



NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY

School of Electrical Engineering and Computer Sciences

Machine Learning

End-Semester Project Report

Submitted by:

Daim Abbas Kazmi 235149

Class: BESE – 8A

Date of Submission: 11th June 2020

Introduction:

We have been asked to predict the new cases of novel Covid-19 in Pakistan the month of June using the dataset present in Kaggle. The first task is to find five countries having similar rise of cases as in Pakistan and then selecting an appropriate model to train the data from Pakistan and the other selected countries. We were limited to the use of only scikit-learn and Pytorch for the implementation of our model.

Approach:

Finding similar Countries:

For the selection of top 5 similar countries, I used corrccoef method in numpy that uses Pearson correlation method to calculate the similarity between the 2 sets of data. I placed two limitations on finding the countries.

- First being that we correlation is calculated from 26/2/20 to 10/5/20. This is the data that we actually need to train.
- Secondly, I only considered those countries that have cumulative cases more than 20,000 so the data more represent that of Pakistan.

Using this, the five countries are following:

India 0.9976172372044213

Mexico 0.9974640778447007

Brazil 0.9973183087147551

Qatar 0.9965421674432555

Saudi Arabia 0.995875821968421

Processing Data:

The time-series data displayed the cumulative (total) cases on a certain date. The next step was to convert this data to show number of cases per specific day. This is easily done by subtracting the cases of one day from that of previous days. Once we have daily cases for each day, I processed this data such as 5 days cases predicted the number of cases for the next day. So, we have 5 features for each label. Training data included data from start of cases in the country to 10th may. Test cases include cases from 10th May to 27th May.

Models:

I used two models to predict this data. First one is MLPClassifier that stands for multi-layer perceptron is a neural network that is available in Sklearn. The idea to use this came from the research paper named 'COVID-19 Outbreak Prediction with Machine Learning' that shows that MLP works better than other models to predict the future cases. An MLP consists of at least three layers of nodes: an input layer, a hidden layer and an output layer. My MLP consists of 3 hidden layers each of 100 neurons. There are basically three steps in the training of the model i.e. Forward pass, Calculate error or loss and Backward pass. Its multiple layers and non-linear activation distinguish MLP from a simple neural network. It can distinguish data that is not linearly separable.

The second model is LSTM RNN. It is known for its ability to maintain knowledge for long time that can be useful in prediction such as the task that was given. LSTM is present in Pytorch library, so build using pytorch. Long Short-Term Memory networks – usually just called “LSTMs” are capable of learning long-term dependencies. The LSTM does have the ability to remove or add information to the cell state, carefully regulated by structures called gates. Gates are a way to optionally let information through. They are composed out of a sigmoid neural net layer and a pointwise multiplication operation. Making them useful in using its own prediction to forecast further.

Forecasting:

The predictions were made of two kind:

1. Between 10th May and 27th May (next 17 days to predict)
2. From 27th May until 27th June (next 31 days to predict)

The models are trained using the training data and then Pakistan's data from 10th May to 27th May is used to test the model and RMSE is calculated between the predicted and actual values. The smaller the error that means model is more accurate and for higher error vice versa.

RMSE for MLP:

```
Pakistan 1716.8655434010107
India 3422.2858202222264
Mexico 1925.3819178053388
Brazil 11381.11069220554
Qatar 1023.8409802997614
Saudi Arabia 1565.203989712448
```

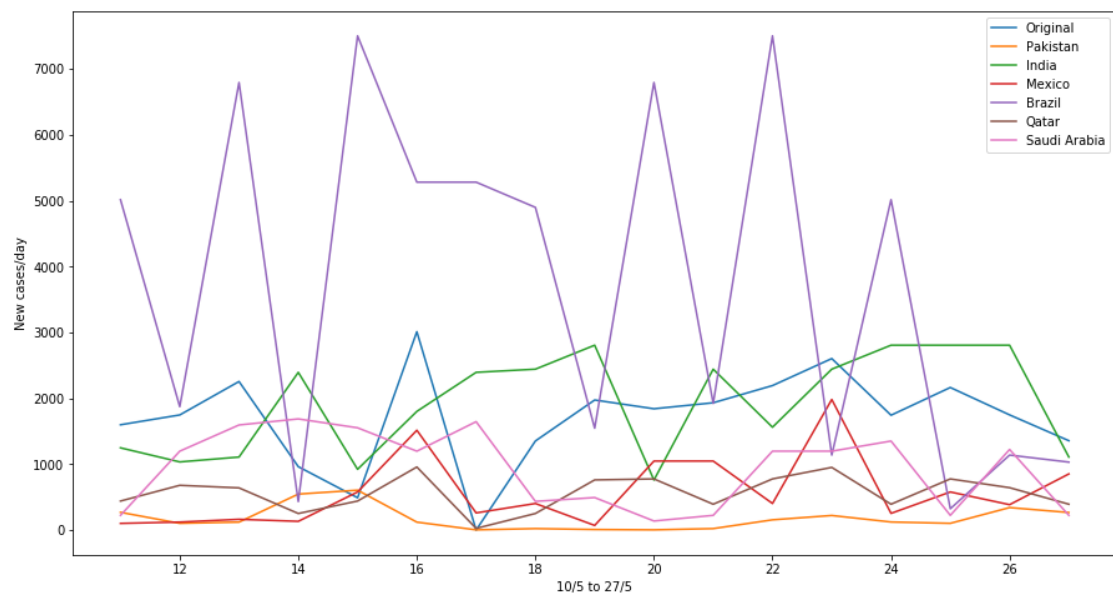
RMSE for LSTM:

```
Pakistan 1266.514136843142
India 1073.9366286485163
Mexico 1366.1819704821858
Brazil 2500.82367607625
Qatar 1829.588446168411
Saudi Arabia 2037.310712968562
Note: This is for 50 epochs
```

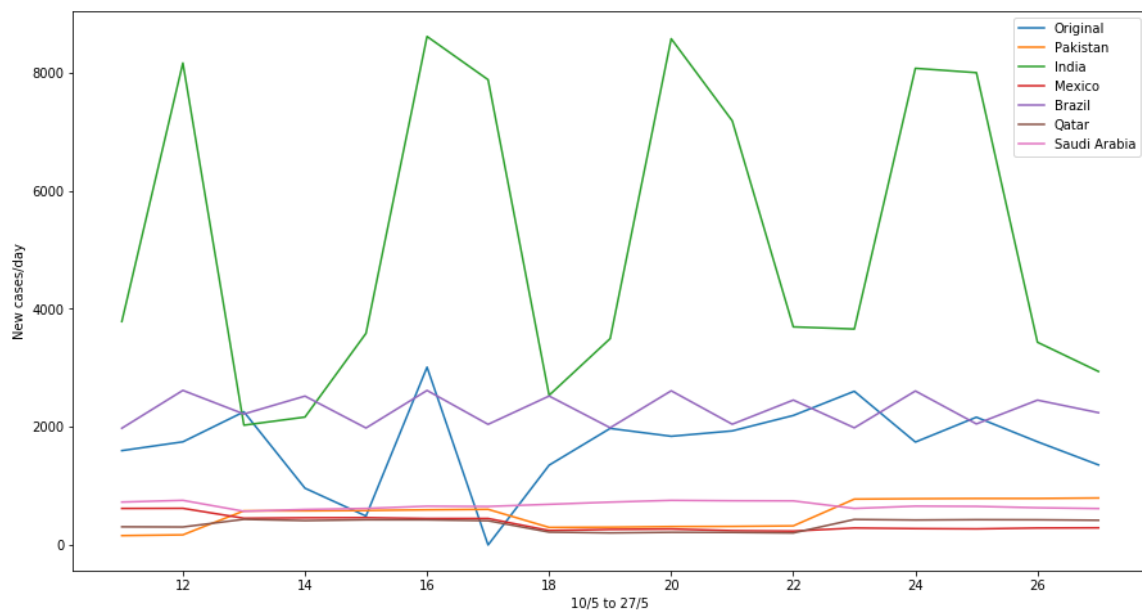
In the second prediction, I took the last sequence from existing data that I have and used to predict the data of the next day. Then I add this prediction into my initial sequence and then removing the first data value giving me a new sequence that is equal to the size of window that I have selected. After I gained results from both of the models, I have took average for each country for each day which gives me better confidence then if I had used a single model to do so.

Results:

MLP results for 10th-27th May:

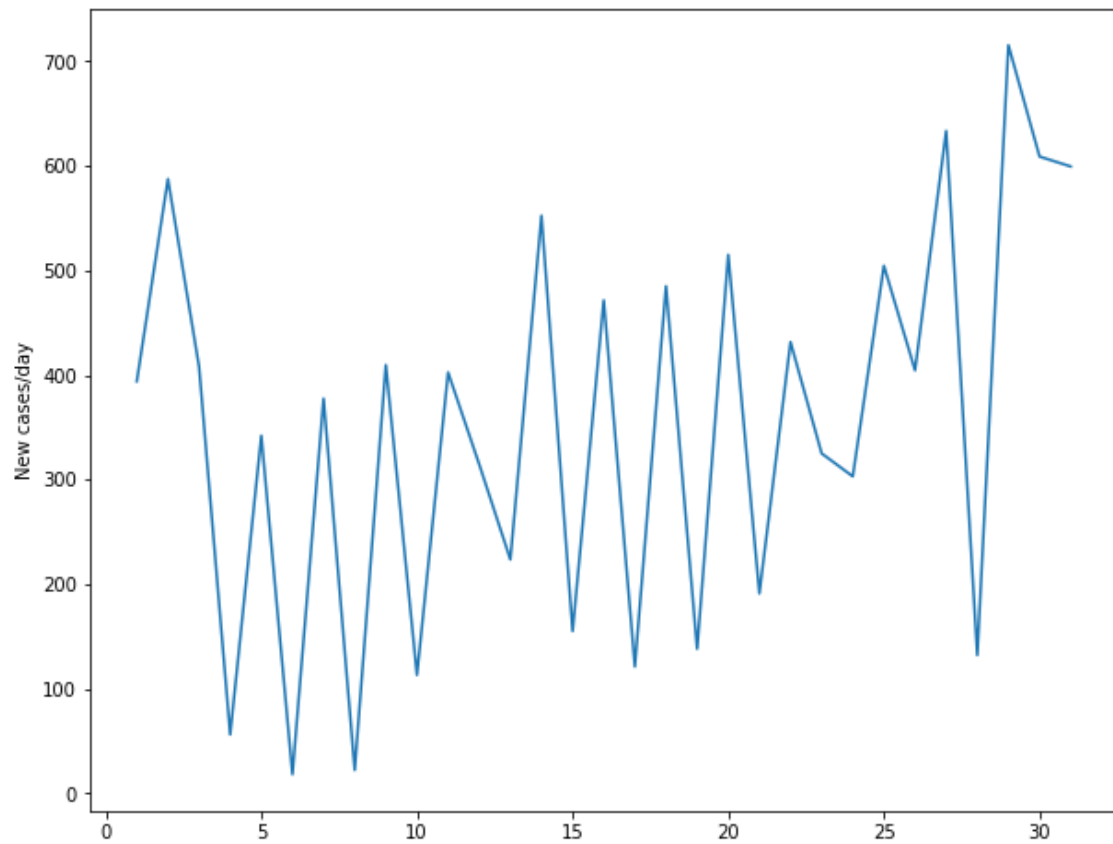


LSTM results for 10th-27th May:

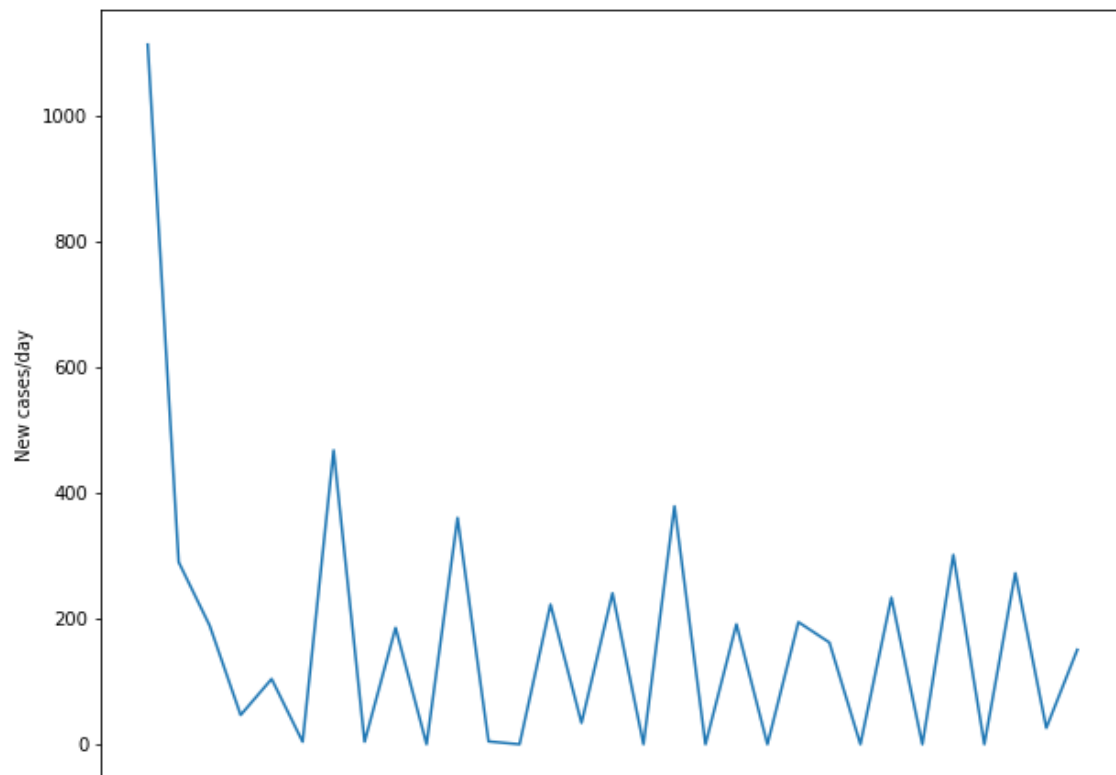


Prediction From 27th May-27th June:

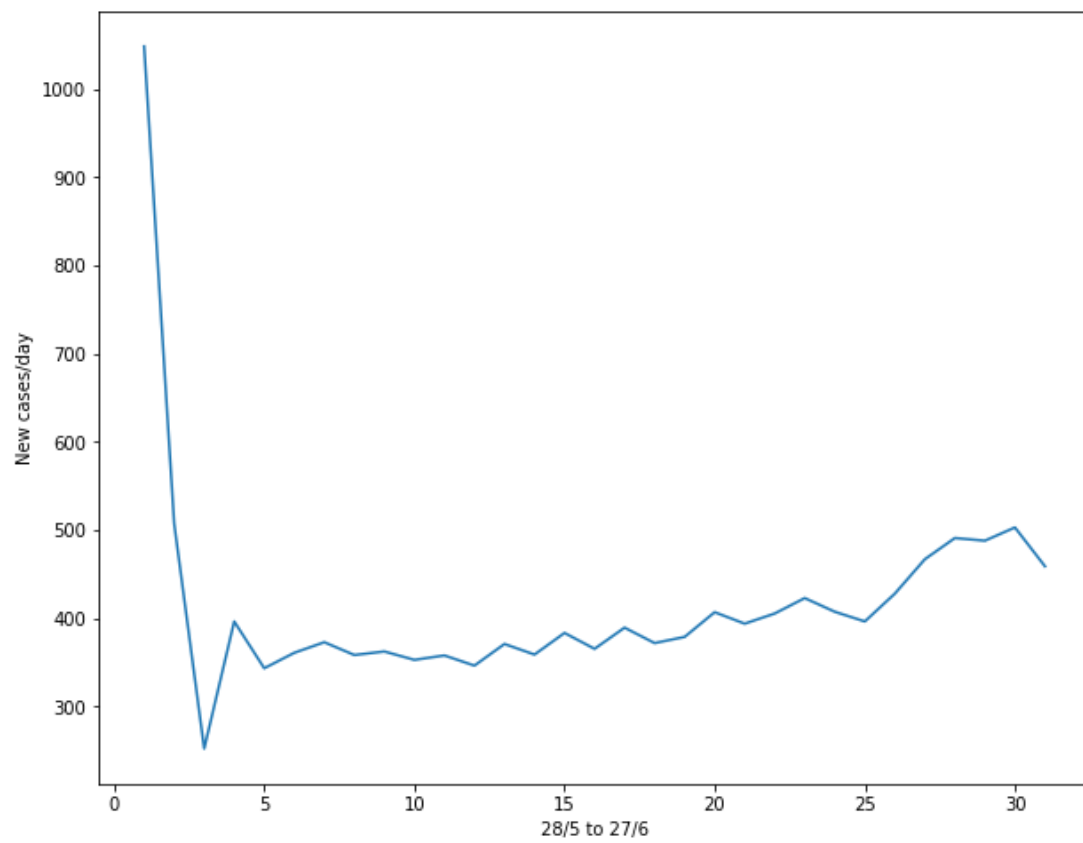
Pakistan



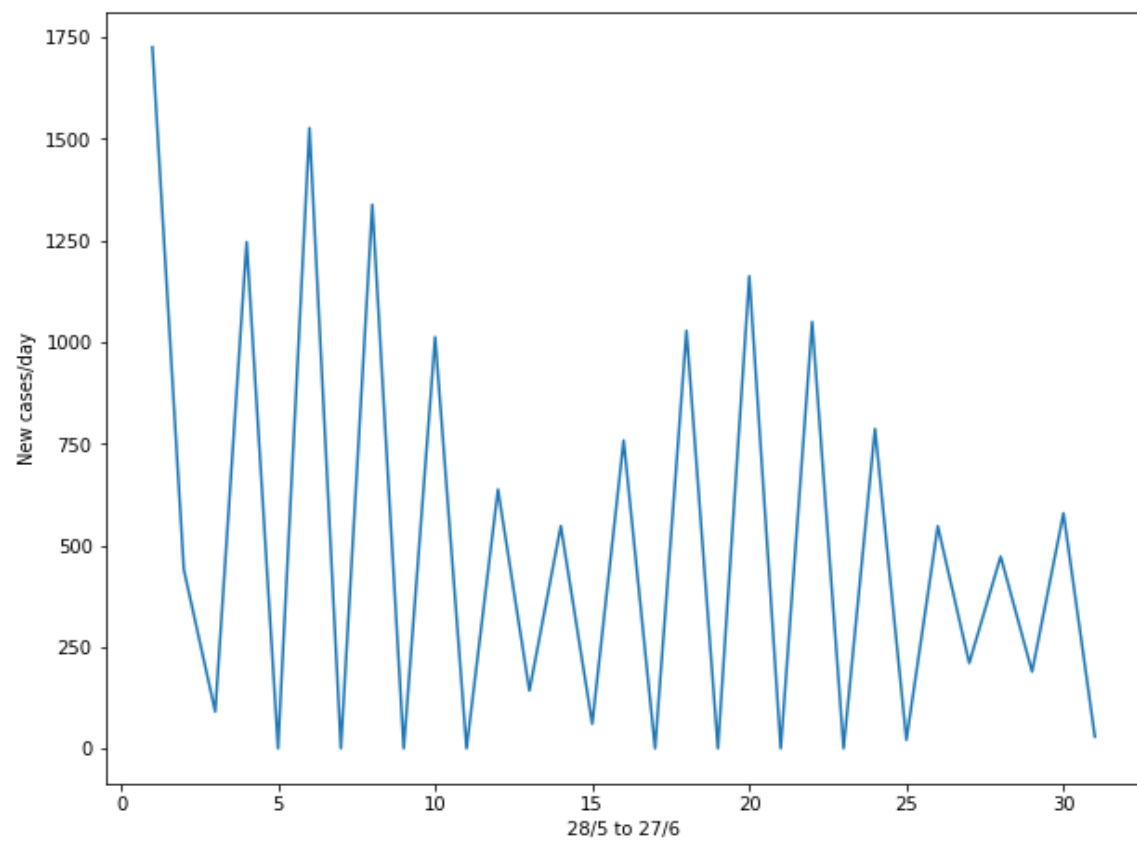
India



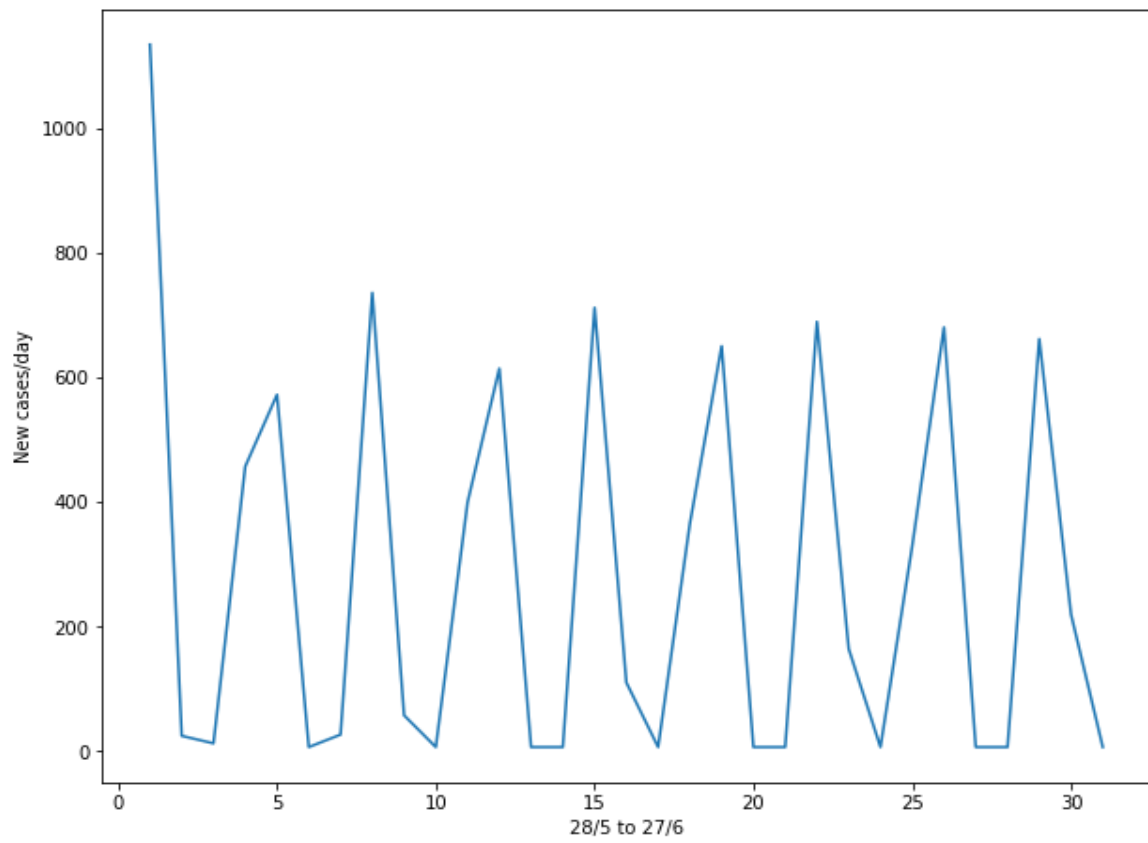
Mexico



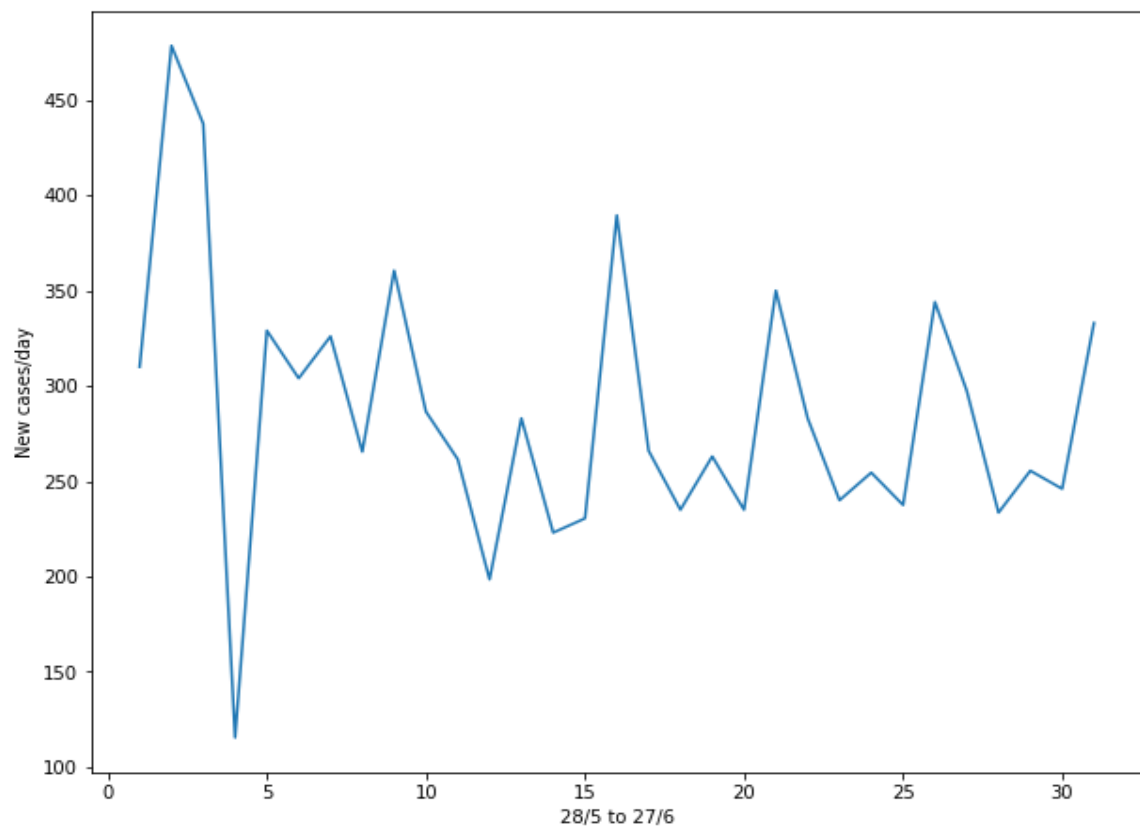
Brazil



Qatar



Saudi Arabia



Analysis:

Based on the data presented above, we can easily see that we get more complex outputs from MLP, which shows it was able to better learn the trend of the cases, however since the scale of cases of countries is different, this cases MLP to make some very high or low predictions for Pakistan, based on learnings from another country's trend. This can also be seen by the RMSE values for MLP.

LSTM on the other hand have a much better RMSE. This shows that they were able to make more smooth predictions, however on the same hand it was incapable of learning higher complexities. This is because LSTM require a lot of training, and we have trained for 50 epochs only. Training for a greater number of epochs will yield better results, however I wanted my LSTM outputs to compliment my MLP outputs for future prediction. Thus, I kept it at 50.




Future predictions clearly show a good trend, we have learned the complexity of data, as well as the general upward and downward trend as well. Which is a very good result for us.

Based on our analysis, we can say that Mexico and Pakistan itself were good predictor of future cases

Git-Hub:

<https://github.com/daim-cell/MLProject>

History:

Uncommitted changes		11 Jun 2020 20:54	*	*
	 master  origin/master Last Commit	11 Jun 2020 20:44	dkazmi <dkazmi.b	ba3ce73
	slight update	11 Jun 2020 16:24	dkazmi <dkazmi.b	c8a2ad4
	june onth prediction made	11 Jun 2020 4:29	dkazmi <dkazmi.b	67c1752
	june prediction remaining	11 Jun 2020 2:40	dkazmi <dkazmi.b	29719e0
	Prediction_done(RNN)	10 Jun 2020 18:23	dkazmi <dkazmi.b	42dd9d9
	prediction works(RNN)	10 Jun 2020 18:22	dkazmi <dkazmi.b	6408c79
	RNN working	9 Jun 2020 13:17	dkazmi <dkazmi.b	87b8fb8
	RunTime Error exists	9 Jun 2020 11:31	dkazmi <dkazmi.b	f9b3dc0
	AttributeError persists	8 Jun 2020 3:59	dkazmi <dkazmi.b	cf0c5f7
	RNN Class implementes	8 Jun 2020 3:20	dkazmi <dkazmi.b	d20be94
	Implementing RNN	6 Jun 2020 21:40	dkazmi <dkazmi.b	c297a66
	Graph comparing original vs predictions	4 Jun 2020 20:54	dkazmi <dkazmi.b	eb29568
	MLP implemented	4 Jun 2020 2:44	dkazmi <dkazmi.b	848dd0e
	Correlation Implemented	2 Jun 2020 1:35	dkazmi <dkazmi.b	22bbac9
	Start	1 Jun 2020 1:17	dkazmi <dkazmi.b	eba1a24