COVID-19 Predictions Outcome, Forecasting and Analysis

Mriga Khanna[1], Anmol Kukreja[2], Richa Singh[3], Dr. Pawan Kumar Shukla[4], Prof. Kunal Lala[5]

[1,2,3]Student Department of Electronics & Communication Engineering R.K.G.I.T, Ghaziabad, (U.P) India

[4,5]Professor Department of Electronics & Communication Engineering R.K.G.I.T, Ghaziabad, (U.P) India

---***---

ABSTRACT

Coronavirus (COVID19), first detected inDecember 2019 in Wuhan, China. The first case was Followed back on 17 November 2019 which was again declared an Epidemic in March 2020.

This is an

infectious disease caused by SARS-CoV-2. Inalmost every affected country, the number of infected patients and those who have died is high; they have been increasing at an alarming rate.

Since earlier forecasts can reduce the spread ofthe virus, it is highly desirable to have intelligent guessing and testing tools. The use of effective forecasting models can help the government in implementing better strategies to prevent the spread of the virus. The proposed project uses 'Fbprophet' to predict the total number of deaths, cases found, aggregate number of validated cases and number of daily cases. Model made in Anaconda Distribution to get predicted numbers of cases to date.

Keywords:Machine Learning, COVID 19, visualisation, pandas.

1. INTRODUCTION:

COVID-19 has become a major global problem since World War II and the world's largest epidemic since the Spanish influenza of 1918-19. The epidemic has had a profound effect on people's lives and the country's economy [4]. Among the many questions related to infection, governments and individuals are most concerned about (i) when will COVID19 infection rate reach its peak? (ii) How long will the epidemic stop and (iii) What will be the total number of people who will eventually become infected? (iv) What will be the death toll? [4] These questions are of great concern to India, a country with a large population and economic differences. The spread of the disease in India is much lower than in China, the USA and other European countries. India is under total closure from March 21, 2020 with the expert's belief that this could be detrimental to reducing the spread of Covid19 among its citizens [4]. As of April, 30 the number of COVID19 cases in India was 36669 and she died in 1229 as a result of Severe Acute Respiratory Syndrome (SARS). The total number of people found COVID19 in India is 140980 as of 4June

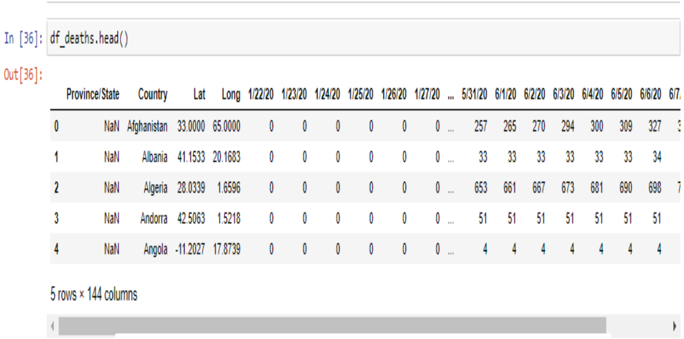
Imprisonment affects the poor and the migrant workers. Staying at home may not be possible immediately because many people may die of starvation and other diseases [4]. Media reports from around the world report on the problem and how it affects people's lives. A lot of research is being done at all levels to quickly gather data, develop mitigation tools and similar methods and applications. Therefore, policymakers and authorities want to have an overview of the current situation and to imagine how quickly it can spread.

This paper discusses the proposed prediction model of COVID19 that is spreading worldwide using machine learning that has been used in Anaconda distribution. The model steps are discussed in the subset section.

2. OBJECTIVE:

This research paper seeks to investigate the potential global impact of Novel Coronavirus (COVID-19) by predicting the prevalence of confirmed cases and the analysis of the number of deaths and acquisitions with the help of machine learning using Python. This paper introduces an objective way to predict the progress of case.

outlining the timeline of the forecasting process which has major implications for doing the planning and decision-making of how it will be going to be implemented in order to get the actual data.

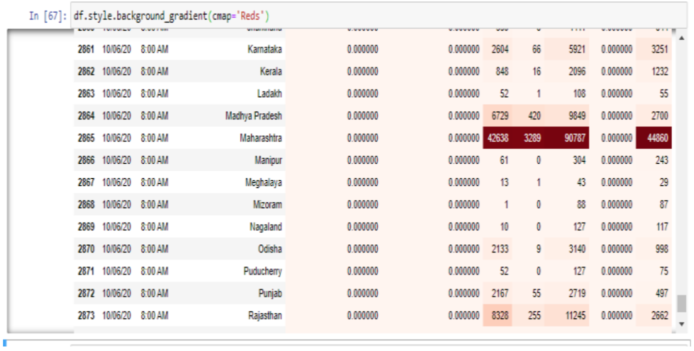


```
In [36]: df_deaths.head()

Out[36]:
   Province/State      Country      Lat     Long 1/22/20 1/23/20 1/24/20 1/25/20 1/26/20 1/27/20 ... 5/31/20 6/1/20 6/2/20 6/3/20 6/4/20 6/5/20 6/6/20 6/7
0             NaN  Afghanistan  33.0000  65.0000       0       0       0       0       0       0 ...     257    265    270    294    300    309    327   3
1             NaN      Albania  41.1533  20.1683       0       0       0       0       0       0 ...      33     33     33     33     33     33     34
2             NaN      Algeria  28.0339   1.6596       0       0       0       0       0       0 ...     653    661    667    673    681    690    698   7
3             NaN      Andorra  42.5063   1.5218       0       0       0       0       0       0 ...      51     51     51     51     51     51     51
4             NaN       Angola -11.2027  17.8739       0       0       0       0       0       0 ...       4      4      4      4      4      4      4

5 rows × 144 columns
```

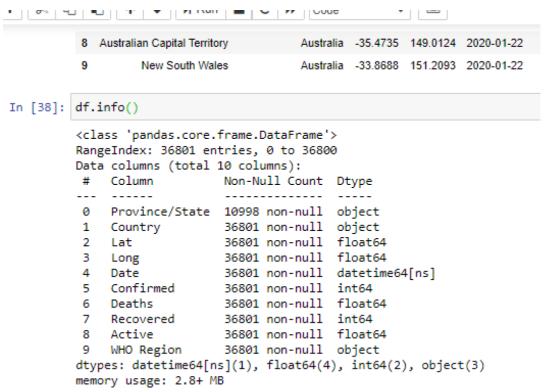
Figure 1



```
In [67]: df.style.background_gradient(cmap='Reds')

2861 10/06/20 8:00 AM      Karnataka  0.000000  0.000000  2604    66   5921  0.000000   3251
2862 10/06/20 8:00 AM         Kerala  0.000000  0.000000   848    16   2096  0.000000   1232
2863 10/06/20 8:00 AM         Ladakh  0.000000  0.000000    52     1    108  0.000000     55
2864 10/06/20 8:00 AM Madhya Pradesh  0.000000  0.000000  6729   420   9849  0.000000   2700
2865 10/06/20 8:00 AM    Maharashtra  0.000000  0.000000 42638  3289  90787  0.000000  44860
2866 10/06/20 8:00 AM        Manipur  0.000000  0.000000    61     0    304  0.000000    243
2867 10/06/20 8:00 AM      Meghalaya  0.000000  0.000000    13     1     43  0.000000     29
2868 10/06/20 8:00 AM        Mizoram  0.000000  0.000000     1     0     88  0.000000     87
2869 10/06/20 8:00 AM       Nagaland  0.000000  0.000000    10     0    127  0.000000    117
2870 10/06/20 8:00 AM         Odisha  0.000000  0.000000  2133     9   3140  0.000000    998
2871 10/06/20 8:00 AM     Puducherry  0.000000  0.000000    52     0    127  0.000000     75
2872 10/06/20 8:00 AM         Punjab  0.000000  0.000000  2167    55   2719  0.000000    497
2873 10/06/20 8:00 AM      Rajasthan  0.000000  0.000000  8328   255  11245  0.000000   2662
```

Figure 2



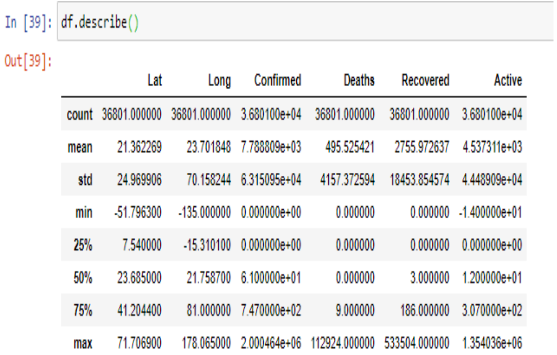
```
8  Australian Capital Territory   Australia  -35.4735  149.0124  2020-01-22
9        New South Wales          Australia  -33.8688  151.2093  2020-01-22
```

In [38]: df.info()

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 36801 entries, 0 to 36800
Data columns (total 10 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Province/State  10998 non-null  object
 1   Country         36801 non-null  object
 2   Lat             36801 non-null  float64
 3   Long            36801 non-null  float64
 4   Date            36801 non-null  datetime64[ns]
 5   Confirmed       36801 non-null  int64
 6   Deaths          36801 non-null  float64
 7   Recovered       36801 non-null  int64
 8   Active          36801 non-null  float64
 9   WHO Region      36801 non-null  object
dtypes: datetime64[ns](1), float64(4), int64(2), object(3)
memory usage: 2.8+ MB
```

Figure 3

So, here we can see that the above dataset provides the record till 10 June 2020, indicating the number of cases found in specific states in the country. Moreover, the dataset contains2666 entries and 9 features. The colour red in the above dataset indicate the highest no. of recovered cases, confirmed cases and death cases which is in Maharashtra .Also if we are doing some analysis on dataset in order to get some statistics values like as count ,mean, standard deviation (std), minimum ,maximum(max). All of these results are shown below :



In [39]: df.describe()

Out[39]:

	Lat	Long	Confirmed	Deaths	Recovered	Active
count	36801.000000	36801.000000	3.680100e+04	36801.000000	36801.000000	3.680100e+04
mean	21.362269	23.701848	7.788809e+03	495.525421	2755.972637	4.537311e+03
std	24.969906	70.158244	6.315095e+04	4157.372594	18453.854574	4.448909e+04
min	-51.796300	-135.000000	0.000000e+00	0.000000	0.000000	-1.400000e+01
25%	7.540000	-15.310100	0.000000e+00	0.000000	0.000000	0.000000e+00
50%	23.685000	21.758700	6.100000e+01	0.000000	3.000000	1.200000e+01
75%	41.204400	81.000000	7.470000e+02	9.000000	186.000000	3.070000e+02
max	71.706900	178.065000	2.000464e+06	112924.000000	533504.000000	1.354036e+06

Figure 4

In this we have imported the data from .csv file and passed the dates to perform date/time operation.

Now to look at the cases for India.Here df is used for the data-frames and some arguments are passed into the function to get the total confirmed, recovered, death cases in different states of the country along the date passed into the function. We have analysed the situation for different states in India and plotted a bar graph using a matplotlib library.

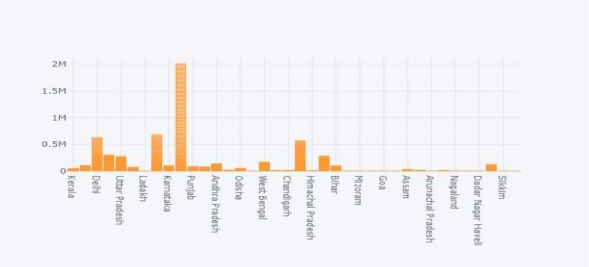


Figure 5

The pace of novel coronavirus has quickened in the past few days. The above graph represents the state wise analysis of India. The rise and fall of the cases in different states can be analysed from the graph. As per the analysis the states like Maharashtra, Gujarat and Tamil Nadu being highly infected by the virus. As single-handed Maharashtra only accounts for 34% of the cases, which now has made the country amongst the top five caseloadcountries.

The health care centres have warned the five most affected states named as - Maharashtra, Tamil Nadu, Delhi, Gujarat, Uttar Pradesh. Maharashtra etc are surpassing 94.01 thousand cases (till June 10) and 120 deaths in a single day whereas, Karnataka is wrapping more influenza like illness caseloads, Delhi in need of 80 thousand beds daily by the July end, Gujarat with the caseload of 21 thousand in which 70% of state burden is from Ahmedabad.

Then, a horizontal bar graph has been plotted to compare the number of cases in India and outside India in order to get more appropriate visualization. The cases outside India have been calculated by subtracting India's confirmed cases from the counties confirmed casesworldwide.We have used plt.barh function to plot the horizontal bars. The title to this plot is given as "Number Of Coronavirus Confirmed Cases" and the following data isobtained:

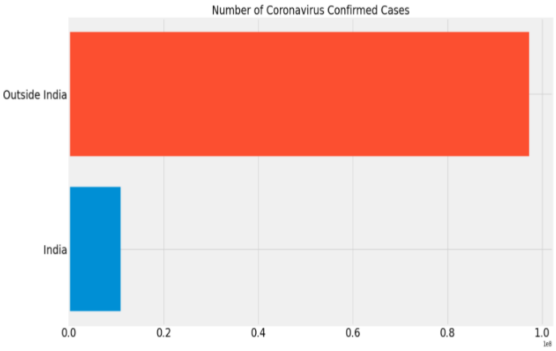


Figure 6

The blue bar represents the total number of confirmed cases in India and the red bar represents the number of confirmed cases from outside India. We have done the visualization for only unique countries with the most confirmed cases and the rest of the countries are grouped into the other categories. For this the two empty lists are created as visual_unique_countries [], visual_confirm_cases []. Then a user defined function is created called plot_bar_graphs () that will plot the bar graph with a title "Number of Covid19 Confirmed Cases in Countries"

So from the above bar graph we can visualize that the United States has the highest number of covid19 confirmed cases and then the countriesare sortedindescendingorderofthecovid19confirmedcases.

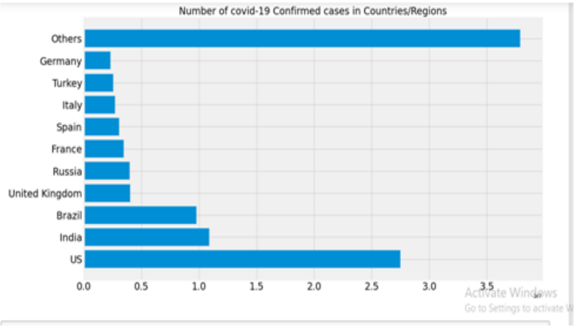


Figure 7

At the top we can visualize the others category too. The above bar graph can be visualized with the help of pie charts to get a more clear picture about the unique countries with the confirmed cases.

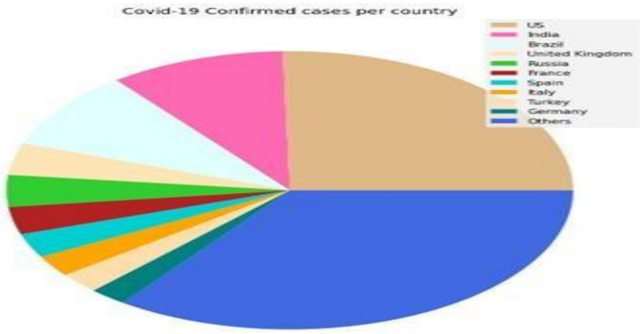


Figure 8

3.2 PREDICTION FOR THECASES

- In this step , we have converted all the dates and cases in the form of a numpy array using the np.array function() .Then we are predicting the cases for next 20 daystogetthevaluesforfutureforecast.

- Next, we are converting all the integers into date time values for better visualization. Then we split the data into training and testing cells. For this we have used the train_test_split function. 75% of the data is used for training the model and 25% for testing themodel.

- Next we will be transforming the dataset for polynomial regression using fit_transform methodtotransformourtraining,testingand future forecast data. Then, we build the polynomial regression model using the Linear Regression function and a predict function is used inside that to predict the test data values.

- The graph is then plotted between the test valuedatasetandthevaluespredictedfromthe polynomial regression model. So the below figure represents the graph in which the blue line represents the test data and the red line represents the polynomial regression predictions.

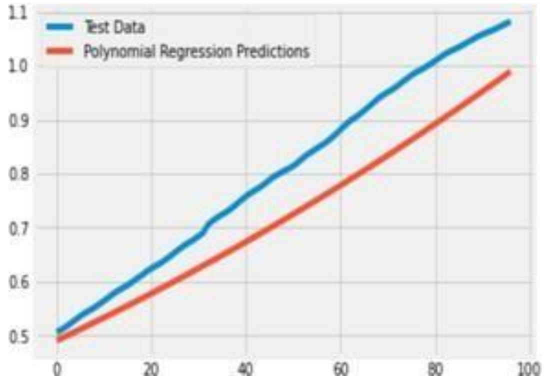


Figure 9

We have created a bar graph with the adjusted dates and world daily increase in confirmed cases and the following output is obtained :

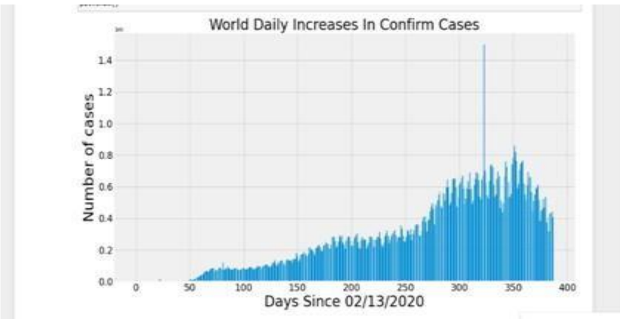


Figure 10

As we can see in some of the days the cases go very high reaching hundreds, thousands in an approx time or in adaytime.

Similarly we have done the visualization with the help of bar graph to get the world daily increase in death and recoveries

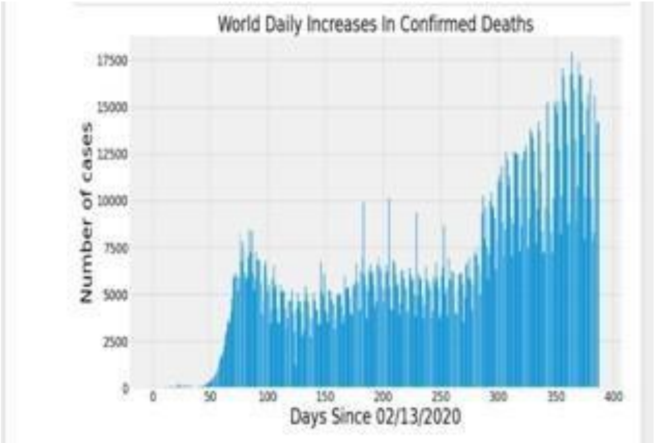
As we can see in the below figure about :

1.World daily increases in confirmed deaths

2.World daily increases in confirmed recoveries

We have the estimated increase and recoveries in the cases in the below figure shown :

Figure11



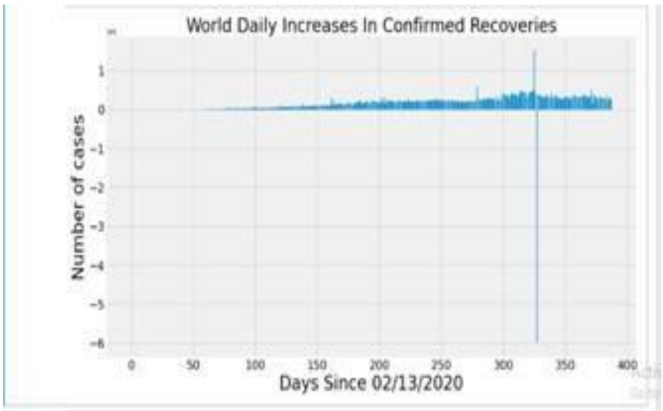


Figure12

3.3 COMPARISON OF THE PLOTS (CONFIRMED, DEATHANDRECOVERIESWORLD WIDE):

The COVID-19 pandemic is spreading its wings across the globe at a surprisingly faster rate and has already resulted in thousands of deaths across all over the other countries. Unfortunately, this number is sure to grow within a short period and healthcare organizations would soon face scarcity of resources. In this sequel, it is important to analyze various forecasting models for COVID-19 to empower allied organizations with more1.Visualize the time series to analyse the trendsappropriate informationpossible.

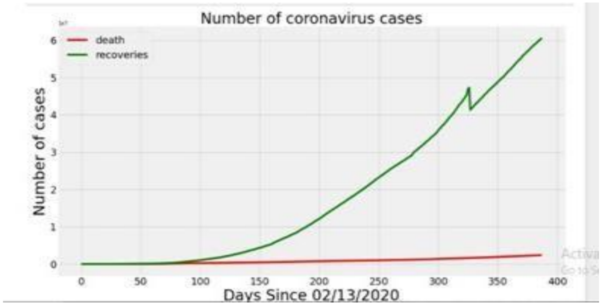


Figure 13

3.4 FORECASTING:

Forecasting involves taking models fit on historical data and using them to predictfuture observations. An important distinction in forecasting is that the future is completely unavailable and must only be estimated from what already happened. When we are using a classical statistic, then the primary concern is to do the analysis of the real time series. Time series analysis provides a body of techniques to better understand a dataset. Following are the steps involved in forecasting the time series analysis:

Fbprophet is available in both Python and R. This covers python installation and implementation of Prophet. The main properties of the Fbprophet are as follows:

1. Accurate andfast

prior to building any kind of time series model.

2. Stantionarize theseries

3. Find the optimal prameter

4. Build the modelrequired

5. Makepredictions.

Therefore, to make time series predictions we will use a library called the Fbprophet and Arima model.

In which ARMIA stands for "autoregressive integrated moving average" . This model is used in statistics and econometrics to measure events that happen over a period of time. It is used to understand the past data or predict future data in a series. These models are fitted to time series data either to better understand the data or to predict future points in the series.

ARIMA uses a number of lagged observations of time series to forecast observations. A weight is applied to each of the past terms and the weights can vary based on how recent they are and AR(x) means x lagged error terms are going to be used in the ARIMA model and it relies on auto regression. The main three properties of ARIMAare:

1. AutoRegressive
2. Integrated
3. MovingAverage

Fbprophet is a library created by Facebook, written in python and it allows us to make time series analysis and make some predictions based on the data we have accumulated over the days. Prophet is a procedure for forecasting time series data based on an additive model where non linear trends are fit with yearly ,weekly,and daily seasonality, plus holiday effects and several seasons of historical data. Prophet is robust to missing data and shifts in the trend, and typically handles outlierswell.

2. Fully automatic 3.Tunable forecastsetc.

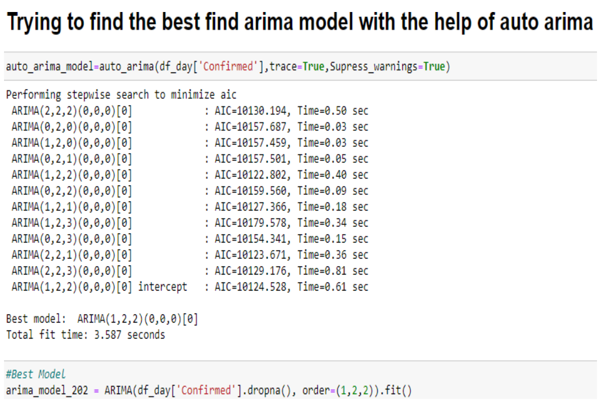
So, we import the libraries. The way prophet works is that it needs 2 columns:

1. DS which stands for datestamp.

2. The other is a variable 'y' which is tryingt o predict the cases.

We are predicting with 95% interval of confidence.

Firstly we have imported the arima model and tried to find out the best arima model with the help of auto arima:



Trying to find the best find arima model with the help of auto arima

```
auto_arima_model=auto_arima(df_day['Confirmed'],trace=True,Supress_warnings=True)
```

```
Performing stepwise search to minimize aic
 ARIMA(2,2,2)(0,0,0)[0]             : AIC=10130.194, Time=0.50 sec
 ARIMA(0,2,0)(0,0,0)[0]             : AIC=10157.687, Time=0.03 sec
 ARIMA(1,2,0)(0,0,0)[0]             : AIC=10157.459, Time=0.03 sec
 ARIMA(0,2,1)(0,0,0)[0]             : AIC=10157.501, Time=0.05 sec
 ARIMA(1,2,2)(0,0,0)[0]             : AIC=10122.802, Time=0.40 sec
 ARIMA(0,2,2)(0,0,0)[0]             : AIC=10159.560, Time=0.09 sec
 ARIMA(1,2,1)(0,0,0)[0]             : AIC=10127.366, Time=0.18 sec
 ARIMA(1,2,3)(0,0,0)[0]             : AIC=10179.578, Time=0.34 sec
 ARIMA(0,2,3)(0,0,0)[0]             : AIC=10154.341, Time=0.15 sec
 ARIMA(2,2,1)(0,0,0)[0]             : AIC=10123.671, Time=0.36 sec
 ARIMA(2,2,3)(0,0,0)[0]             : AIC=10129.176, Time=0.81 sec
 ARIMA(1,2,2)(0,0,0)[0] intercept   : AIC=10124.528, Time=0.61 sec

Best model:  ARIMA(1,2,2)(0,0,0)[0]
Total fit time: 3.587 seconds
```

```
#Best Model
arima_model_202 = ARIMA(df_day['Confirmed'].dropna(), order=(1,2,2)).fit()
```

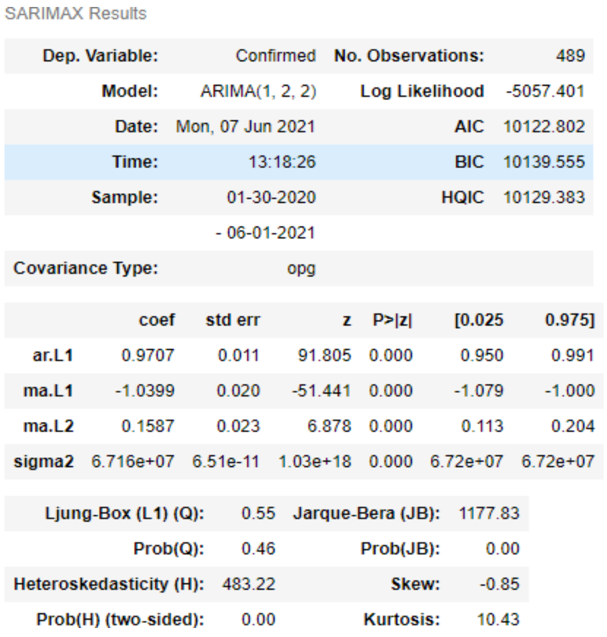
Figure:14

Then we have plotted the arima modelstatistics withthefollowingwarnings:

[1] Covariance matrix calculated using the outer product of gradients(complex-step).

[2] Covariance matrix is singular or near-singular, with condition number 4.66e+33. Standard errors may beunstable.

Figure 15



SARIMAX Results

Dep. Variable:	Confirmed	No. Observations:	489
Model:	ARIMA(1, 2, 2)	Log Likelihood	-5057.401
Date:	Mon, 07 Jun 2021	AIC	10122.802
Time:	13:18:26	BIC	10139.555
Sample:	01-30-2020	HQIC	10129.383
	- 06-01-2021		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.9707	0.011	91.805	0.000	0.950	0.991
ma.L1	-1.0399	0.020	-51.441	0.000	-1.079	-1.000
ma.L2	0.1587	0.023	6.878	0.000	0.113	0.204
sigma2	6.716e+07	6.51e-11	1.03e+18	0.000	6.72e+07	6.72e+07

Ljung-Box (L1) (Q):	0.55	Jarque-Bera (JB):	1177.83
Prob(Q):	0.46	Prob(JB):	0.00
Heteroskedasticity (H):	483.22	Skew:	-0.85
Prob(H) (two-sided):	0.00	Kurtosis:	10.43

3.5 FORECASTING THE OCCURRENCE OF THIRD WAVEIN INDIA WITH THE HELP OF ARIMA MODEL:

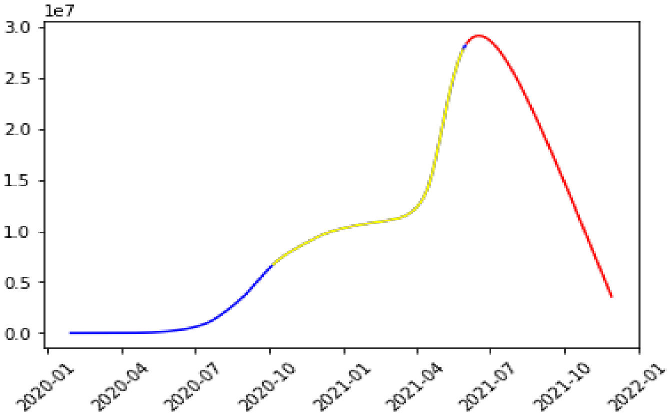
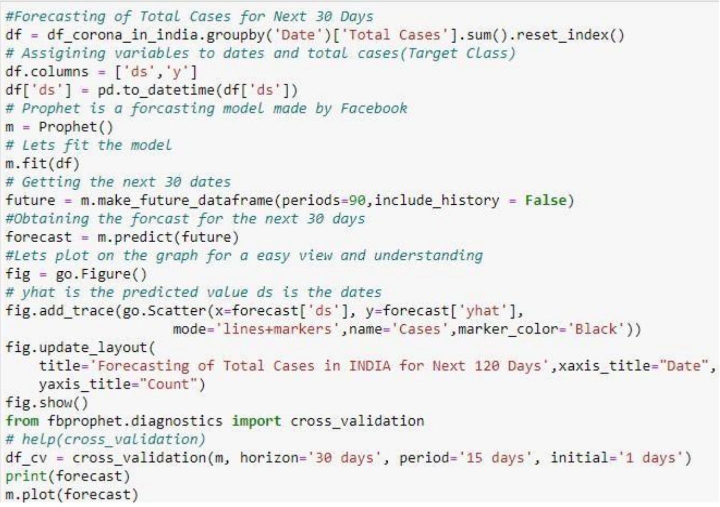


Figure 16

3.6 FORECASTING THE THIRD WAVE WITH THE HELP OF FBPROPHET:

In this we can see about the forecasting of the third wave occurring and the graphical representation of the increase in number of



```
#Forecasting of Total Cases for Next 30 Days
df = df_corona_in_india.groupby('Date')['Total Cases'].sum().reset_index()
# Assigining variables to dates and total cases(Target Class)
df.columns = ['ds','y']
df['ds'] = pd.to_datetime(df['ds'])
# Prophet is a forcasting model made by Facebook
m = Prophet()
# Lets fit the model
m.fit(df)
# Getting the next 30 dates
future = m.make_future_dataframe(periods=90,include_history = False)
#Obtaining the forcast for the next 30 days
forecast = m.predict(future)
#Lets plot on the graph for a easy view and understanding
fig = go.Figure()
# yhat is the predicted value ds is the dates
fig.add_trace(go.Scatter(x=forecast['ds'], y=forecast['yhat'],
                    mode='lines+markers',name='Cases',marker_color='Black'))
fig.update_layout(
    title='Forecasting of Total Cases in INDIA for Next 120 Days',xaxis_title="Date",
    yaxis_title="Count")
fig.show()
from fbprophet.diagnostics import cross_validation
# help(cross_validation)
df_cv = cross_validation(m, horizon='30 days', period='15 days', initial='1 days')
print(forecast)
m.plot(forecast)
```

Figure 17

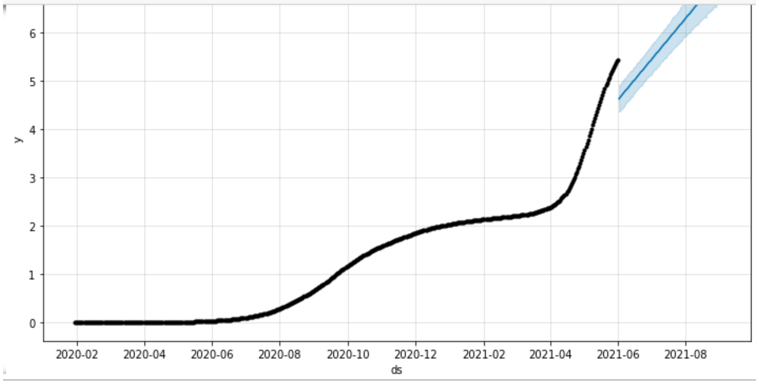


Figure 18

4. RESULTS:

1. By doing predictions for the total number of confirmed cases and forecasting the plots we can analyse there is a steady increase in the number of cases from figure 8. The graph is rising exponentially. At its early trend it is increasing slowly but from mid-March 2020, the curve has taken a sharp turn and its increasing gradually till June2020

2.Similarly by doing predictions for total number of death cases and forecasting the plots we can analyse that there is a steady increases in the cases figure 9.Also by visualising the trend we can analyse that despite having lowerfatality rate, SARS-CoV-2 caused thrice the total of deaths when compared to the combined statistics of deaths caused by both MERS and SARS-CoV.

3. Likely by doing the predictions for total number of recovered cases and forecasting the plots we can analyse that the number of recovery is also going to take a rapid increase figure 10 as there are more number of patients get introduced .If we analyse the trend, it looks like the number of patients who have recovered matches with the active number of cases. So, if thetotalnumberofconfirmedcasesincreasesthe recovery rate alsoincreases.

5. CONCLUSIONS:

The above proposed methodology usually predicts the total number of COVID-19 infected cases, total number of recovered cases, and total number of deaths all over thecountry.Weeklypredictions have also been done for the confirmed, recovered, death cases. Based on the recent trends, the future trends have been predicted and the plots are visualized for the confirmed,recovery,death case, using machine learning. The methodology used has 95%accuracy in predicting the confirmeddeaths and recovered cases. The machine learning approaches are useful in forecasting the impact of COVID-19 on different sectors which may help the government in implementing proper policies to overcome the economic crisis [6]. Therefore, to empower the government and healthcare sector, it is necessary to analyse various forecasting and prediction tools. Moreover, the accuracy of prediction tools can be enhanced by the usage of advanced computing intelligent approaches such as ensemble method like bagging, stacking etc., application of optimization techniques, usage of artificial neuralnetworks and higher order neural networks inthe screening and prediction of COVID-19 which is considered as further scope of research [6]. The public health officials and government should take different preventive measures to control the rapid increase of the COVID-19 [3]. Besides the officials, the general public should also maintain social distance and use precautions in order to ensure their safety and control the disease from further spreading[3].

REFERENCES:

1. Putra M, Kesavan MM, Brackney K, Hackney DN, Roosa MK. Forecasting the Impact of Coronavirus Disease During Delivery Hospitalization: An Aid for Resources Utilization. American Journal of Obstetrics & Gynecology MFM. 2020 Apr25:100127.
2. ChakrabortyT,Ghosh I. Real-time forecasts and risk assessment of novel coronavirus (COVID-19) cases: A data- driven analysis. Chaos, Solitons & Fractals. 2020Apr30:109850.
3. YousafMohammad, Zahir Siddiqui, Riaz Mohammad, Hussain SM, Shah Khan. Statistical Analysis of Forecasting COVID-19 for the Upcoming Month in Pakistan. Chaos, Solitons & Fractals. 2020 May25:109926.
4. Parbat D. A Python based SupportVectorRegression Model for prediction of Covid19 cases in India. Chaos and Solitons & the Fractals. 2020 May 31:109942.
5. Panwar H, Gupta PK, Siddiqui MK, Morales-Menendez R, SinghV.Application of Deep Learning for Fast Detection of COVID-19 in X-Rays using CONVnet. Chaos, Solitons &Fractals. 2020 May 28:109944.
6. Rekha Hanumanthu S. Role of Intelligent Computing in COVID-19 Prognosis: A State-of-the-ArtReview.Chaos, Solitons & Fractals. 2020 May29:109947.
7. Nishiura H, Linton NM, Akhmetzhanov AR. Serial interval of novel coronavirus (COVID-19) infections. International journal of infectious diseases. 2020 Mar 4.
8. Lee H, Park SJ, Lee GR, Kim JE, Lee JH, JungY,NamEW.The relationship between the trends in COVID-19 prevalence and traffic levels in South Korea. International Journal of Infectious Diseases. 2020 Jul;96:399.
9. TuliS,TuliS,TuliR,GillSS.Predictingthe actual Growth andTrendof COVID-19 and Pandemic using Machine Learning and Cloud Computing and Internet of Things. 2020 May12:100222.
10. Hasaann N. A Methodological Approach for Predicting COVID-19 Epidemic Using EEMD-ANN Hybrid Model. and the Internet of Things. 2020 May28:100228.
11. Lansbury L, Lim B, BaskaranV,Lim WS. Co-infections in people with COVID-19: a systematic review and meta-analysis. Journal of Infection. 2020 May27.
12. Chintalapudi N, Battineni G, AmentaF.COVID-19 disease outbreak forecasting of registered and recovered cases after sixty-day lockdown in Italy: A data driven model approach. Journal of Microbiology, Immunology and Infection. 2020 Apr13.
13. WangL, Li J, Guo S, Xie N,YaoL, CaoY,DaySW,Howard SC, Graff JC, GuT,Ji J. Real-time estimation and prediction of mortality caused by COVID-19with patient information-based algorithm. Science of the Total Environment. 2020 Apr 8:138394.
14. Duffey RB, Zio E. Prediction ofCoVid-19infection, transmission and recovery rates: a new analysis and global societal comparisons. Safety Science. 2020 May 28:104854.
15. TagliazucchiE, BalenzuelaP,Travizano M, Mindlin GB, Mininni PD. Lessons from being challenged by COVID-19. Chaos, Solitons & Fractals. 2020 May23:1099