Spread Visualization and Prediction of the Novel Coronavirus Disease COVID-19 Using Machine Learning

*Zahra Taheri**

July 2020

Abstract

The rapid spread of the novel coronavirus disease 2019 (COVID-19) has become a health challenge worldwide. At this time, spread forecasting using Artificial intelligence and Machine learning methods is an important task to track the growth of the pandemic. The main focus of this project is to visualize the spreading of the virus country-wise as well as globally, and to perform Linear regression, Support vector machine, Ensemble methods, Multilayer perceptron, Recurrent neural network-LSTM, ARIMA and Prophet on the COVID-19 data to forecast the future effects of the pandemic. Moreover, we are going to study the impact of some new parameters such as population statistics and life expectancy, etc., in prediction of COVID-19 spread.

Keywords: COVID-19, Machine learning, Pandemic, Predictive modeling, Visualization.

1 Introduction

The current destructive pandemic of coronavirus disease 2019 (COVID-19), caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) [13], was first reported in Wuhan, China, in December 2019 [14, 6]. The outbreak has affected millions of people around the world and the number of infections and mortalities has been growing at an alarming rate. As of date, confirmed COVID-19 cases are more than 15 millions in 187 countries. In such a situation, forecasting and proper study of the pattern of disease spread can inspire design

*Email Address: `zahra.taheri518@gmail.com`

Github Repository: `https://github.com/zata213/Applied_Machine_Learning_S20_Final_Project` better strategies to make more efficient decisions. Moreover, such studies play an important role in achieving accurate predictions.

Machine learning has numerous tools that can be used for visualization and prediction, and nowadays it is used worldwide to study the pattern of COVID-19 spread, e.g., see [4, 7, 8, 15, 11]. One of the main focus of the study in this project is to use machine learning techniques to analyse and visualize the spreading of the virus country-wise as well as globally during a specific period of time by considering confirmed cases, recovered cases and fatalities.

The global impact of the outbreak on various aspects of life has been the focus of many studies, e.g., see [12, 1, 5, 3]. On the other hand, a pandemic can be forecast by considering a variety of parameters such as the impact of environmental factors, quarantine, age, gender and a lot more, e.g., see [2, 9, 10].

The forecasting accuracy depends on the availability of proper data to base its predictions and provide an estimate of uncertainty. A challenge to use machine learning techniques for the current outbreak is that the datasets are not yet standardized by any standardization organization and the statistical anomalies are not considered. Also, the appropriate selection of parameters and the selection of the best machine learning model for prediction are other challenges involved in training a model.

In this project, we are going to perform Linear regression, Support vector machine, Ensemble methods, Multilayer perceptron, Recurrent neural network-LSTM, ARIMA and Prophet, etc., on the Johns Hopkins University's COVID-19 data to anticipate the future effects of COVID-19 pandemic in the world, Iran and some other countries. Moreover, we are going to study the impact of some other parameters such as environmental factors, life expectancy, population statistics, etc., in prediction of COVID-19 spread.

2 Experimental data and results

The data is provided by the Johns Hopkins University Center for Systems Science and Engineering(JHU-CSSE) and contains three time series with the number of reported daily confirmed cases, recovered cases and deaths by country. This dataset is updated automatically on daily basis.

In this project we employed data from 22 January 2020 up to 21 July 2020. Initially, data preprocessing was almost challenging and much time was required because the dataset was not standard and many data cleaning processes were required. This part was done carefully and some appropriate dataframes were prepared, such as follows,

	Date	Country/Region	Confirmed	Deaths	Recovered	Active	New confirmed	New deaths	New recovered	WHO region
34403	2020-07-23	West Bank and Gaza	9744	67	2720	6957	346	1	770	EMRO
34404	2020-07-23	Western Sahara	10	1	8	1	0	0	0	AFRO
34405	2020-07-23	Yemen	1654	461	762	431	14	3	11	EMRO
34406	2020-07-23	Zambia	3789	134	1677	1978	206	6	0	AFRO
34407	2020-07-23	Zimbabwe	2124	28	510	1586	90	2	0	AFRO

	Country/Region	Confirmed	Deaths	Recovered	Active	New confirmed	New deaths	New recovered	Recovery rate(per 100)	Mortality rate(per 100)	WHO region
0	Afghanistan	35928	1211	24550	10167	201	21	626	68.33	3.37	EMRO
1	Albania	4466	123	2523	1820	108	3	60	56.49	2.75	EURO
2	Algeria	25484	1124	17369	6991	612	13	386	68.16	4.41	AFRO
3	Andorra	889	52	803	34	0	0	0	90.33	5.85	EURO
4	Angola	851	33	236	582	39	0	15	27.73	3.88	AFRO

	Date	Confirmed	Deaths	Recovered	Active	New confirmed	New deaths	New recovered	Recovery rate(per 100)	Mortality rate(per 100)	Number of countries
0	2020-01-22	555	17	28	510	0	0	0	5.05	3.06	6
1	2020-01-23	654	18	30	606	99	1	2	4.59	2.75	8
2	2020-01-24	941	26	36	879	287	8	6	3.83	2.76	9
3	2020-01-25	1434	42	39	1353	493	16	3	2.72	2.93	11
4	2020-01-26	2118	56	52	2010	684	14	13	2.46	2.64	13

After exploring the data, we performed some visualizations on the data in order to get a better understanding of the data and how the pandemic is affecting all of us. For example, in Figure 1, we can see the latest status of cases in the world.

	Confirmed	Recovered	Deaths	Active	Recovery rate(per 100)	Mortality rate(per 100)	Number of countries
183	15510436.00	8710976.00	633381.00	6166079.00	56.16	4.08	187.00

Figure 1

Also in Figure 2, we can see that the latest global recovery rate per 100 cases is 56.16 whereas the mortality rate per 100 cases is 4.08, that is a good news because at the start point of this project, the recovery rate was around 54 percent whereas the mortality rate was around 5 percent.

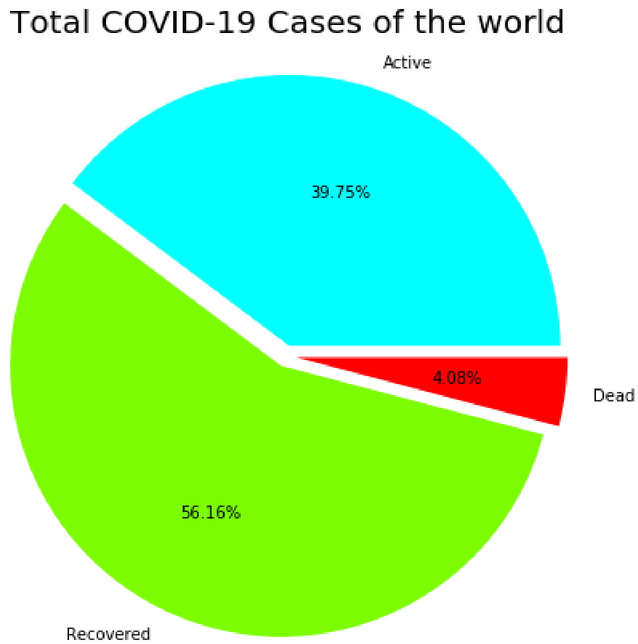

Figure 2

Also as an example, Figure 3 shows comparisons between the latest COVID-19 cases status of 10 most affected countries, i.e., US, Brazil, India, Russia, Peru, South Africa, Mexico, Chile, United Kingdom, and Iran.

Some of our conclusions based on the analysis from the above observations and some others, which can be found in the project's Github repository, are as follows:

1. Even though the total number of confirmed cases and deaths in the world are monotonically (almost exponentially) increasing, the recovery rate shows some increase whereas the mortality rate shows some decrease.

2. Although US has shown the greatest rise in the number of confirmed cases and deaths, its death curve is flattening.

3. Between 10 most affected countries, Brazil shows the greatest rise in the number of recovered cases, whereas United Kingdom shows very few recoveries.

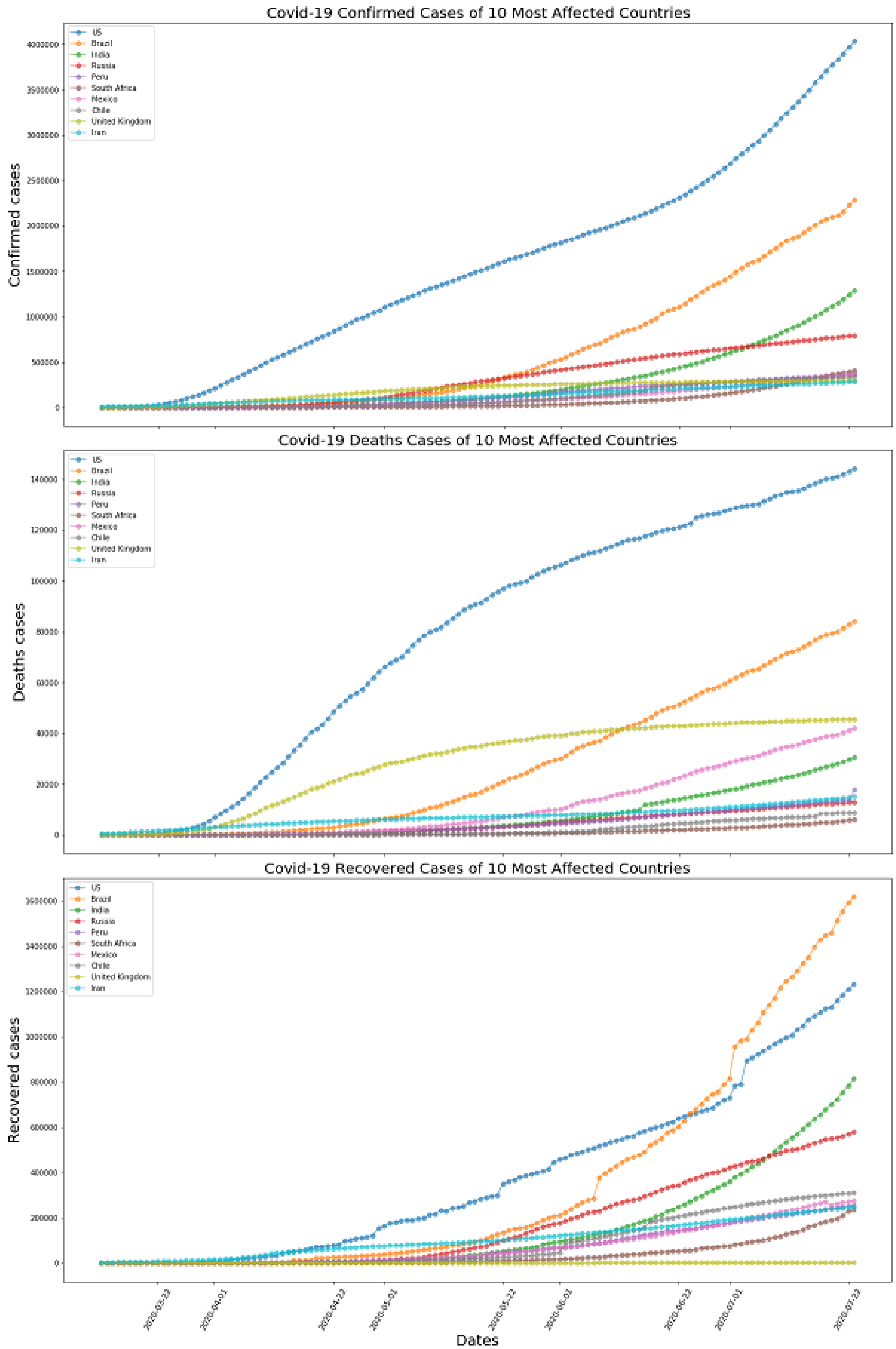
Covid-19 Confirmed Cases of 10 Most Affected Countries
US
Brazil
India
Russia
Peru
South Africa
Mexico
Chile
United Kingdom
Iran
Confirmed cases
Covid-19 Deaths Cases of 10 Most Affected Countries
US
Brazil
India
Russia
Peru
South Africa
Mexico
Chile
United Kingdom
Iran
Deaths cases
Covid-19 Recovered Cases of 10 Most Affected Countries
US
Brazil
India
Russia
Peru
South Africa
Mexico
Chile
United Kingdom
Iran
Recovered cases
Dates

Figure 3

After visualization, we investigated data modeling and prediction based on univariate time series, using Linear regression, Support vector machine, Random forests, XGBoost, Multilayer perceptron (MLP), and a recurrent neural network, Long Short-Term Memory network (LSTM-RNN) to forcast the number of confirmed cases and deaths in the world and some other countries such as Iran. Some of our results are summarized in the following tables:

Table 1: Prediction errors of total confirmed cases of the world

Regressor	RMSE
Support vector machine	84855.25
Linear regression	880383.54
Random Forests	3254934.58
XGBoost	3172578.3

Table 2: Prediction errors of total deaths of the world

Regressor	RMSE
Support vector machine	140122.16
Linear regression	19337.55

Table 3: Accuracy of predicting the total cases of Iran using MLP and LSTM-RNN

Neural Network	MAPE	Accuracy(in percent)
MLP	0.24124079849664185	99.99758759201504
LSTM-RNN	0.6490237741213257	99.99350976225878

Table 4: Accuracy of predicting the total deaths of Iran using MLP and LSTM-RNN

Neural Network	MAPE	Accuracy(in percent)
MLP	0.10785990390804867	99.99892140096092
LSTM-RNN	1.1043418430881735	99.98895658156911

Table 5: Accuracy of predicting the total cases of the world using MLP and LSTM-RNN

Neural Network	MAPE	Accuracy(in percent)
MLP	0.47257501056634127	99.99527424989434
LSTM-RNN	1.3621032695517894	99.98637896730449

Table 6: Accuracy of predicting the total deaths of the world using MLP and LSTM-RNN

Neural Network	MAPE	Accuracy(in percent)
MLP	0.14795993064612276	99.99852040069354
LSTM-RNN	0.6681058141637979	99.99331894185836

	confirmed	confirmed_predicted
2020-07-12	257303	258255.216509
2020-07-13	259652	260653.501362
2020-07-14	262173	262996.952808
2020-07-15	264561	265338.761990
2020-07-16	267061	267673.333040
2020-07-17	269440	269929.288252
2020-07-18	271606	272183.205840
2020-07-19	273788	274376.329207
2020-07-20	276202	276674.740013
2020-07-21	278827	278935.530784
2020-07-22	NaN	281163.993262
2020-07-23	NaN	283303.834251
2020-07-24	NaN	285419.558273
2020-07-25	NaN	287508.336982
2020-07-26	NaN	289579.659022
2020-07-27	NaN	291621.268229
2020-07-28	NaN	293655.578481

MLP-Iran

	confirmed	confirmed_predicted
2020-07-12	257303	255792.153422
2020-07-13	259652	258122.649222
2020-07-14	262173	260456.064604
2020-07-15	264561	262793.220701
2020-07-16	267061	265128.460822
2020-07-17	269440	267473.828245
2020-07-18	271606	269819.560615
2020-07-19	273788	272159.149698
2020-07-20	276202	274530.793361
2020-07-21	278827	276931.298313
2020-07-22	NaN	279352.240342
2020-07-23	NaN	281908.669240
2020-07-24	NaN	284502.626936
2020-07-25	NaN	287137.945382
2020-07-26	NaN	289818.608592
2020-07-27	NaN	292549.452124
2020-07-28	NaN	295333.547621

LSTM(RNN)-Iran

Figure 4: Iran-Prediction of confirmed cases using Neural Networks

	Deaths	Deaths_predicted
2020-07-12	12829	12823.306296
2020-07-13	13032	13012.394314
2020-07-14	13211	13196.052448
2020-07-15	13410	13391.281282
2020-07-16	13608	13583.963116
2020-07-17	13791	13783.641293
2020-07-18	13979	13982.397670
2020-07-19	14188	14187.267750
2020-07-20	14405	14388.072571
2020-07-21	14634	14596.819634
2020-07-22	NaN	14807.348543
2020-07-23	NaN	15020.703101
2020-07-24	NaN	15235.859085
2020-07-25	NaN	15455.117984
2020-07-26	NaN	15676.330436
2020-07-27	NaN	15901.074949
2020-07-28	NaN	16128.361943

MLP-Iran

	Deaths	Deaths_predicted
2020-07-12	12829	12689.374158
2020-07-13	13032	12867.127951
2020-07-14	13211	13049.990807
2020-07-15	13410	13238.513997
2020-07-16	13608	13433.313561
2020-07-17	13791	13634.541627
2020-07-18	13979	13840.276272
2020-07-19	14188	14053.563050
2020-07-20	14405	14272.026670
2020-07-21	14634	14500.660216
2020-07-22	NaN	14738.160816
2020-07-23	NaN	14993.416612
2020-07-24	NaN	15260.722083
2020-07-25	NaN	15541.136095
2020-07-26	NaN	15835.815415
2020-07-27	NaN	16146.121658
2020-07-28	NaN	16473.514340

LSTM(RNN)-Iran

Figure 5: Iran-Prediction of deaths using Neural Networks

	confirmed	confirmed_predicted
2020-07-12	12914656	1.286562e+07
2020-07-13	13107435	1.307987e+07
2020-07-14	13328887	1.329376e+07
2020-07-15	13560006	1.351194e+07
2020-07-16	13812547	1.373167e+07
2020-07-17	14054586	1.395792e+07
2020-07-18	14292221	1.418730e+07
2020-07-19	14506868	1.441855e+07
2020-07-20	14713646	1.464807e+07
2020-07-21	14947101	1.487898e+07
2020-07-22	NaN	1.511205e+07
2020-07-23	NaN	1.535376e+07
2020-07-24	NaN	1.559830e+07
2020-07-25	NaN	1.584609e+07
2020-07-26	NaN	1.609707e+07
2020-07-27	NaN	1.635157e+07
2020-07-28	NaN	1.660911e+07

MLP-World

	confirmed	confirmed_predicted
2020-07-12	12914656	1.274731e+07
2020-07-13	13107435	1.295299e+07
2020-07-14	13328887	1.316294e+07
2020-07-15	13560006	1.337692e+07
2020-07-16	13812547	1.359755e+07
2020-07-17	14054586	1.382632e+07
2020-07-18	14292221	1.405874e+07
2020-07-19	14506868	1.429583e+07
2020-07-20	14713646	1.453770e+07
2020-07-21	14947101	1.478465e+07
2020-07-22	NaN	1.503869e+07
2020-07-23	NaN	1.531663e+07
2020-07-24	NaN	1.560625e+07
2020-07-25	NaN	1.590865e+07
2020-07-26	NaN	1.622516e+07
2020-07-27	NaN	1.655706e+07
2020-07-28	NaN	1.690573e+07

LSTM(RNN)-World

Figure 6: World-Prediction of confirmed cases using Neural Networks

	Deaths	Deaths_predicted
2020-07-12	568994	569614.212665
2020-07-13	572809	574632.031882
2020-07-14	578469	579745.631318
2020-07-15	583962	584786.351573
2020-07-16	589761	589979.902569
2020-07-17	596504	595341.304694
2020-07-18	602131	600621.610206
2020-07-19	606160	605893.159439
2020-07-20	610208	611178.516651
2020-07-21	616432	616460.977555
2020-07-22	NaN	621769.370515
2020-07-23	NaN	627160.072495
2020-07-24	NaN	632593.074038
2020-07-25	NaN	638046.619157
2020-07-26	NaN	643544.417163
2020-07-27	NaN	649077.172927
2020-07-28	NaN	654629.259860

MLP-World

	Deaths	Deaths_predicted
2020-07-12	568994	566623.606691
2020-07-13	572809	571354.422033
2020-07-14	578469	576100.998677
2020-07-15	583962	580896.138990
2020-07-16	589761	585799.453024
2020-07-17	596504	590804.474604
2020-07-18	602131	595822.361177
2020-07-19	606160	600905.111570
2020-07-20	610208	606024.907757
2020-07-21	616432	611199.733788
2020-07-22	NaN	616459.158943
2020-07-23	NaN	621976.826965
2020-07-24	NaN	627610.751427
2020-07-25	NaN	633370.496879
2020-07-26	NaN	639264.146040
2020-07-27	NaN	645298.434509
2020-07-28	NaN	651481.377648

LSTM(RNN)-World

Figure 7: World-Prediction of deaths using Neural Networks

Also, we did some predictions of confirmed cases and deaths related to 5 most affected countries and Iran using ARIMA and Prophet. As examples, Figures 8 and 9 show some of such predictions.

Moreover, we did some predictions on multivariate time series using Linear regression, Support vector machine, Ensemble methods, etc., and at the end, by examining the correlations between the features, we studied the impact of adding some new parameters such as life expectancy, GDP per capita, social support, freedom to make life choices, generosity, and population in prediction of COVID-19 spread.

3 Conclusions and future works

As a conclusion based on the analysis of the observations, it seems that even though the total number of confirmed cases and deaths in the world are monotonically (almost exponentially) increasing, the recovery rate shows some increase whereas the mortality rate shows some decrease. On the other hand, by data modeling and prediction based on univariate time series, using Linear regression, Support vector machine, Random forests and XGBoost we concluded that Support vector machine and Random forests performed the best and the worst accuracy, respectively. Moreover, both of Multilayer perceptron and LSTM-RNN performed high accuracy, more than 99.98 in percent.

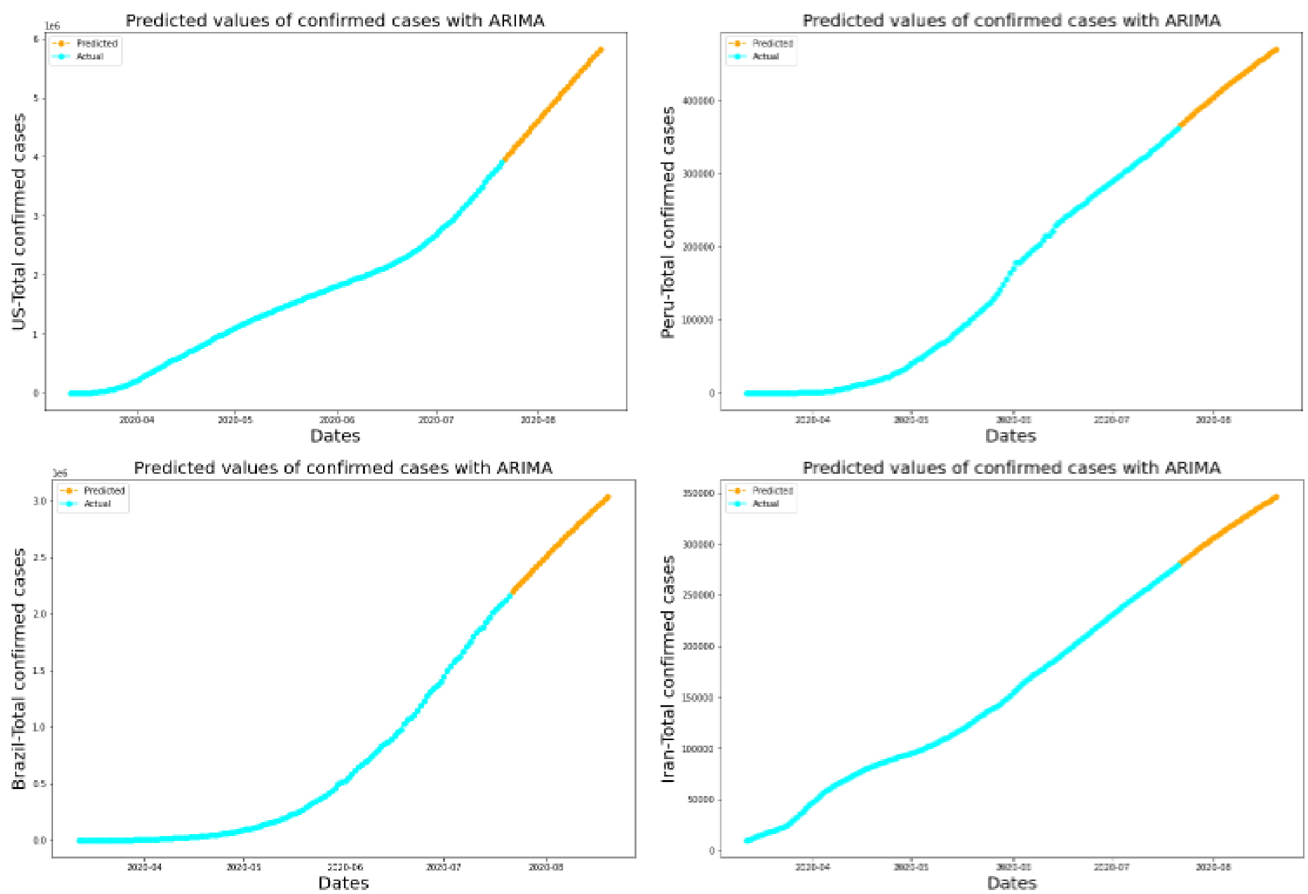

Figure 8: ARIMA-Confirmed

Furthermore, by examining the correlations between the features, it seems that there exist week correlations between the new parameters, life expectancy, GDP per capita, social support, freedom to make life choices, generosity, and the primary ones, confirmed, deaths, recovered, active cases, recovery rate and mortality rate. Also, it seems that the correlation between population and confirmed and, population and active cases is moderate (near 0.4).

As future works, by considering the population of each country, we may investigate the percentage of total populations that will be affected by COVID-19. Also, the impact of some other parameters in prediction of COVID-19 spread can be considered. Moreover, data modeling and prediction based on multivariate time series using Multilayer perceptron and LSTM-RNN can be considered.

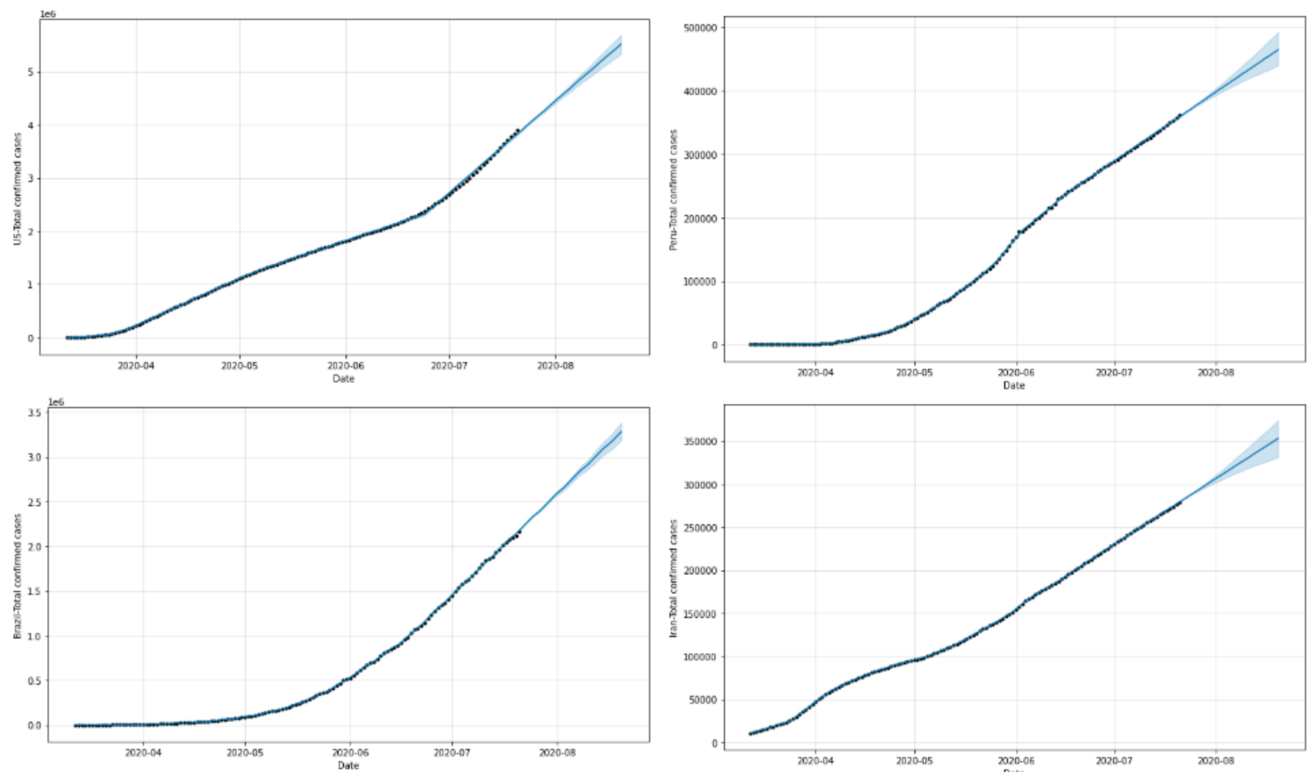

Figure 9: Prophet-Confirmed

References

[1] Alon, T. M., et al., The impact of COVID-19 on gender equality, National Bureau of Economic Research, (2020), no. w26947.

[2] Chen, B., et al., Roles of meteorological conditions in COVID-19 transmission on a worldwide scale, MedRxiv, (2020).

[3] Fernandes, N., Economic effects of coronavirus outbreak (COVID-19) on the world economy, Available at SSRN 3557504, (2020).

[4] Fong, S. J., Li, G., and Dey, N., Finding an Accurate Early Forecasting Model from Small Dataset: A Case of 2019-nCoV Novel Coronavirus Outbreak, Int. j. interact. multimed. artif. intell., 6(2020), no. 1, 132–140.

[5] Ho, C. S., Chee, C. Y., and Ho, R. C., Mental health strategies to combat the psychological impact of COVID-19 beyond paranoia and panic, Ann Acad Med Singapore, 49(2020), no. 1, 1–3.

[6] Huang, C., Wang, Y., Li, X., et al., Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China, The Lancet , 395(2020), no. 10223, 497–506.

[7] Jia, L., Li, K., Jiang, Y., and Guo, X., Prediction and analysis of coronavirus disease 2019, arXiv preprint, (2020).

[8] Kumar, J., and Hembram, K. P. S. S., Epidemiological study of novel coronavirus (COVID-19), arXiv preprint, (2020).

[9] Ma, Y., et al., Effects of temperature variation and humidity on the mortality of COVID-19 in Wuhan, MedRxiv, (2020).

[10] Shi, P., et al., The impact of temperature and absolute humidity on the coronavirus disease 2019 (COVID-19) outbreak-evidence from China, MedRxiv, (2020).

[11] Sujath, R., et al., A machine learning forecasting model for COVID-19 pandemic in India, Stochastic Environmental Research and Risk Assessment, (2020), no. 34, 959–972.

[12] Walker, P., et al., Report 12: The global impact of COVID-19 and strategies for mitigation and suppression, (2020).

[13] World Health Organization (WHO), Naming the coronavirus disease (COVID–19).

[14] World Health Organization (WHO), Novel Coronavirus–China, Retrieved 9 April 2020.

[15] Yang, Z., et al., Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions, J Thorac Dis, 12(2020), no. 3, 165.