Countries with Zero Deaths

Here we find an excellent scenario where these countries such as Cambodia, Bhutan etc have the highest Recovery rates without any deaths.

4.1.4 Clustering

I will be clustering different countries based on the Mortality and Recovery rate of individual countries. This is because COVID-19 has a different Mortality Rate in different countries based on different factors , similar is the case with Recovery Rate because of different pandemic controlling practices followed by the individual country. Also Mortality Rate and Recovery Rate both together takes into account all types of cases Confirmed, Recovered and Deaths.

K-Means clustering with the Elbow method is applied. We obtain K=3 to be the optimal number of clusters.

Country/Region	Confirmed	Recovered	Deaths	Mortality	Recovery	Active Cases	Outcome Cases	Clusters
US	10903890.00	4148444.00	245598.00	2.25	38.05	6509848.00	4394042.00	1.00
France	1915713.00	139760.00	42600.00	2.22	7.30	1733353.00	182360.00	1.00
Spain	1458591.00	150376.00	40769.00	2.80	10.31	1267446.00	191145.00	1.00
United Kingdom	1347907.00	3108.00	51858.00	3.85	0.23	1292941.00	54966.00	1.00
Italy	1144552.00	411434.00	44683.00	3.90	35.95	688435.00	456117.00	1.00
Yemen	2072.00	1394.00	605.00	29.20	67.28	73.00	1999.00	2.00
MS Zaandam	9.00	0.00	2.00	22.22	0.00	7.00	2.00	2.00
India	8814579.00	8205728.00	129635.00	1.47	93.09	479216.00	8335363.00	0.00
Brazil	5848959.00	5279452.00	165658.00	2.83	90.26	403849.00	5445110.00	0.00
Russia	1887836.00	1415213.00	32536.00	1.72	74.96	440087.00	1447749.00	0.00
Argentina	1304846.00	1119366.00	35307.00	2.71	85.79	150173.00	1154673.00	0.00
Colombia	1191634.00	1097576.00	33829.00	2.84	92.11	60229.00	1131405.00	0.00

Fig.11. Clustering of Countries for K=3

```
Avergae Mortality Rate of Cluster 0:  2.028899643460227
Avergae Recovery Rate of Cluster 0:  85.69068746546172
Avergae Mortality Rate of Cluster 1:  1.8267819090831703
Avergae Recovery Rate of Cluster 1:  35.25814272384334
Avergae Mortality Rate of Cluster 2:  25.71053196053196
Avergae Recovery Rate of Cluster 2:  33.63899613899614
```

Cluster 0 is a set of countries which have a really Low Mortality Rate and really High Recovery Rate. These are the set of countries who have been able to control the ill effects of COVID-19 by following pandemic controlling practices rigorously. Eg- India, Brazil, Russia and Argentina.

Cluster 1 is a set of countries which have a Low Mortality Rate and Low Recovery Rate. These countries need to pace up their Recovery Rate. Some of these countries have a really high number of infected cases but Low Mortality which is a positive sign. Eg- US, France, Spain and UK.

Cluster 2 is a set of countries which have a really High Mortality Rate and low Recovery Rate. These countries have already seen the worst of this pandemic but are now recovering with a healthy Recovery Rate. eg- Yemen and MS Zaandam.

4.1.5 Training and Testing

Here Time Series Forecasting is applied for the number of confirmed cases in India. Data is divided into training and validation data. The model is trained with the number of confirmed cases since the epidemic began until mid november and tested with data until the 1st week of December.

```
Fit ARIMA: order=(2, 2, 1); AIC=5841.038, BIC=5859.590, Fit time=0.073 seconds
Fit ARIMA: order=(2, 2, 2); AIC=5793.287, BIC=5815.550, Fit time=0.424 seconds
Fit ARIMA: order=(2, 2, 3); AIC=nan, BIC=nan, Fit time=nan seconds
Fit ARIMA: order=(3, 2, 1); AIC=5843.028, BIC=5865.290, Fit time=0.143 seconds
Fit ARIMA: order=(3, 2, 2); AIC=5844.640, BIC=5870.613, Fit time=0.262 seconds
Fit ARIMA: order=(3, 2, 3); AIC=5811.883, BIC=5841.567, Fit time=0.537 seconds
Total fit time: 2.263 seconds
```

Here the ARIMA model is fit with training data and validated.

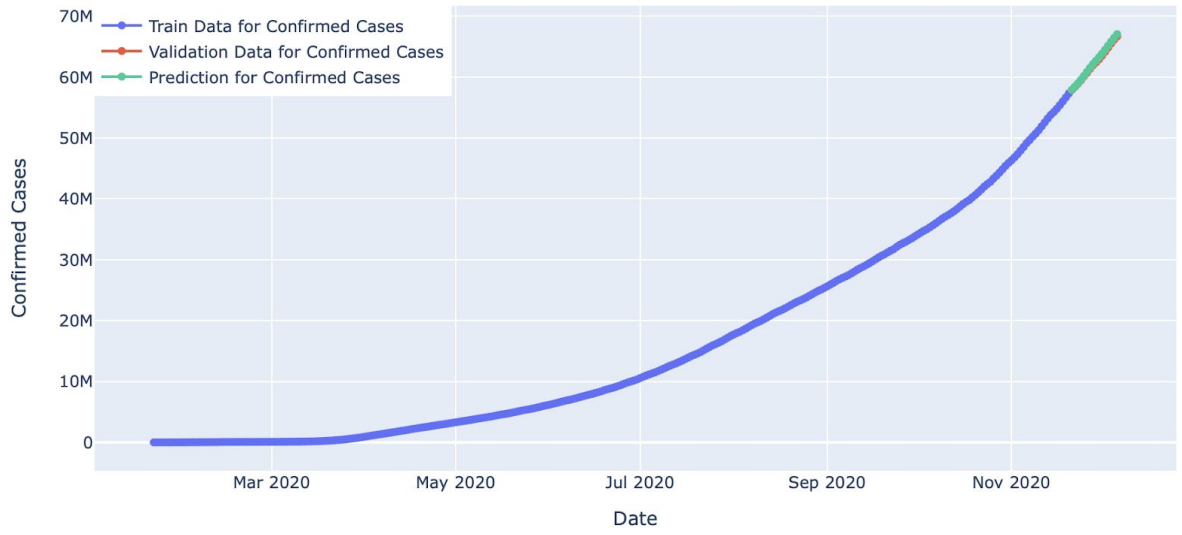

Fig.12. Confirmed cases ARIMA Model Prediction

The model is able to predict the number of confirmed cases accurately. Prediction is very close to actual values.

Table 3: Error analyzed for proposed and existing model

Model	Root Mean Squared Error	Mean Absolute Error (MAE)	Training dataset	Ratio(MAE per data point)
Arima Model Proposed	81637.26956868572	65828.26535279735	9,677,203	0.008
Arima model from Paper[9]	50.83	16.07	64	0.251

We have used root mean square error (RMSE), mean absolute error (MAE), to evaluate the predictive performance of the model used in this study. The paper [9] predicts the confirmed cases for India containing 64 observations and gets an error per data point of 0.251 for which there is a considerable improvement with 0.008 for the proposed Arima model.

Table 4. ARIMA Model Forecast Data for future Dates

	Deaths	**ARIMA Model Death Forecast**
1	2020-12-07	9884192.658758
2	2020-12-08	9935270.814755
3	2020-12-09	9987774.014331
4	2020-12-10	10036930.247790

5	2020-12-11	10086532.878144
6	2020-12-12	10139623.689411
7	2020-12-13	10191097.745191

5 Conclusions

COVID-19 doesn't have a very high mortality rate which is continuously dropping. The recovery rate is constantly rising which is a very positive take away. The only matter of concern is the exponential growth rate of infection due to which the number of Confirmed cases is rapidly increasing.

Countries like the USA, India, Spain, Brazil and the United Kingdom are facing some serious troubles in containing the disease. It's really important to perform COVID-19 pandemic controlling practices like Testing, Contact Tracing and Quarantining efficiently with a speed greater than the speed of disease spread at each country level. The growth of Confirmed and Death Cases seems to have slowed down since the past few days which is a really good sign. The USA, India and Brazil seem to be the three hotspots of Covid cases and hopefully there shouldn't emerge any other.

The time forecasting model indicates an considerable increase in the number of confirmed cases in India with over a combined 10 million cases in the next few days.

References

1.The Weather Channel; https://weather.com/en-IN/india/coronavirus/news/2020-04-09-four-stages-of-virus-transmission-stage-india-currently-finds

2. CSSEGISandData: GITHUB: https://github.com/CSSEGISandData/COVID-19

3.Pandas: https://pandas.pydata.org/docs/getting_started/overview.html

4.Gupta,A.:GeeksforGeeks: https://www.geeksforgeeks.org/elbow-method-for-optimal-value-of-k-in-kmeans/

5. Sangarshanam: Time series Forecasting — ARIMA models: https://towardsdatascience.com/time-series-forecasting-arima-models-7f221e9eee06

6. https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/

7. Hashmi, Abrar.A., Wahed, Abdul: Road Accident Prediction and Classification, 2019: https://hashmiabrar1.github.io

8. Davit Gondauri, Ekaterine Mikautadze, Mikheil Batiashvili: Research on COVID-19 Virus Spreading Statistics based on the Examples of the Cases from Different Countries, 2020: https://doi.org/10.29333/ejgm/7869

9. Tanujit Chakrabortya, Indrajit Ghosh: Real-time forecasts and risk assessment of novel coronavirus (COVID-19) cases, 2020: A data-driven analysis: https://doi.org/10.1016/j.chaos.2020.109850

10. https://medium.com/human-in-a-machine-world/mae-and-rmse-which-metric-is-better-e60ac3bde13d