Trend Analysis and Forecasting of Covid-19 outbreak in India

**Project Report**

**Reported By- Meghdut Nandy**

**How Covid-19 Cases have**

**Risen in India since January**

**INTRODUCTION**

COVID-19 is spreading really fast around the world. The current study describes the situation of the outbreak of this disease in India. This project also discusses the regional analysis of Indian states and presents the preparedness level of India Combating this outbreak. Till now it has infected more than 24000 people in India till the end of the day on 30th April, 2020. India is the leading region in South Asian region which is the second most populous country in the world after china.

**Objective**

The objective of this project is to analyze the COVID-19 outbreak situation in India and asses the trends in near future. The scope of this project is to building forecasting models that predicts the spread of the virus next 30 days across different states/UTs, Time series based forecasting methods which are easy to build and easy to understand in these kind of critical conditions. The project does not include forecasting for any other nation suffering from COVID-19 outbreak.

**Techniques**

The current analysis has been divided into various phase. Firstly the trend analysis of number of Confirmed Cases, Death cases and Recover cases for different states & Union territories in the country. Secondly forecasting methods have been applied on the data collected from the www.kaggle.com site, the data has been considered for Indian region from 30-Jan-2020 onwards when the first case of COVID-19 was reported up to 26th April 2020, when complete lockdown of the nation has been imposed by the Govt. of India.

Based on the popular methods available for modeling and forecasting the time series data, ARIMA modeling and Exponential Smoothing methods have been implemented in this Project with the use of following development tools.

**Development Tools:-**

All the Code have been implemented using Python 3 in Google colab and Jupyter Notebook. The data files have been downloaded from www.kaggle.com in csv format.

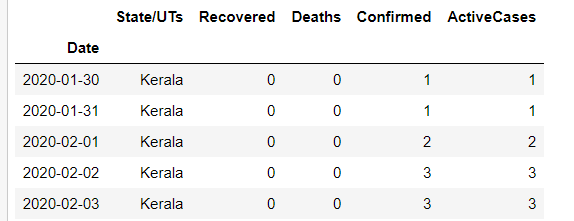
**Dataset Overview:-**

Let’s take an overview of the dataset.



**Fig.1.1**

Above dataset contains total 9 features with ranges from 30.01.2020 to 26.04.2020.Applying some Exploration Analysis & Missing value treatment by deleting some features which are not use full for prediction.so here is the final dataset consists 1350 values with 5 features by taking some values shown in Fig.1.2

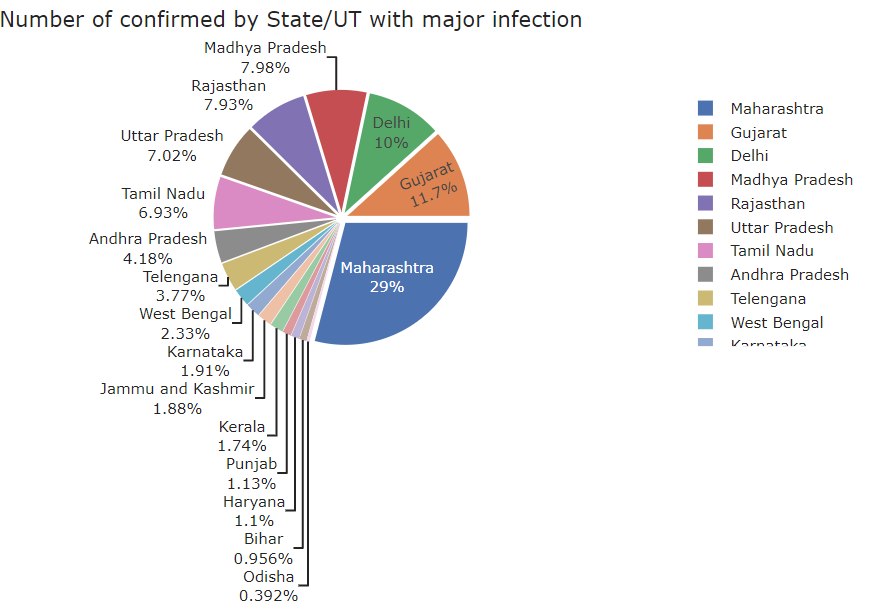


**Fig.1.2**

**Visualization:-**

No we are trying to understand the data by placing it in a visual context to easily understand. Python offers multiple great graphing libraries like seaborn, Matplotlib & so many more that come packed with lots of different features which are used for creating graphs, charts.

Now, plotting pie chart to see total number of Confirmed, Death & Recovered Cases by all State/UTs with major infection.

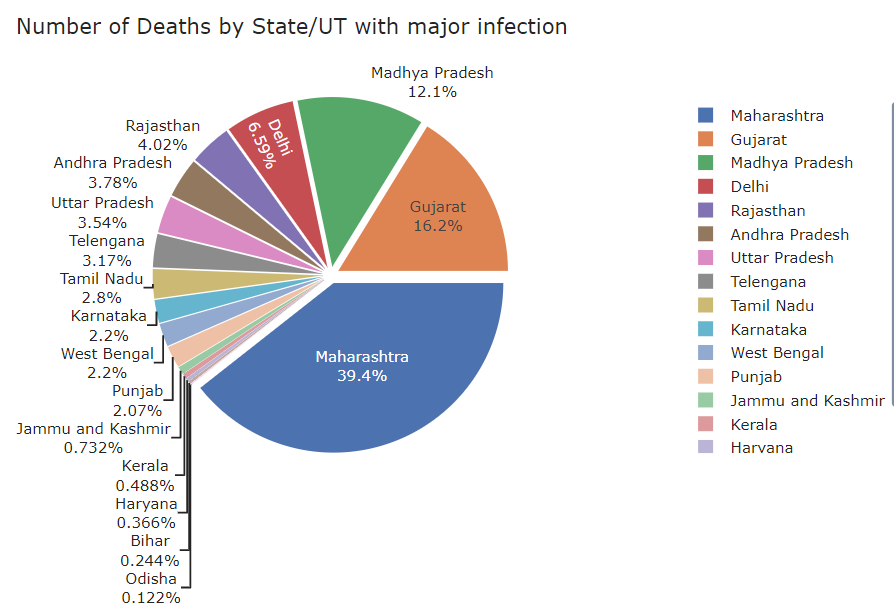
****

**Fig. 1.3**

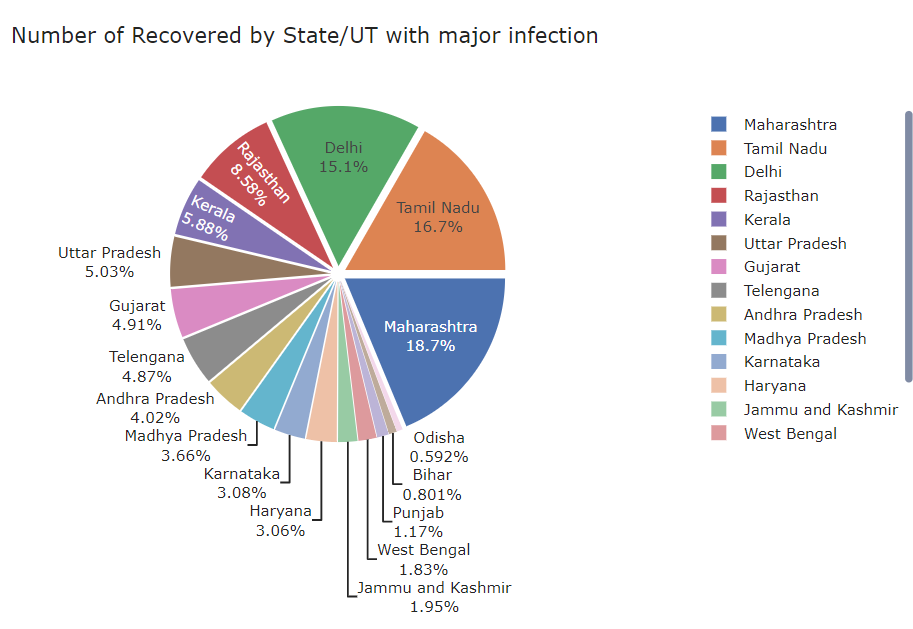
As seen from the Fig. 1.3 States having more than 100 confirmed cases are affected by virus. Maharashtra with 7628 confirmed cases, Gujarat with 3071 confirmed cases & Delhi with 2625 confirmed cases are reported that are worst affected states by the coronavirus and other states which are not present in pie chart like Andaman & Nicobar, Assam, Goa, and Manipur these states are less affected by virus.

Fig.1.4 pie chart is the total number of deaths by State/UT with major infection, here Maharashtra with 323 death cases, Gujarat with 133 death cases & Madhya Pradesh with 99 death cases. Rest other region have less than 50 death cases

Fig.1.5 pie chart shows Maharashtra and Tamil Nadu are at the top currently with more than 1000 cases are recovered from the infection.

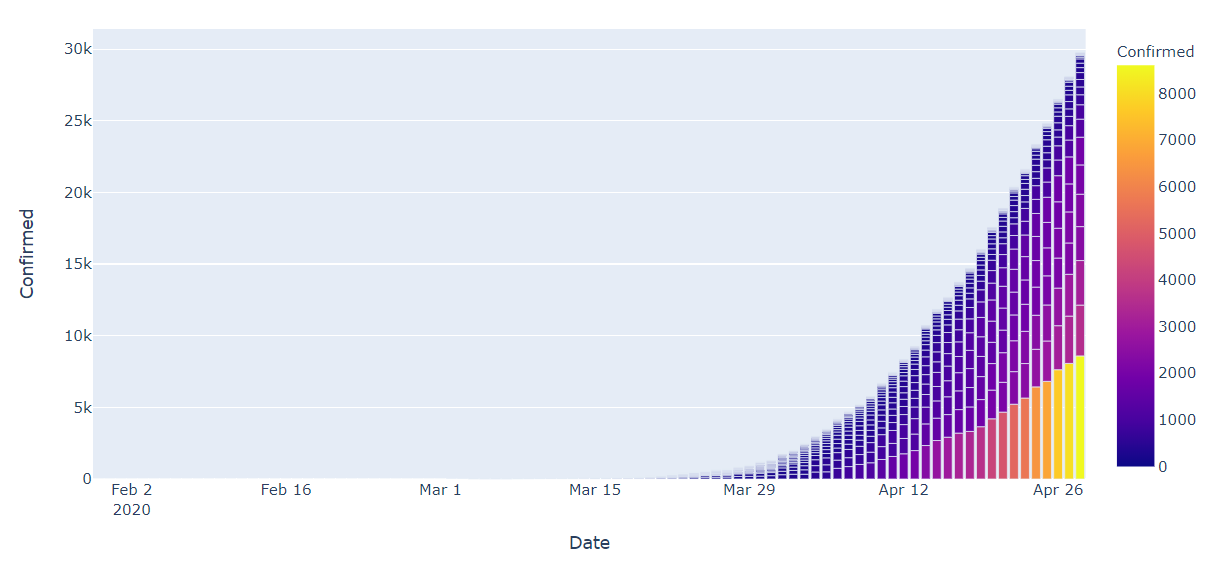
****

**Fig. 1.4**

****

**Fig. 1.5**

**Bar chart:-**



As seen from the first two figure and trends depicted by the dataset for COVID-19 outbreak, the number of infected people increased in last one week as on 28th march there are only 176 reported cases of infections and 0 deaths were reported. Suddenly in one week, the number of infection people and number of deaths increased by almost 3 times.

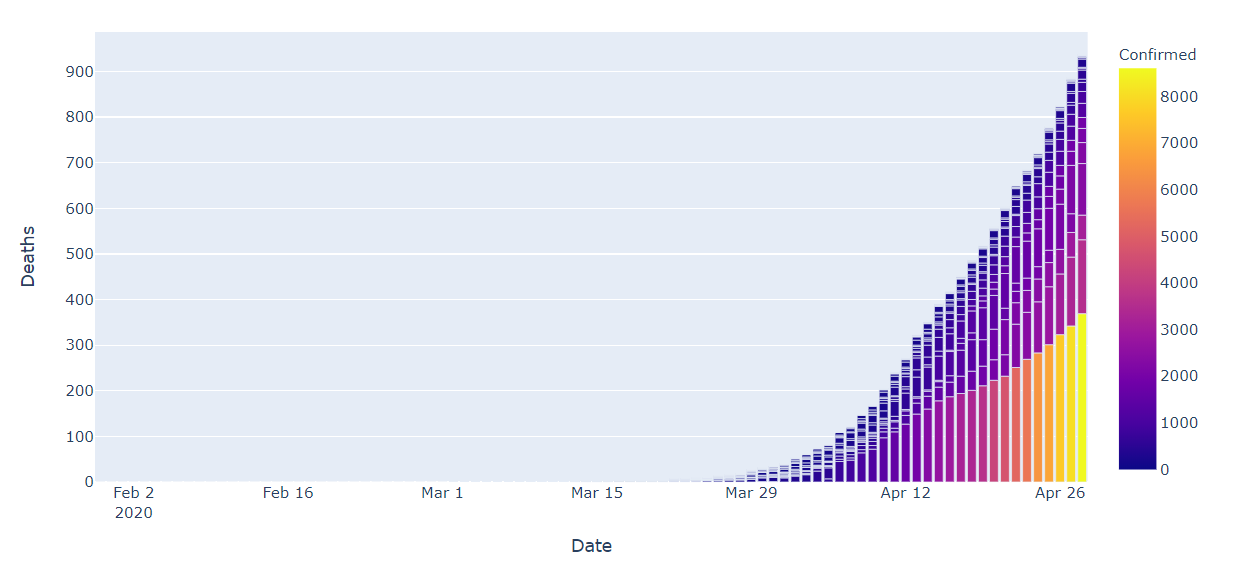
**ARIMA modelling based Forecasts in India:-**

With the number of infected people rising sharply in last one week, ARIMA model, which is a combination of 2 models AR (auto regressive) & MA (moving average) it has 3 hyper parameter –p (auto regressive lags), d (order of differentiation), Q (moving average) which comes from the AR, I & MA components, gives the predictive range of possible infections patients in India. Last figure shows the natural log of the number of infected cases as there was an exponential upward trend seen in the original data.

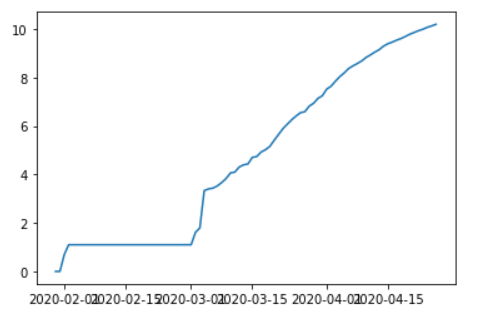
Before applying any statistical model on a Time series, the series has to be stationary, which mean that, over time periods, it should have constant mean, standard deviation & auto-covariance, Trend & Seasonality are two reasons why time series is not stationary & hence need to be correct.

Even after performing the natural log transformation on data the upward trend was still visible. Since, this can hamper the stationarity of the series, so exponentially weighted moving average were applied to eliminate the trend & seasonality from the data.

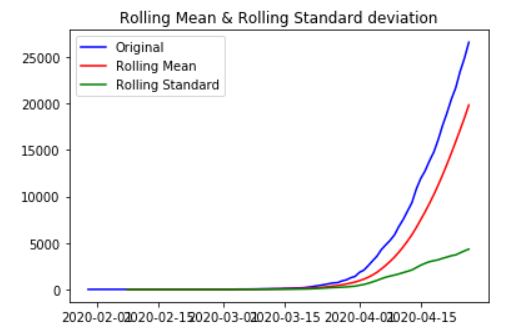
. **Fig. 1.6**



**Fig. 1.7**



**Fig. 1.8**

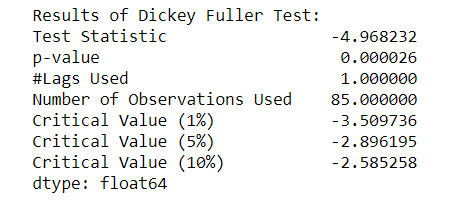


**Rolling Statistics:-**

Rolling statistics check whether the time series stationary or not, plot the moving average or moving standard deviation to see if it varies with time and it is visual technique. In figure rolling mean and standard is not constant with respect to time. So our data is not stationarity.

**Dickey-Fuller test:-**

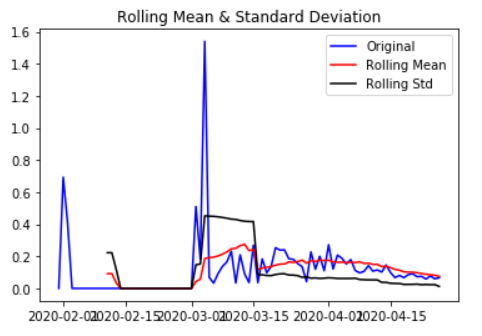
Augmented Dickey-Fuller test is used to gives us various values that can help in identifying stationarity. The Null hypothesis says that a TS is non-confidence levels. If the Test statistics is less than the critical values, we can reject the null hypothesis & says that the series is stationary. The ADCF test also gives us a p-value.



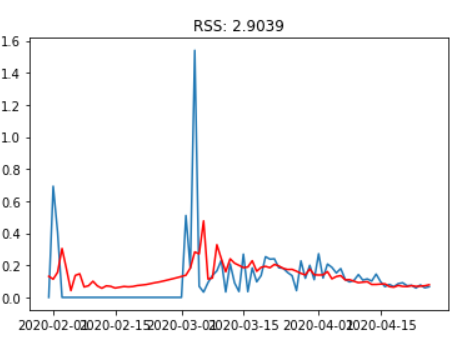
As shown in above table , the first order auto regressive component & second order moving average component were found to be statistical significant for the ARIMA(5, 1, 3) model. Therefore, model can be considered for the predictions.

Last figure show the predictability of the model and spread over the original data. After removing the trends & seasonality from the log transformed data, the forecast line follows the data line. Also, originally the log transformed data series has been well forecasted by the ARIMA model.

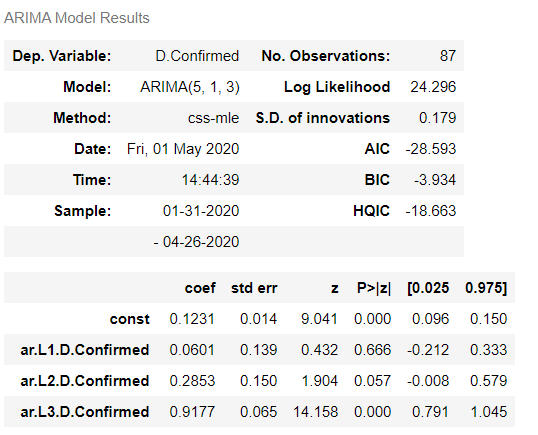
**Fig. 1.9**



**Fig. 2.0**

****

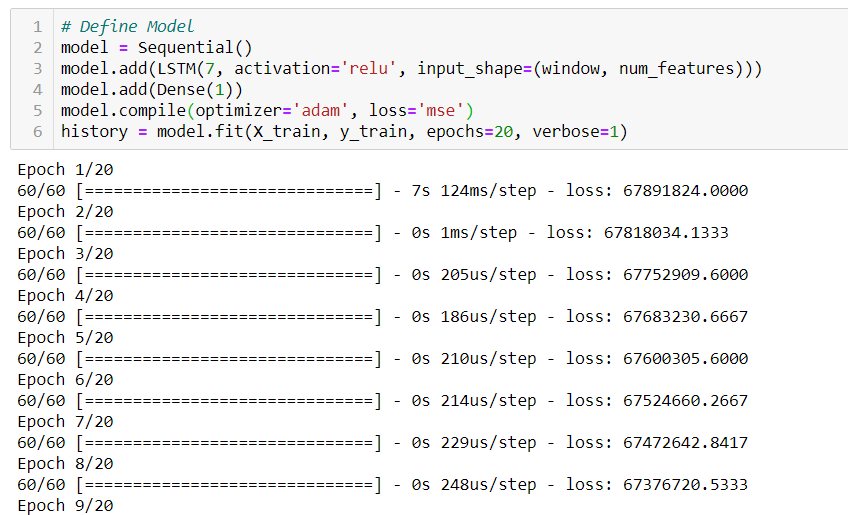
**Fig. 2.1**

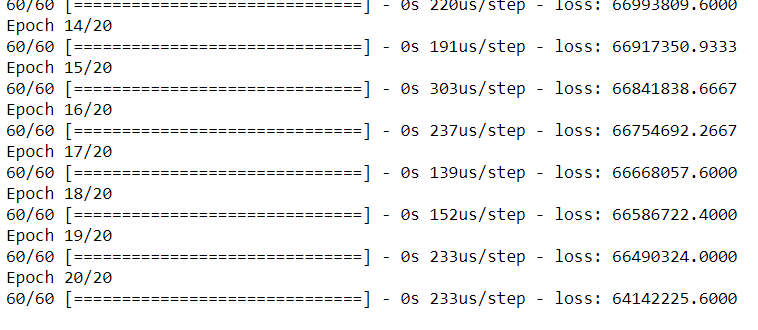
****

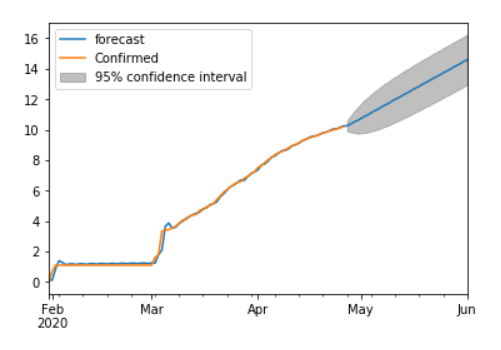
**Fig 2.2**

**LSTM (Long-Short Term Memory):-**

LSTM models can be used to diagnose the behavior of the model by converting the data into proper time steps data from which ML model can learn something.





****

**Fig. 2.3**

**Figure. Graph showing the forecasted band at 95% confidence interval for the next 30-35 days**

Above figure shows the future trend predictions which seem to be quite alarming at 95% confidence interval. Looking at the graph range for next 30 days, the optimistic scenario looks to be in control while the pessimistic scenario looks quite horrifying, with the average case being in the normal growth scenario.

The range for the predicted log values comes out to be around 15.5 in the worst case scenario while it remains around 10 in the best case scenario.

**Conclusion:-**

Although the growth rate of virus in India looks dangerous as India heading towards exponential growth, as historical data is very less for the model to be really accurate, so the training of the model may not be really accurate for now, even if the projections are 60-70% accurate, then also it would be really difficult days for the country.