COVID-19 ANALYSIS AND FORECASTING-INDIA

BACKGROUND

The novel corona virus was first identified in China in December 2019. In the early 2020s, it became pandemic all over the world. In India, it became a serious issue by mid of March 2020. A machine learning model which would predict the approximate number of probable COVID-19 positive cases, fatalities and recovered cases would be helpful. The algorithm would be best appreciated if it could forecast the numbers of Confirmed cases, recoveries and fatalities accurately.

OBJECTIVE

To estimate the probable number of confirmed cases, fatalities and recoveries of COVID-19 in India between July 22nd and August 8th, form past data, also to find their respective cumulative numbers for the given dates.

METHODS

The SARIMA model is used with the parameters ‘order’ and ‘seasonal order’ and their best values were found using AIC metric. On fitting the model on variables (Daily Confirmed, Daily Recovered, Daily Deceased) with best combination of order and seasonal order, the model is tested on test data and it’s performance is measured using RMSE metric.

RESULTS

Using the entire available time series data of infected patients, fatalities and recoveries from January 30th to July 21st 2020, we built an SARIMA model and forecasted the approximate number of confirmed cases, fatalities and recoveries and their respective cumulatives on the dates of July 22nd to August 8th.

CONCLUSION

The SARIMA model forecasts a continuous increase in numbers of active cases, recoveries and deceased from 22nd july. From the results, the percentage of active cases has steady raise in the period (22nd july to 8th August) while percentage confirmed cases that are either recovered or deceased declines with small margin.

1. INTRODUCTION

Corona Virus Disease also known as COVID- 19 was discovered in Wuhan, China in December 2019. It became pandemic all over the world in the early 2020s. The number of people tested positive for the disease are multiplying day by day in almost all the parts of the world, whereas on the other side the death toll also keeps multiplying. On 30 January 2020, the World Health Organization (WHO) director-general declared the coronavirus disease 2019 outbreak a public-health emergency of international concern. Advanced computational models, such as those based on machine learning, have shown great potential in tracing the source or predicting the future spread of infectious diseases. A machine learning model to globally forecast the probable number of confirmed cases and fatalities that would occur in the forthcoming days would be helpful. This problem could be better handled using Time series models.

On an average, out of 1 million enrolments in WHO, almost seventy percent of cases belong to COVID-19. The number of countries implementing additional health measures that significantly interfere with international traffic has increased since the declaration of COVID-19 as a public health emergency of international concern. The United Nations World Tourism Organization launched a Crisis Committee to review the impact of the outbreak on the aviation, shipping and tourism sectors and propose innovative solutions for recovery. WHO has shared information with Member States every week since 6 February 2020 through the Event Information Site, a secure platform accessible by national IHR focal points and United Nations (UN) agencies. The majority of measures relate to the denial of entry of passengers from countries experiencing outbreaks, followed by flight suspensions, visa restrictions, border closures, and quarantine measures.

The ultimate aim of this project is to predict the number of positive cases, fatalities and recoveries due to COVID-19 between July 22nd and August 8th, provided the past data.

2. DATA AND METHODS

2.1 Data

The source of the data is Kaggle. This data represents the overall spread of COVID-19 in India. This is a time series data giving information about the confirmed cases, fatalities and recoveries in India from 30th January to 21st July 2020. The data contains 177 records denoting 177 days.

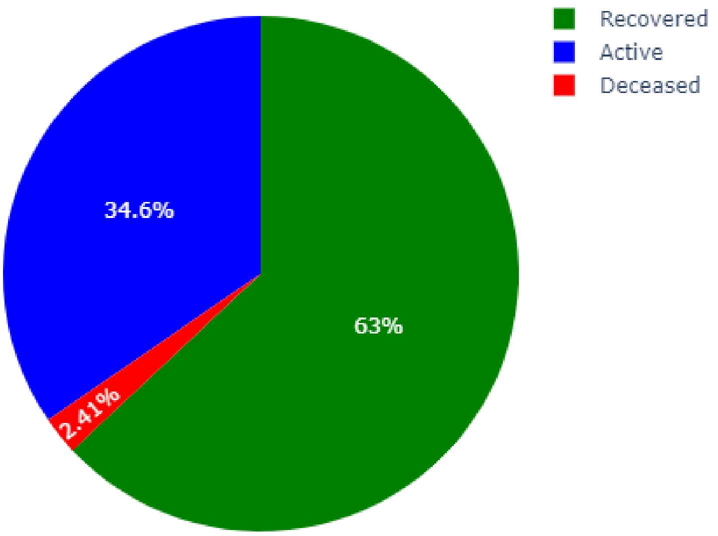

Fig (2.1.1)
Representation of Corona cases as of July 21st
Active - 4,13,030 Recovered -7,52,284

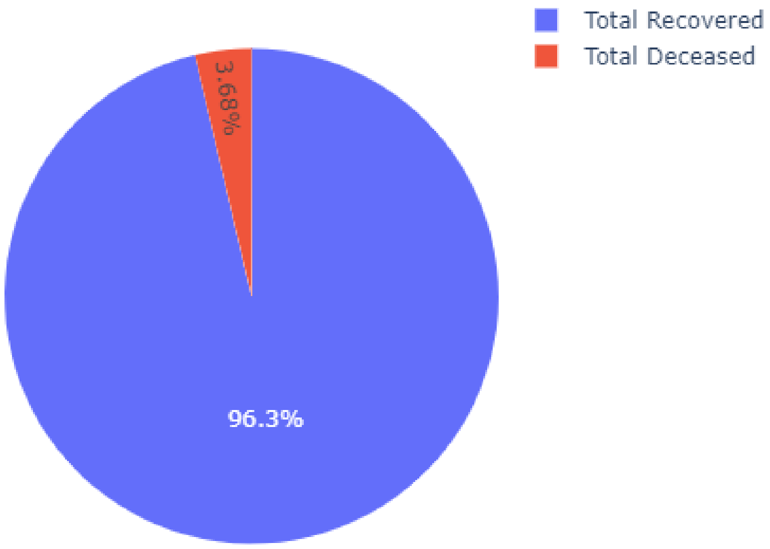

Fig (2.1.2)
Recovered v/s Deceased as of July 21st
Recovered -7,52,284 Deceased -28,772

2.2 Methods

Once when the dataset is imported as a data frame in the python notebook, Exploratory Data Analytics (EDA) is done to study general facts about the data, treat missing data and to get the statistical information contained in it. Since this is a time series problem, we use the time series model called SARIMA for which we import the library called 'statsmodels'. In this project we predict the daily and cumulative number of confirmed cases, fatalities and recoveries.

2.2.1 SARIMA Time series model

To predict the three variables (Confirmed, Recovered, Deceased), we fit them into SARIMA model with the parameter order (p,d,q) where p ,d, q represent order of Auto regression, Differencing and Moving average respectively and parameter seasonal order (P,D,Q,S) where P,D,Q are order of seasonal Auto regression, seasonal Differencing and seasonal Moving average. S represents time span of repeating seasonal pattern. The curves of Daily confirmed, Daily recovered, Daily deceased showed a repetitive trend for every seven days. This could be probably due to underreporting of cases on weekends. So, we have chosen 7 as Seasonality.

We have considered (0,1,2) as possible values for each of p,d,q and P,D,Q. Model is built with all possible combinations (27*27) and corresponding AIC values are calculated. The lower the AIC value, greater will be the performance of the model, So the combination with which the model's AIC is least is selected. Data is split into train and test, and Sarima model with the best combination of order and seasonal order is fit on train data and its predictions are compared with the test data. Root mean square error (RMSE) is used to measure the accuracies of the models.

2.2.2 Daily Confirmed

For Daily Confirmed the least AIC was found at (0,2,2)*(2,2,1,7) order of SARIMA. Summary of the model is shown in *table (1).* Forecasted values are plotted against the actual values which is in *fig(2.2.2.1)*.the residuals are calculated and plotted in *fig(2.2.2.2)*.RMSE of the model is calculated to be 4364.24.

Dep. Variable:	Daily Confirmed	**No. Observations:**	147
Model:	SARIMAX(0, 2, 2)x(2, 2, 1, 7)	**Log Likelihood**	-976.596
Date:	Fri, 24 Jul 2020	**AIC**	1965.191
Time:	17:21:24	**BIC**	1982.443
Sample:	01-30-2020	**HQIC**	1972.201
	- 06-24-2020		
Covariance Type:	opg		

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
ma.L1	-1.3557	0.112	-12.126	0.000	-1.575	-1.137
ma.L2	0.3557	0.072	4.952	0.000	0.215	0.497
ar.S.L7	-0.8553	0.070	-12.139	0.000	-0.993	-0.717
ar.S.L14	-0.5913	0.077	-7.724	0.000	-0.741	-0.441
ma.S.L7	-0.6758	0.081	-8.360	0.000	-0.834	-0.517
sigma2	1.505e+05	8.98e-07	1.67e+11	0.000	1.5e+05	1.5e+05

Ljung-Box (Q):	59.77	**Jarque-Bera (JB):**	37.07
Prob(Q):	0.02	**Prob(JB):**	0.00
Heteroskedasticity (H):	1915.75	**Skew:**	0.02
Prob(H) (two-sided):	0.00	**Kurtosis:**	5.61

Table 1 - Summary of SARIMA model built on train data of Daily Confirmed

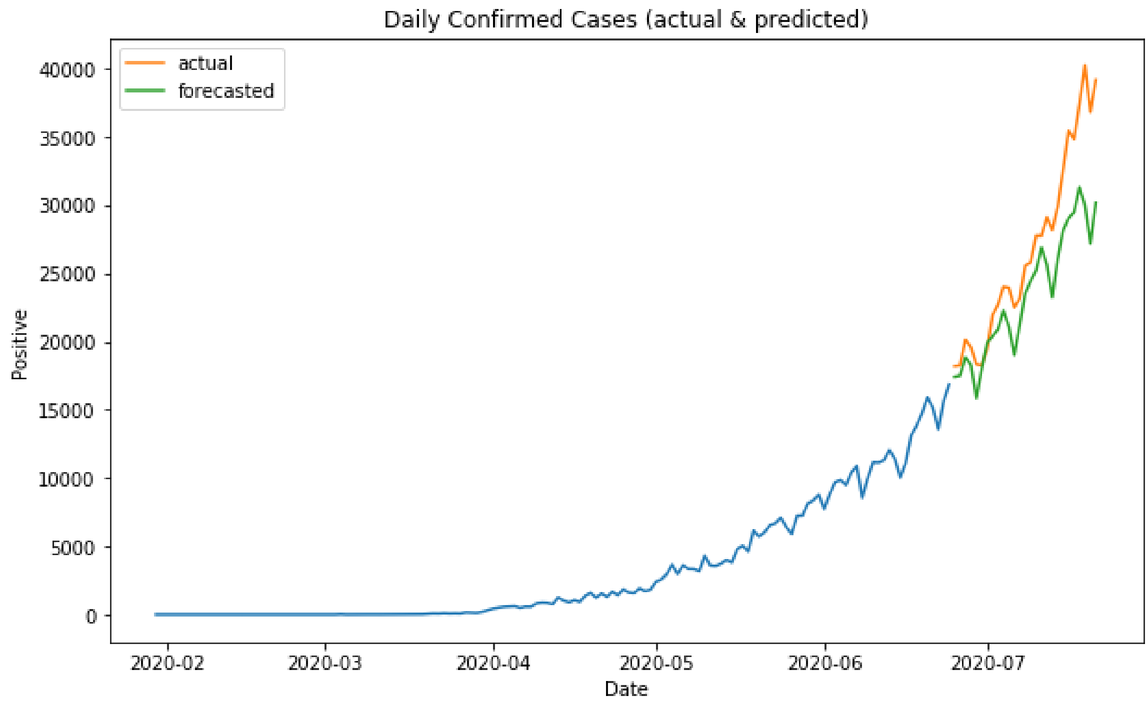

Figure 2.2.2.1 - Daily Confirmed Cases (actual and predicted)

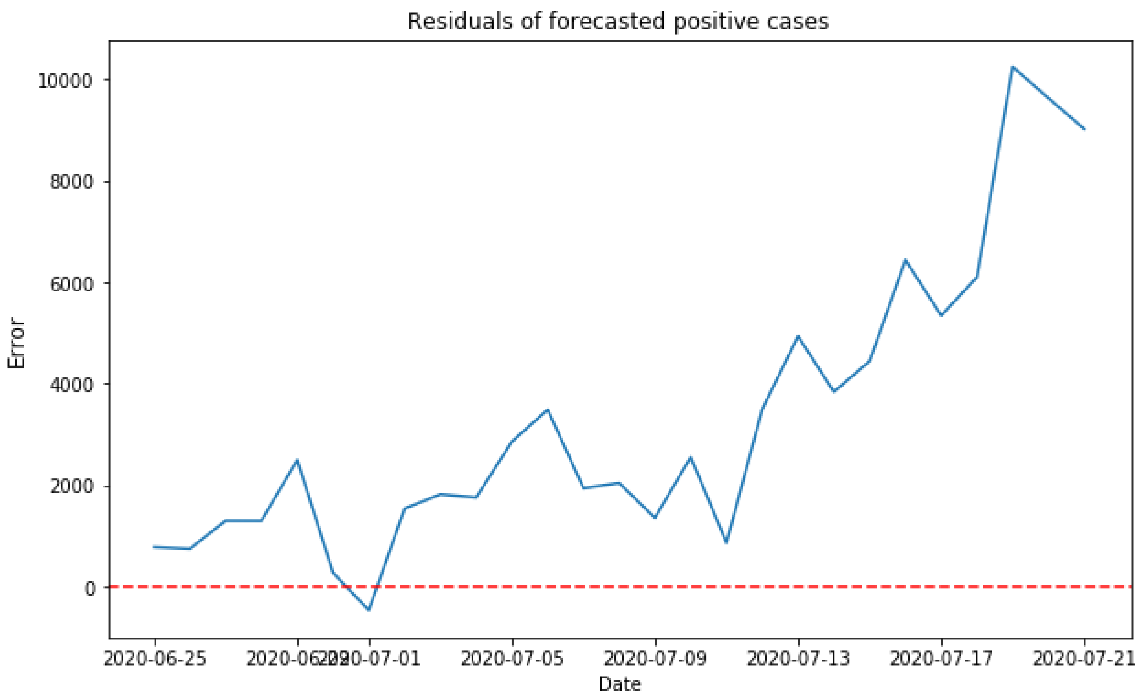

Figure 2.2.2.2 - Residuals of forecasted Daily confirmed cases

2.2.3 Daily Recovered

For Daily Recovered the least AIC was found at (1,2,1)*(0,2,2,7) order of SARIMA. Summary of the model is shown in *table (2).* Forecasted values are plotted against the actual values which is in *fig(2.2.3.1)*.the residuals are calculated and plotted in *fig(2.2.3.2)*.RMSE of the model is calculated to be 739.47

Dep. Variable:	Daily Recovered	No. Observations:	147
Model:	SARIMAX(1, 2, 1)x(0, 2, 2, 7)	Log Likelihood	-967.831
Date:	Fri, 24 Jul 2020	AIC	1945.662
Time:	17:14:57	BIC	1960.038
Sample:	01-30-2020	HQIC	1951.503
	- 06-24-2020		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.6544	0.062	-10.617	0.000	-0.775	-0.534
ma.L1	-1.0000	0.283	-3.537	0.000	-1.554	-0.446
ma.S.L7	-1.8964	0.350	-5.424	0.000	-2.582	-1.211
ma.S.L14	0.9475	0.391	2.423	0.015	0.181	1.714
sigma2	1.058e+05	2.67e-06	3.96e+10	0.000	1.06e+05	1.06e+05

Ljung-Box (Q):	38.49	Jarque-Bera (JB):	4946.81
Prob(Q):	0.54	Prob(JB):	0.00
Heteroskedasticity (H):	330847.09	Skew:	4.30
Prob(H) (two-sided):	0.00	Kurtosis:	31.85

Table 2 - Summary of SARIMA model built on train data of Daily Recovered

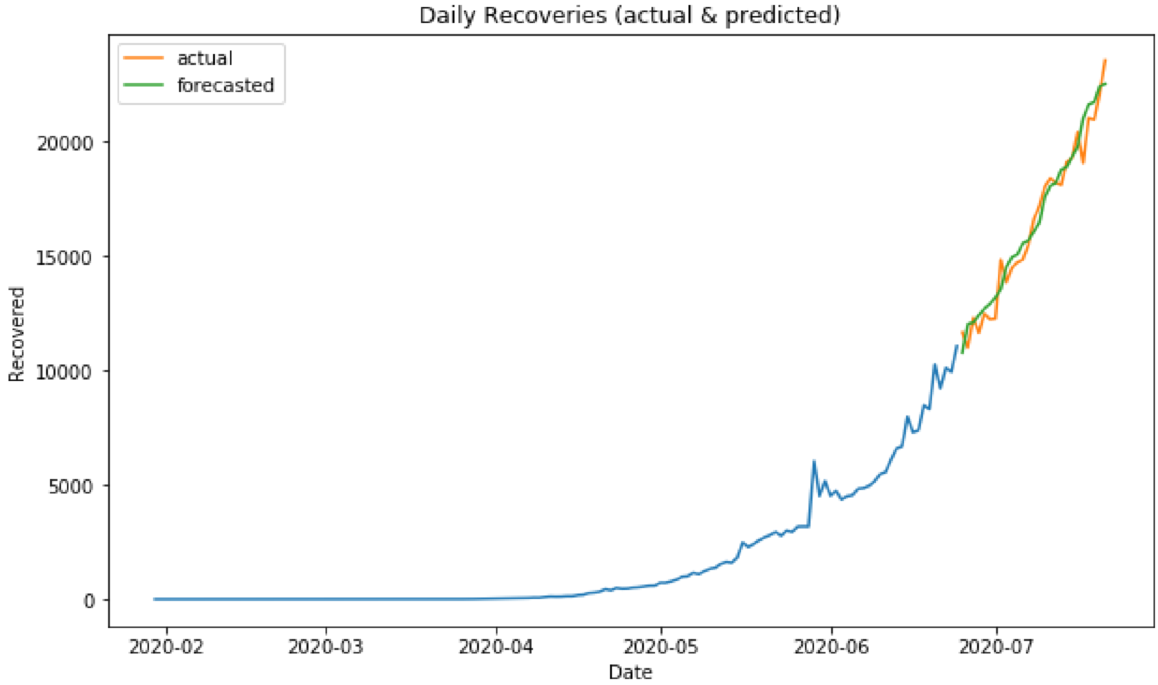

Figure 2.2.3.1 - Daily Recovered (actual and predicted)

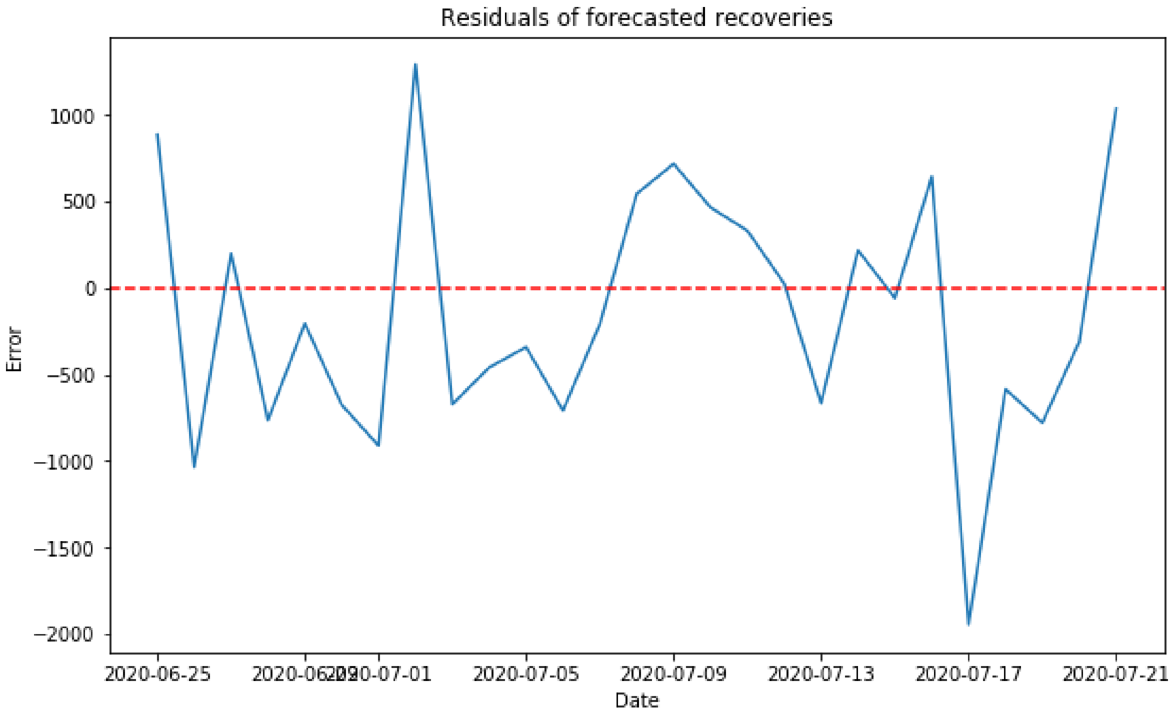

Figure 2.2.3.2 - Residuals of forecasted Daily Recovered

2.2.4 Daily Deceased

For Daily Deceased the least AIC was found at (2,1,0)*(1,2,2,7) order of SARIMA. Summary of the model is shown in *table (3)*. Forecasted values are plotted against the actual values which is in *fig(2.2.4.1)*.the residuals are calculated and plotted in *fig(2.2.4.2)*. RMSE of the model is calculated to be 130.27.

Dep. Variable:	Daily Deceased	**No. Observations:**	137
Model:	SARIMAX(2, 1, 0)x(1, 2, 2, 7)	**Log Likelihood**	-265.978
Date:	Fri, 24 Jul 2020	**AIC**	543.957
Time:	19:13:15	**BIC**	560.781
Sample:	01-30-2020	**HQIC**	550.790
	- 06-14-2020		
Covariance Type:	opg		

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
ar.L1	0.0604	0.070	0.858	0.391	-0.078	0.198
ar.L2	0.5370	0.063	8.483	0.000	0.413	0.661
ar.S.L7	-0.0969	0.230	-0.421	0.674	-0.548	0.354
ma.S.L7	-1.2586	0.216	-5.828	0.000	-1.682	-0.835
ma.S.L14	0.4610	0.270	1.705	0.088	-0.069	0.991
sigma2	4.0511	0.259	15.613	0.000	3.543	4.560

Ljung-Box (Q):	49.33	**Jarque-Bera (JB):**	382.07
Prob(Q):	0.15	**Prob(JB):**	0.00
Heteroskedasticity (H):	2064.45	**Skew:**	1.55
Prob(H) (two-sided):	0.00	**Kurtosis:**	11.10

Table 3 - Summary of SARIMA model built on train data of Daily Deceased

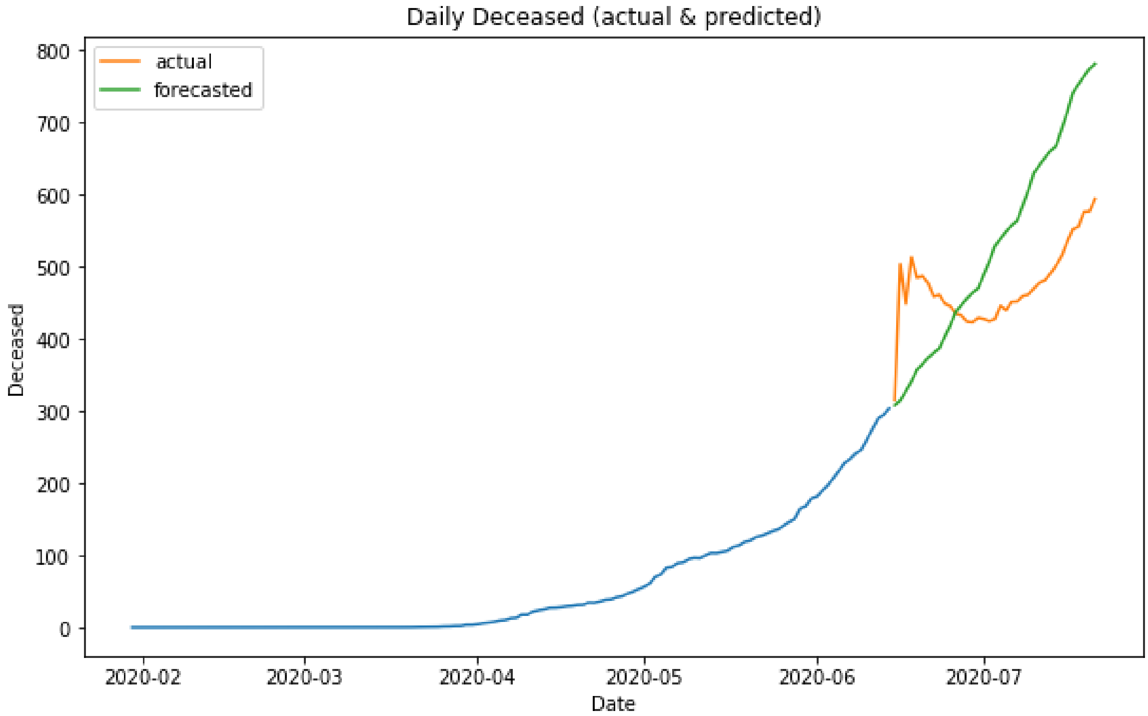

Figure 2.2.4.1 - Daily Deceased (actual and predicted)

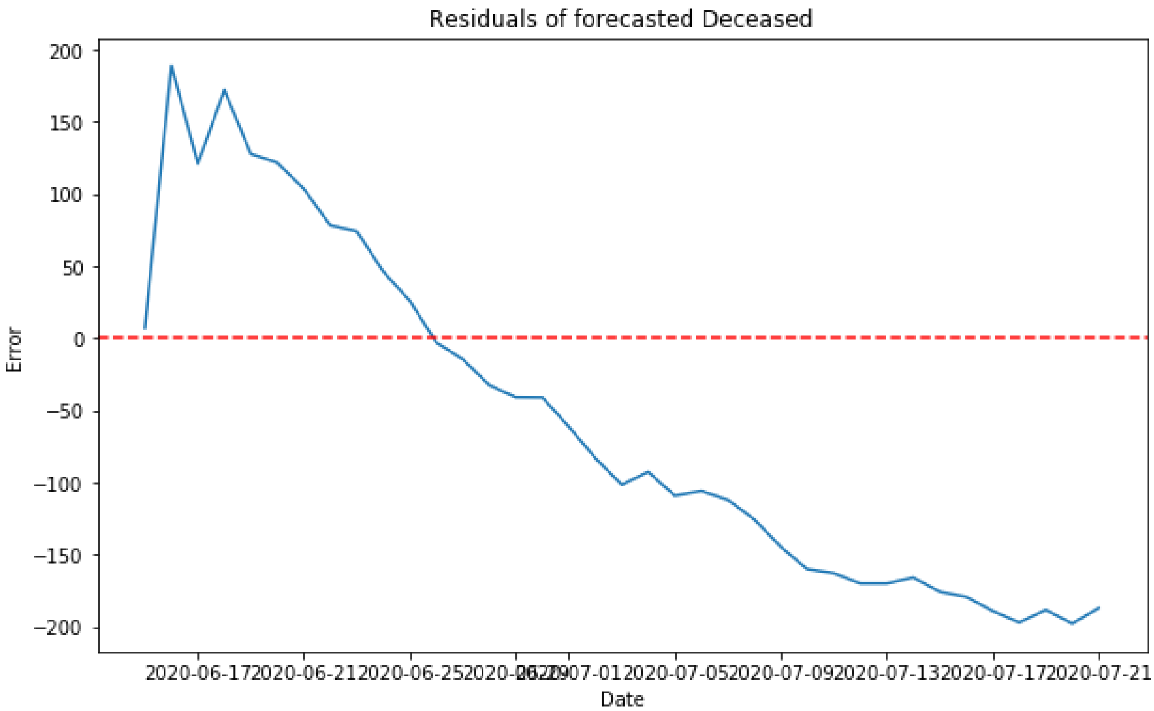

Figure 2.2.4.2 - Residuals of forecasted Daily Deceased

3. RESULTS

Finally, The entire available data (30^{th} Jan to 21^{st} July) of variables (Daily Confirmed, Daily Recovered, Daily Deceased) is used and SARIMA model is built with best order and seasonal order found for each variable previously. The values of variables are forecasted from 22^{nd} July to 8^{th} August. Cumulative Confirmed, Cumulative Recovered and Cumulative Deceased are calculated from their forecasted daily numbers.

3.1 Daily Confirmed

Summary of the SARIMA (0,2,2)*(2,2,1,7) built on entire data is shown table(4).the forecasted values with 95 % Confidence interval is shown in fig(3.1.1) where dotted lines represent upper and lower limit.

Dep. Variable:	Daily Confirmed	No. Observations:	174
Model:	SARIMAX(0, 2, 2)x(2, 2, 1, 7)	Log Likelihood	-1240.619
Date:	Fri, 24 Jul 2020	AIC	2493.237
Time:	09:58:07	BIC	2511.613
Sample:	01-30-2020	HQIC	2500.700
	- 07-21-2020		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ma.L1	-1.4208	0.046	-31.120	0.000	-1.510	-1.331
ma.L2	0.4544	0.052	8.720	0.000	0.352	0.557
ar.S.L7	-0.6621	0.066	-9.996	0.000	-0.792	-0.532
ar.S.L14	-0.3927	0.094	-4.158	0.000	-0.578	-0.208
ma.S.L7	-0.9115	0.065	-14.027	0.000	-1.039	-0.784
sigma2	3.46e+05	2.77e+04	12.484	0.000	2.92e+05	4e+05

Ljung-Box (Q):	43.67	Jarque-Bera (JB):	65.38
Prob(Q):	0.32	Prob(JB):	0.00
Heteroskedasticity (H):	414.77	Skew:	0.08
Prob(H) (two-sided):	0.00	Kurtosis:	6.15

Table 4 - Summary of SARIMA model built on entire data of Daily Confirmed

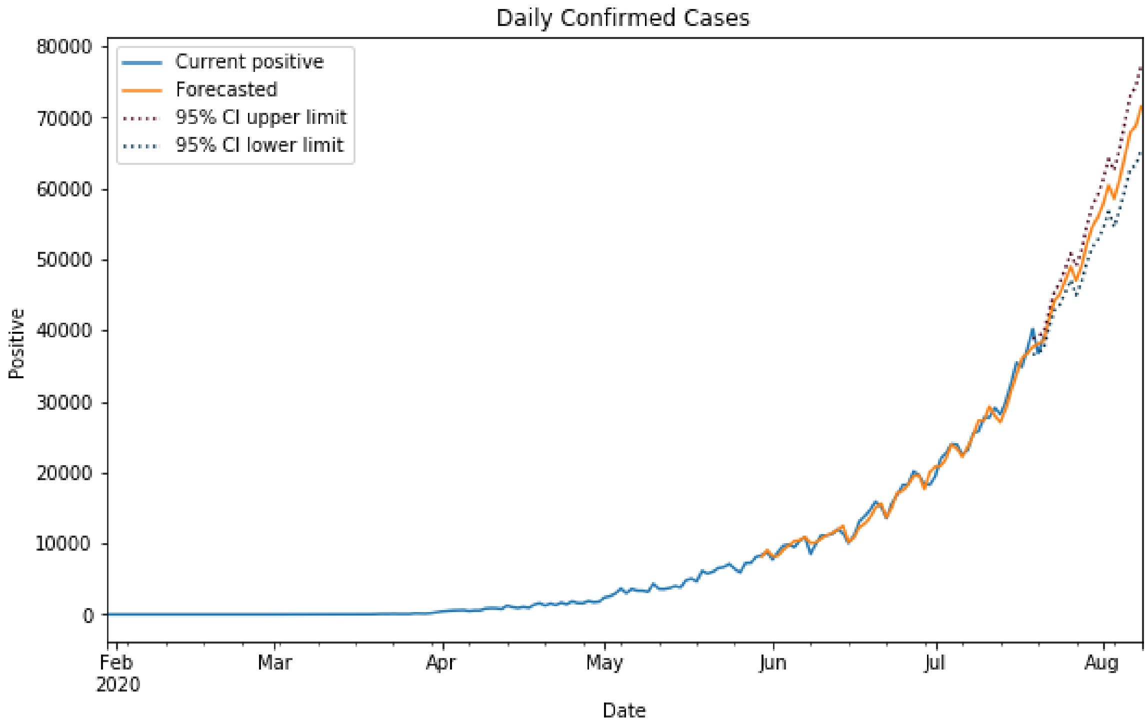

Figure 3.1.1- Forecast of Daily Confirmed from 22nd July to 8th August with 95% CI

3.2 Cumulative Confirmed

Cumulative Confirmed is calculated from forecasted daily confirmed in 3.1.the forecast of cumulative curve for confirmed in shown in fig(3.2.1) with 95 % confidence interval, where dotted lines represent upper and lower limits

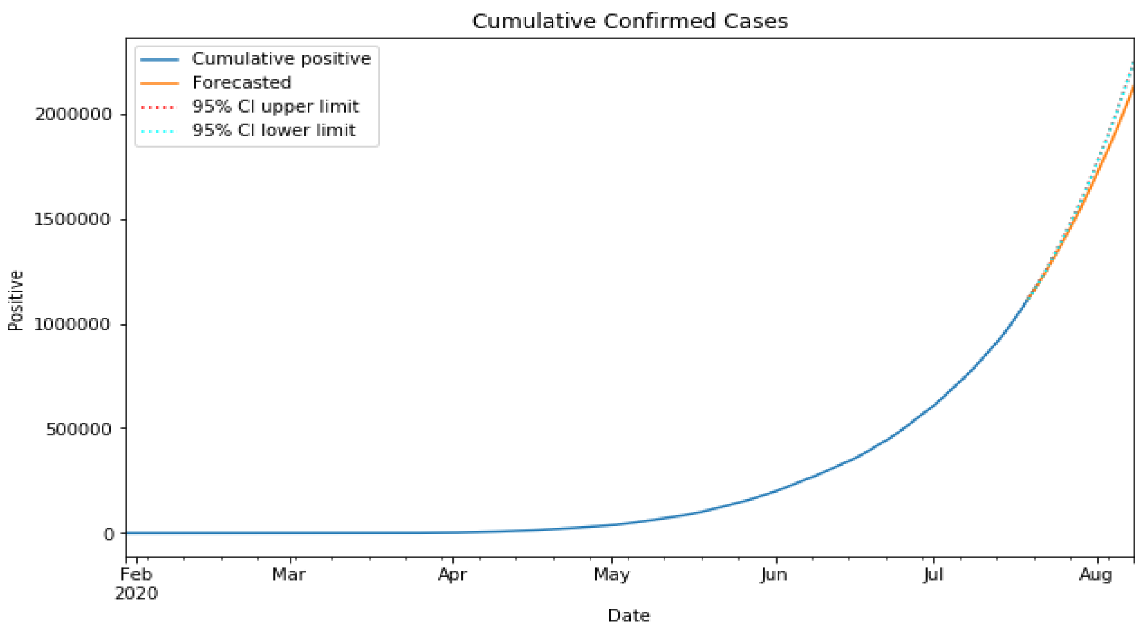

Figure 3.2.1 - Forecast of Cumulative Confirmed from 22nd July to 8th August with 95% CI

3.3 Daily Recovered

Summary of the SARIMA (1,2,1)*(0,2,2,7) built on entire data is shown table(5).the forecasted values with 95 % Confidence interval is shown in fig(3.3.1) where dotted lines represent upper and lower limit.

Dep. Variable:	Daily Recovered	No. Observations:	174
Model:	SARIMAX(1, 2, 1)x(0, 2, 2, 7)	Log Likelihood	-1203.918
Date:	Fri, 24 Jul 2020	AIC	2417.835
Time:	17:46:45	BIC	2433.148
Sample:	01-30-2020	HQIC	2424.054
	- 07-21-2020		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.5546	0.063	-8.773	0.000	-0.679	-0.431
ma.L1	-1.0000	0.072	-13.879	0.000	-1.141	-0.859
ma.S.L7	-1.6739	0.053	-31.369	0.000	-1.779	-1.569
ma.S.L14	0.7059	0.051	13.935	0.000	0.607	0.805
sigma2	1.966e+05	3.67e-07	5.36e+11	0.000	1.97e+05	1.97e+05

Ljung-Box (Q):	60.08	Jarque-Bera (JB):	707.80
Prob(Q):	0.02	Prob(JB):	0.00
Heteroskedasticity (H):	89331.99	Skew:	1.59
Prob(H) (two-sided):	0.00	Kurtosis:	12.87

Table 5 - Summary of SARIMA model built on entire data of Daily Recovered

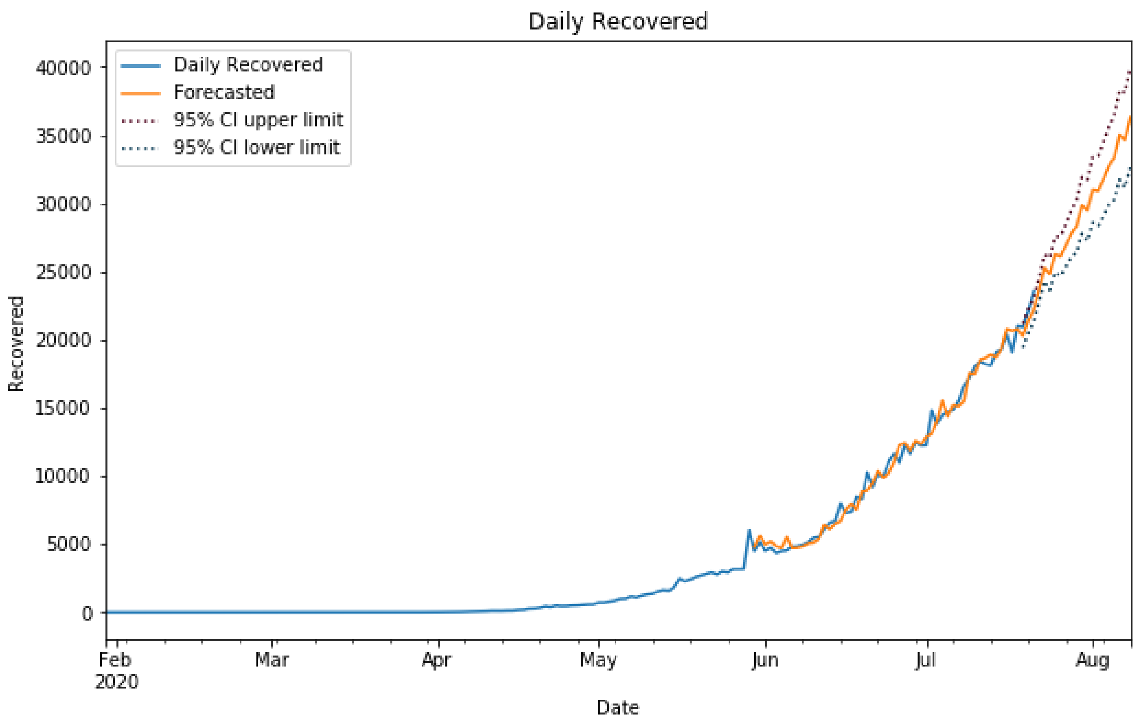

Figure 3.3.1- Forecast of Daily Recovered from 22nd July to 8th August with 95% CI

3.4 Cumulative Recovered

Cumulative Recovered is calculated from forecasted daily recovered in 3.3.the forecast of cumulative curve for confirmed in shown in *fig(3.4.1)* with 95 % confidence interval, where dotted lines represent upper and lower limit.

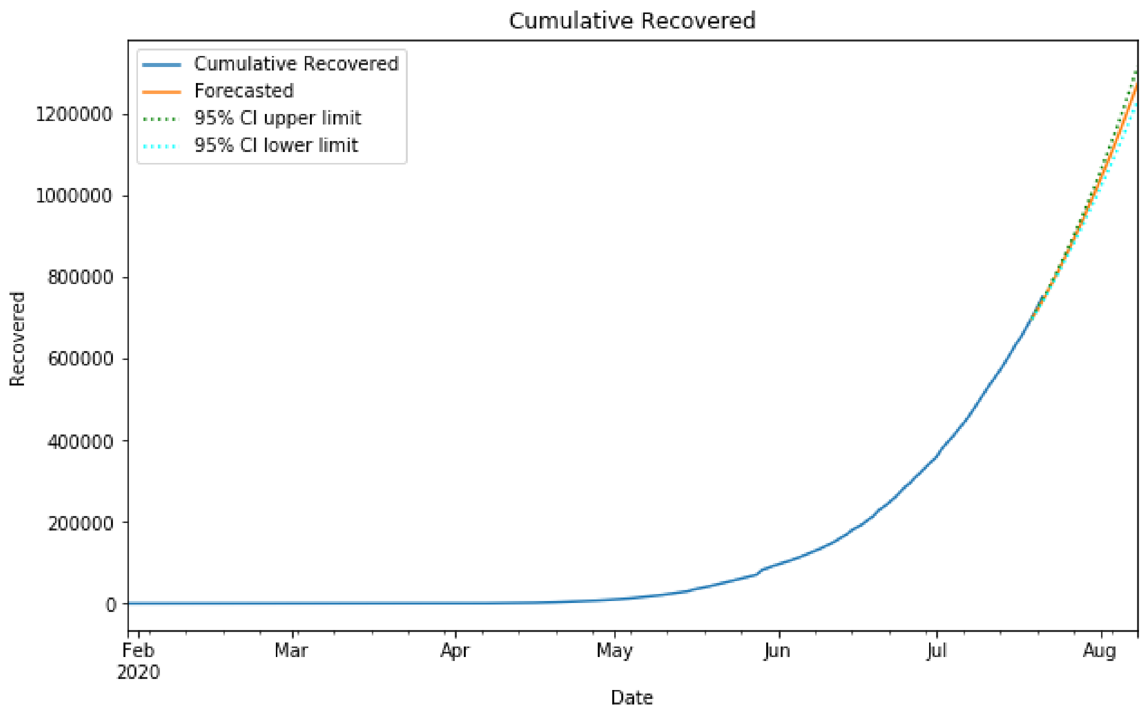

Figure 3.4.1 - Forecast of Cumulative Recovered from 22nd July to 8th August with 95% CI

3.5 Daily Deceased

Summary of the SARIMA (2,1,0)*(1,2,2,7) built on entire data is shown *table(6)*.the forecasted values with 95 % Confidence interval is shown in *fig(3.5.1)* where dotted lines represent upper and lower limit.

Dep. Variable:	Daily Deceased	**No. Observations:**	174
Model:	SARIMAX(2, 1, 0)x(1, 2, 2, 7)	**Log Likelihood**	-689.849
Date:	Fri, 24 Jul 2020	**AIC**	1391.697
Time:	19:16:57	**BIC**	1410.111
Sample:	01-30-2020	**HQIC**	1399.175
	- 07-21-2020		
Covariance Type:	opg		

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
ar.L1	-0.1983	0.050	-3.940	0.000	-0.297	-0.100
ar.L2	0.2864	0.117	2.455	0.014	0.058	0.515
ar.S.L7	0.0186	0.101	0.184	0.854	-0.180	0.217
ma.S.L7	-1.9802	6.947	-0.285	0.776	-15.595	11.635
ma.S.L14	0.9981	6.964	0.143	0.886	-12.652	14.648
sigma2	231.2479	1626.853	0.142	0.887	-2957.325	3419.821

Ljung-Box (Q):	18.40	**Jarque-Bera (JB):**	67325.88
Prob(Q):	1.00	**Prob(JB):**	0.00
Heteroskedasticity (H):	7365.49	**Skew:**	8.82
Prob(H) (two-sided):	0.00	**Kurtosis:**	102.25

Table 6 - Summary of SARIMA model built on entire data of Daily Deceased

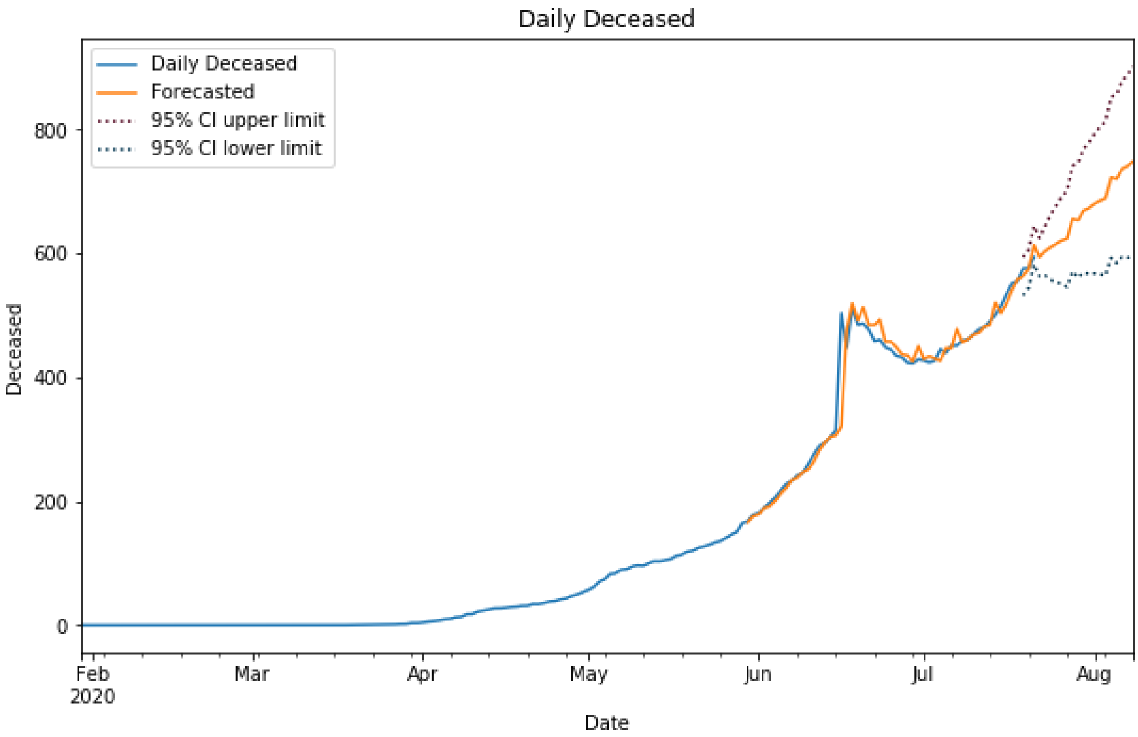

Figure 3.5.1- Forecast of Daily Deceased from 22nd July to 8th August with 95% CI

3.6 Cumulative Deceased

Cumulative Recovered is calculated from forecasted daily deceased in 3.5.the forecast of cumulative curve for confirmed in shown in *fig(3.6.1)* with 95 % confidence interval, where dotted lines represent upper and lower limit.

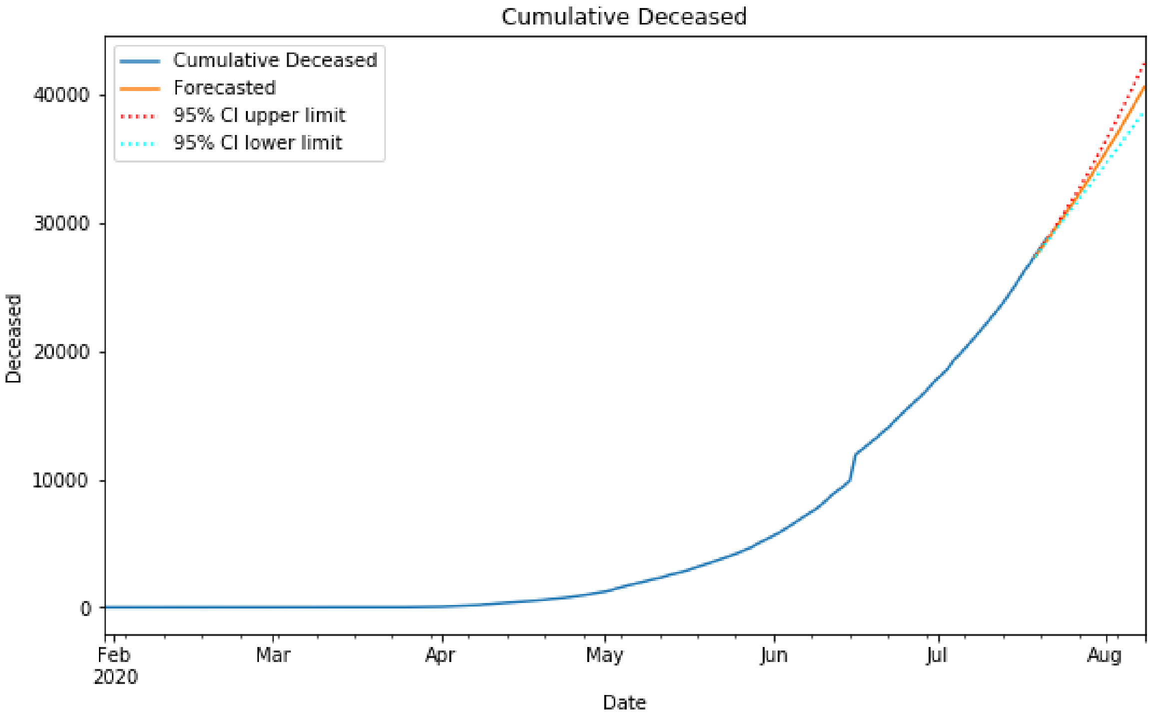

Figure 3.6.1 - Forecast of Cumulative Recovered from 22nd July to 8th August with 95% CI

3.7 COVID -19 Situation by 31st July

From the forecasted cumulative confirmed , cumulative recovered and cumulative deceased, percentages of active, recovered and deceased in total number of infected cases by end of july are calculated as shown *fig(3.7.1)*.Composition of recovered and deceased in cases with outcome is shown in *fig(3.7.2)* .

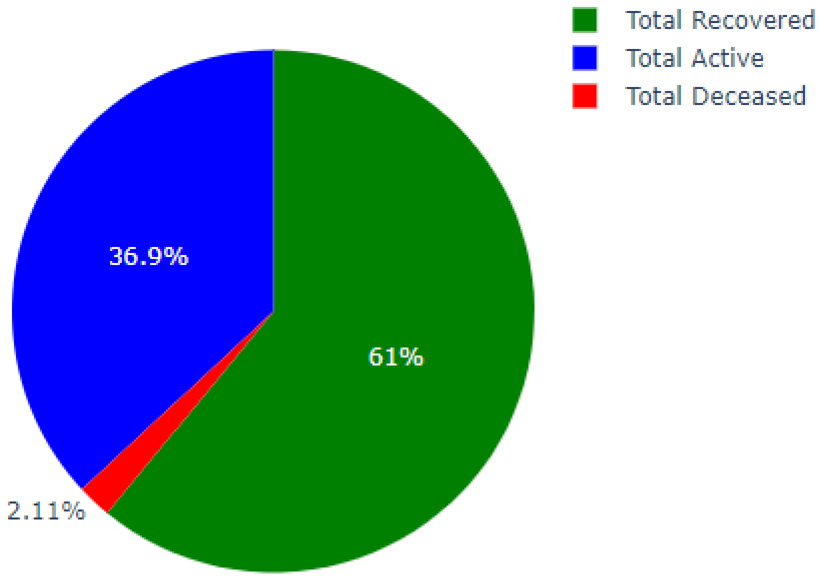

Figure 3.7.1 – COVID-19 Cases by end of July (recovered -10,09,786 active – 6,11,743 deceased – 34,903)

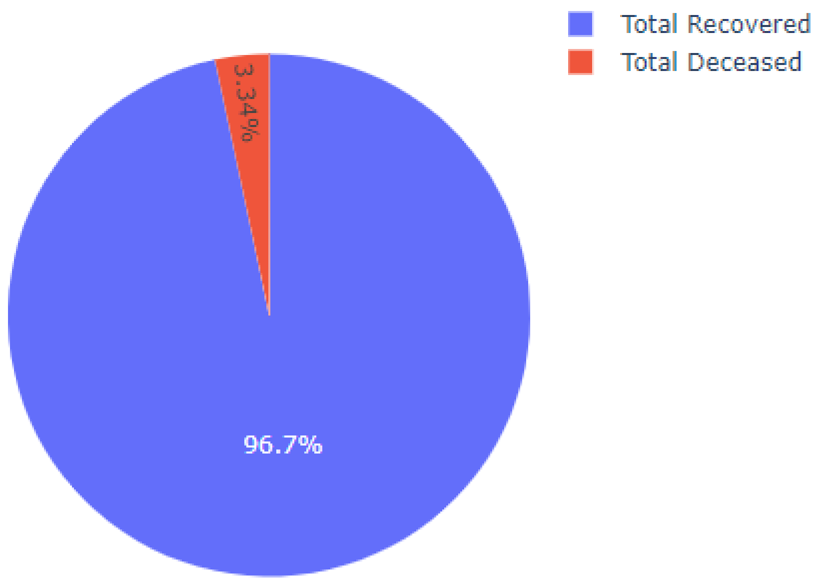

Figure 3.7.2 – COVID-19 Cases by end of July (recovered -10,09,786 deceased – 34,903)

3.8 COVID -19 Situation by 8th August

Similarly as mentioned in 3.7, percentages of active, recovered and deceased in total number of infected cases by 8th August are calculated as shown *fig(3.8.1)*.Composition of recovered and deceased in cases with outcome is shown in *fig(3.8.2)* .

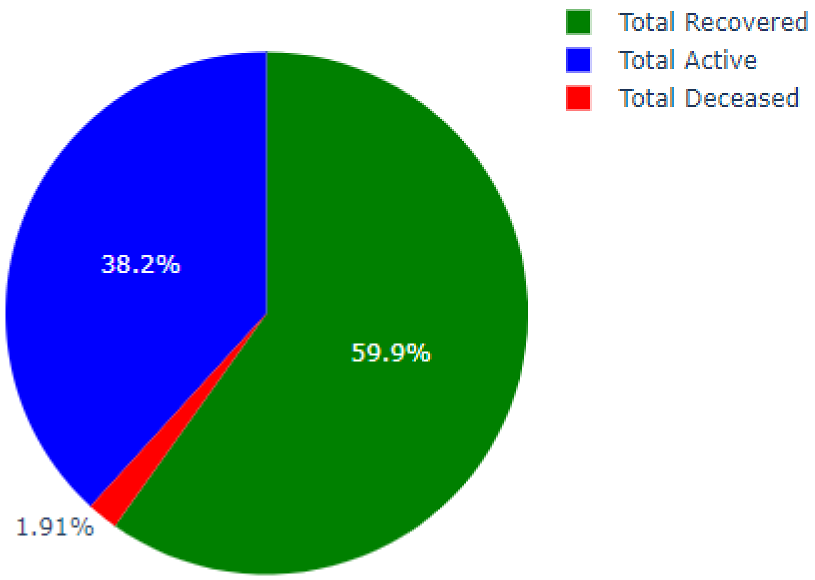

Figure 3.8.1 – COVID-19 Cases by 8th of August (recovered -12,75,437 active – 8,14,379 deceased – 40,629)

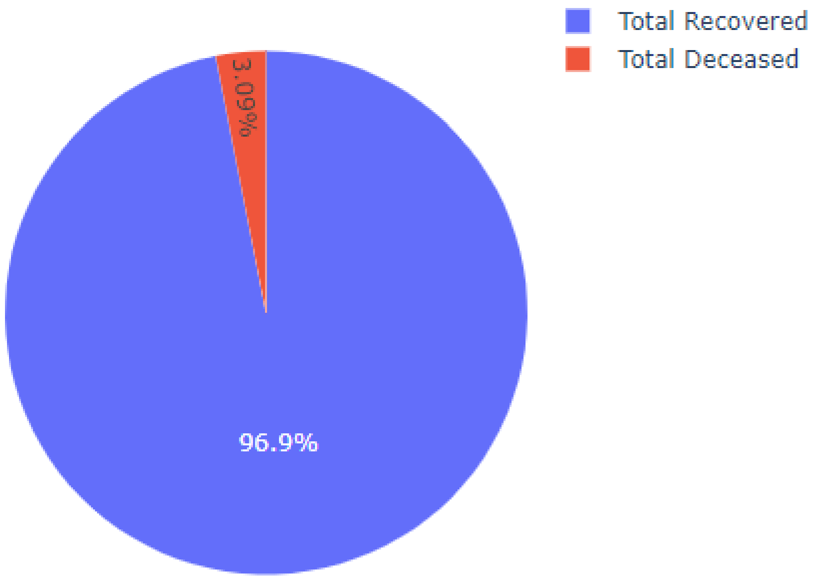

Figure 3.8.2 – COVID-19 Cases by 8th of August (recovered -12,75,437 deceased – 40,629)

4. CONCLUSION

Hence using the past data of confirmed cases, fatalities and recoveries from COVID-19 dated January 30 2020 to July 21 2020, we were able to predict the probable number of confirmed cases, fatalities and recoveries on daily basis and cumulative basis. From the above results, conclusions, charts and trends we are able to observe clearly that by the end of July, we may expect around 6,11,743 active cases of COVID-19 with 10,09,786 people being recovered and an unfortunate 34,903 people being deceased by this disease. Another prediction on August 8^{th} clearly shows that around 8,14,379 people are likely to be tested positive for the disease, with 12,75,437 people being recovered from this disease and around 40,627 people being deceased by this disease. These numbers are alarming especially the active cases, as it poses a high risk of systemic health care failure in India. Finally, the model is as good as the underlying data. Because of real time change in data daily, the predictions will accordingly change. Therefore, the results from this study can be used only for qualitative understanding and reasonable estimate of the pandemic.

5. REFERENCES

[1] Rajesh Ranjan. Predictions for covid-19 outbreak in India using epidemiological models medRxiv preprint doi: https://doi.org/10.1101/2020.04.02.20051466

[2] Anastassopoulou C, Russo L, Tsakris A, Siettos C (2020) Data-base danalysis, modelling and forecasting of the COVID-19 outbreak. PLoS ONE 15(3): e0230405. https://doi.org/10.1371/journal.pone.0230405

[3] Wu P, Hao X, Lau EHY, Wong JY, Leung KSM, Wu JT, et al. Real-time tentative assessment of the epidemiological characteristics of novel coronavirus infections in Wuhan, China, as at 22 January 2020.Eurosurveillance. 2020;25(3). https://doi.org/10.2807/1560-7917.ES.2020.25.3.2000044

[4] Soudeep Deb, Manidipa Majumdar. A time series method to analyze incidence pattern and estimate reproduction number of COVID-19. arXiv:2003.10655v1 [stat.AP] 24 Mar 2020

[5] Hiteshi Tandon, Prabhat Ranjan, Tanmoy Chakraborty, Vandana Suhag. Coronavirus (COVID-19): ARIMA based time-series analysis to forecast near future.

[6] Petropoulos F, Makridakis S (2020) Forecasting the novel coronavirus COVID-19. PLoS ONE 15(3): e0231236. https://doi.org/10.1371/journal.pone.0231236

[7] Artificial intelligence and machine learning to fight COVID-19 https://journals.physiology.org/doi/full/10.1152/physiolgenomics.00029.2020

[8]Covid19 article by WHO

https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200327-sitrep-67-covid-19.pdf?sfvrsn=b65f68eb_4