PROJECT REPORT

Machine Learning

Final Module of Covid-19 Project

Syed Jawad Bukhari
207414
Bese-8b

Syed Bukhari
Sbukhari.bese17seecs@seecs.edu.pk

Technique to find top5 same-trend countries with Pakistan:

- I applied some data cleaning techniques at first, such that get number of increasing cases for each country in specific column of days and indexing of each country

```
[ ] (188, 127)
```

Country/Region	1/22/20	1/23/20	1/24/20	1/25/20	1/26/20	1/27/20	1/28/20	1/29/20	1/30/20	1/31/20	2/1/20	2/2/20	2/3/20
Afghanistan	0	0	0	0	0	0	0	0	0	0	0	0	0
Albania	0	0	0	0	0	0	0	0	0	0	0	0	0
Algeria	0	0	0	0	0	0	0	0	0	0	0	0	0
Andorra	0	0	0	0	0	0	0	0	0	0	0	0	0
Angola	0	0	0	0	0	0	0	0	0	0	0	0	0

5 rows × 127 columns

- Then, I extract the Pakistan Row's data and remove trailing zeros from it and count the number of days till 27th May, which are 92.

```
column counts are:  92
```

Country/Region	2/26/20	2/27/20	2/28/20	2/29/20	3/1/20	3/2/20	3/3/20	3/4/20	3/5/20	3/6/20	3/7/20	3/8/20	3/9/20	3/10/20
Pakistan	2	2	2	4	4	4	5	5	5	6	6	6	6	16

1 rows × 92 columns

- Then, Using pandas technique, I shift all the trailing zeros for each country to it's right side and extract first 92 columns of all the countries whose first 92 columns do not have any zero. In this way I get a dataframe for only those countries which have more or equal number of days for coronavirus spread than Pakistan. Thus, I can compare day1 of every country with day 1 of Pakistan.


```
[ ]
```

Country/Region	1/22/20	1/23/20	1/24/20	1/25/20	1/26/20	1/27/20	1/28/20	1/29/20	1/30/20	1/31/20	2/1/20	2/2/20	2/3/20
Afghanistan	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Algeria	1.0	1.0	1.0	1.0	1.0	1.0	3.0	5.0	12.0	12.0	17.0	17.0	19.0
Australia	4.0	5.0	5.0	6.0	9.0	9.0	12.0	12.0	12.0	13.0	13.0	14.0	15.0
Austria	2.0	2.0	3.0	3.0	9.0	14.0	18.0	21.0	29.0	41.0	55.0	79.0	104.0
Bahrain	1.0	23.0	33.0	33.0	36.0	41.0	47.0	49.0	49.0	52.0	55.0	60.0	85.0
Belgium	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Brazil	1.0	1.0	1.0	2.0	2.0	2.0	2.0	4.0	4.0	13.0	13.0	20.0	25.0
Cambodia	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Canada	1.0	1.0	2.0	2.0	2.0	4.0	4.0	4.0	4.0	4.0	5.0	5.0	7.0
China	548.0	643.0	920.0	1406.0	2075.0	2877.0	5509.0	6087.0	8141.0	9802.0	11891.0	16630.0	19716.0
Croatia	1.0	3.0	3.0	5.0	6.0	7.0	7.0	9.0	10.0	10.0	11.0	12.0	12.0
Diamond Princess	61.0	61.0	64.0	135.0	135.0	175.0	175.0	218.0	285.0	355.0	454.0	542.0	621.0
Egypt	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Finland	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
France	2.0	3.0	3.0	3.0	4.0	5.0	5.0	5.0	6.0	6.0	6.0	6.0	6.0

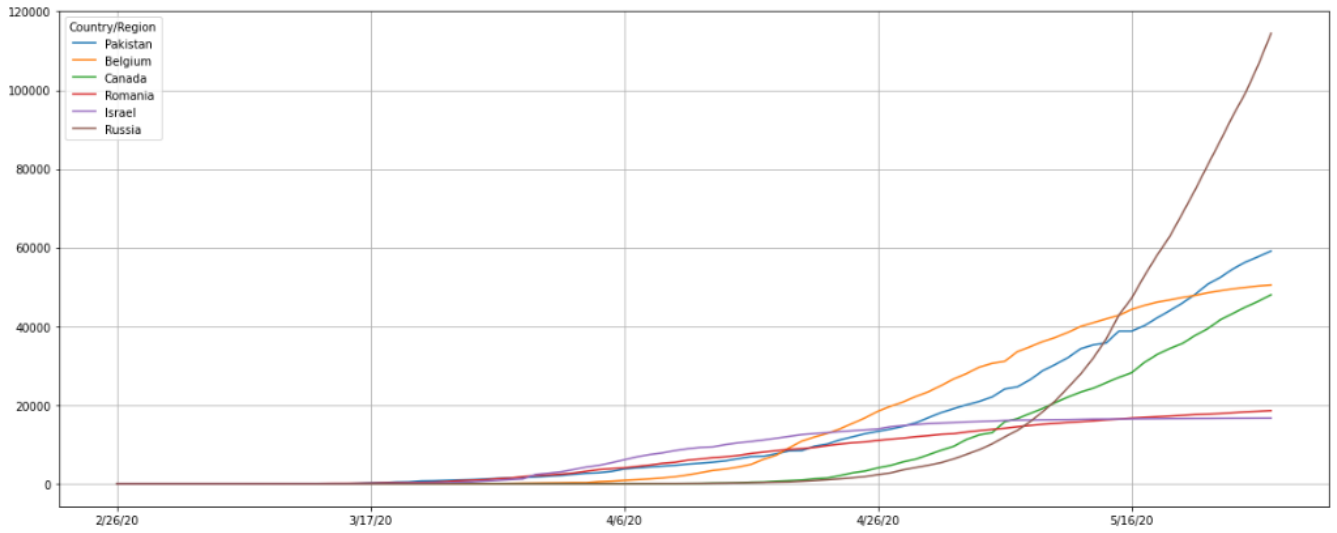
- Then, I subtract each country's corresponding cell with Pakistan and stored the result into a Score column. And select top 5 countries that have minimum absolute difference with Pakistan. And plot their result.

```
[ ] top 5 trend related countries with pakistan:
    ['Belgium' 'Canada' 'Romania' 'Israel' 'Russia']
```

Country/Region	2/26/20	2/27/20	2/28/20	2/29/20	3/1/20	3/2/20	3/3/20	3/4/20	3/5/20	3/6/20	3/7/20	3/8/20	3/9/20
Pakistan	2.0	2.0	2.0	4.0	4.0	4.0	5.0	5.0	5.0	6.0	6.0	6.0	6.0
Belgium	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Canada	1.0	1.0	2.0	2.0	2.0	4.0	4.0	4.0	4.0	4.0	5.0	5.0	7.0
Romania	1.0	1.0	3.0	3.0	3.0	3.0	3.0	4.0	6.0	9.0	9.0	15.0	15.0
Israel	1.0	1.0	1.0	1.0	1.0	2.0	3.0	4.0	7.0	10.0	10.0	12.0	15.0
Russia	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0

6 rows × 92 columns

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f123ff7fbe0>
```



Modeling and Prediction Techniques:

The easiest and better way I thought to perform modelling on the data was RNN model.

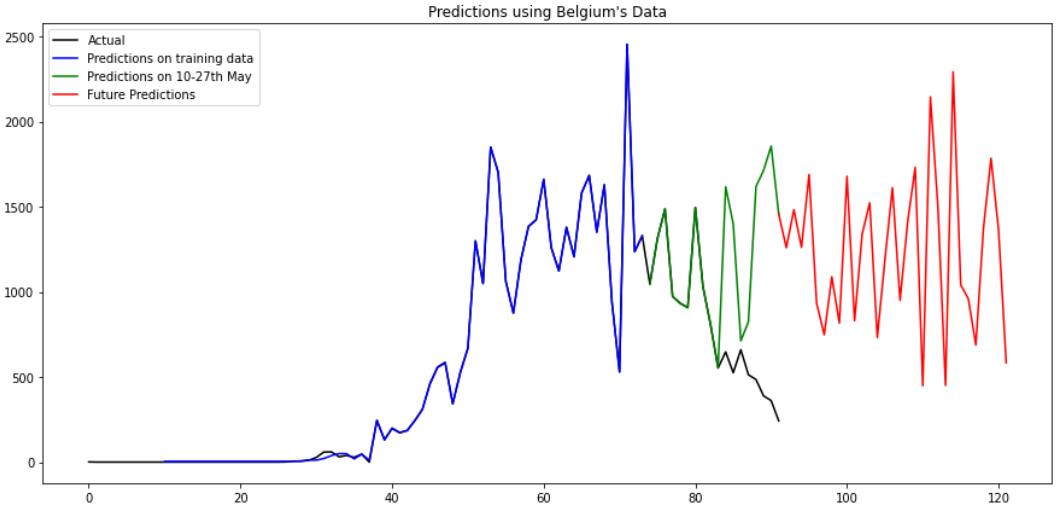
- I scale the final data frame of the top5 countries using minmax scalar. Since the cases come out to be more than three thousand for a single day.
- I convert the data frame in input and output in such a way that for particular number of days say 10 (i.e. stored in look_back variable=10), the 11th day record will be stored as output of those 10 days. In this way I loop across whole training data and stored an output data separately.
- The model used for forecasting is **Recurrent Neural Network**.
    - It consists of one input, one output and 2 hidden layers with 200 neurons each.
    - To introduce non-linearity in model, **ReLU** is used.
    - Model is trained with 0.006 learning rate for 1500 epochs.

- Mean squared error metric is used to compute training loss.
- *Look_back* used is 10.
- The hyper-parameters for this network are chosen using trial and error approach.

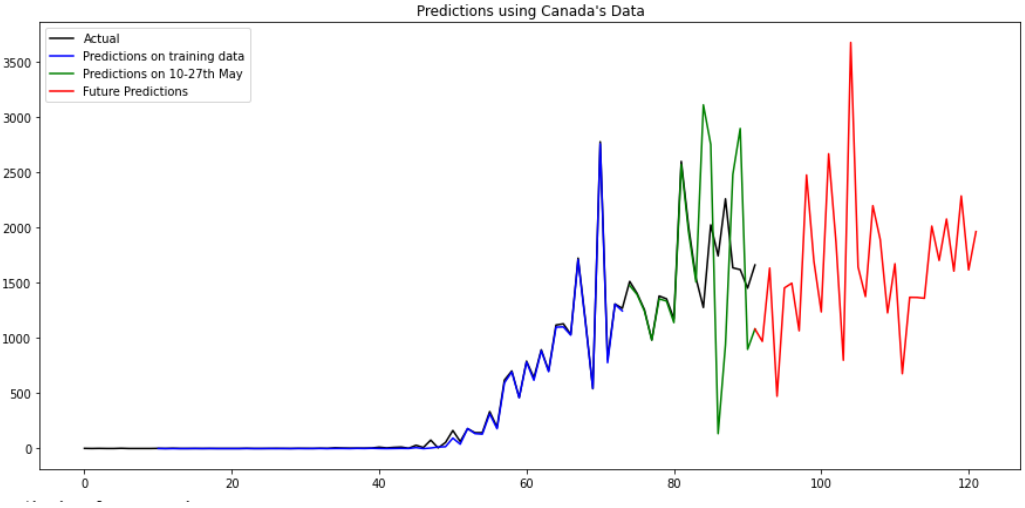
- My model predicts the training data value and calculates error. Then, predicts the validation data from 10th May to 27th May. And finally on the bases of that validation values as input, It predicts the future values from 27th May to 27th June.
- After that, the original data, prediction done on validation data i.e. 10th May to 27th May, and forecasted data are plotted using **Matplotlib**.
- At last, a model is trained on Pakistan's data and then evaluated on validation data and then used to forecast cases for the next month.

Results:

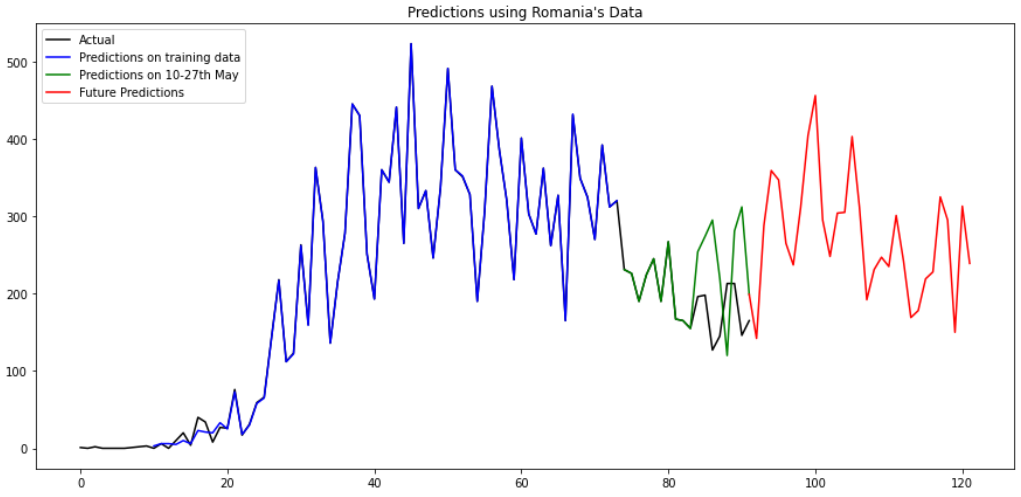
```
[ ] Prediction for  Belgium :
Epoch: 0 Loss: tensor(0.1146)
Epoch: 400 Loss: tensor(0.0001)
Epoch: 800 Loss: tensor(0.0001)
Epoch: 1200 Loss: tensor(2.0545e-05)
Epoch: 1500 Loss: tensor(7.4629e-06)
Each day prediction from 28May to 27June for Belgium: [1456 1259 1482 1261 1689  932  748 1088  816 1679  830 1339 1523  732
 1187 1612  950 1410 1731  448 2145 1495  450 2292 1041  960  688 1372
 1785 1360  583]
```



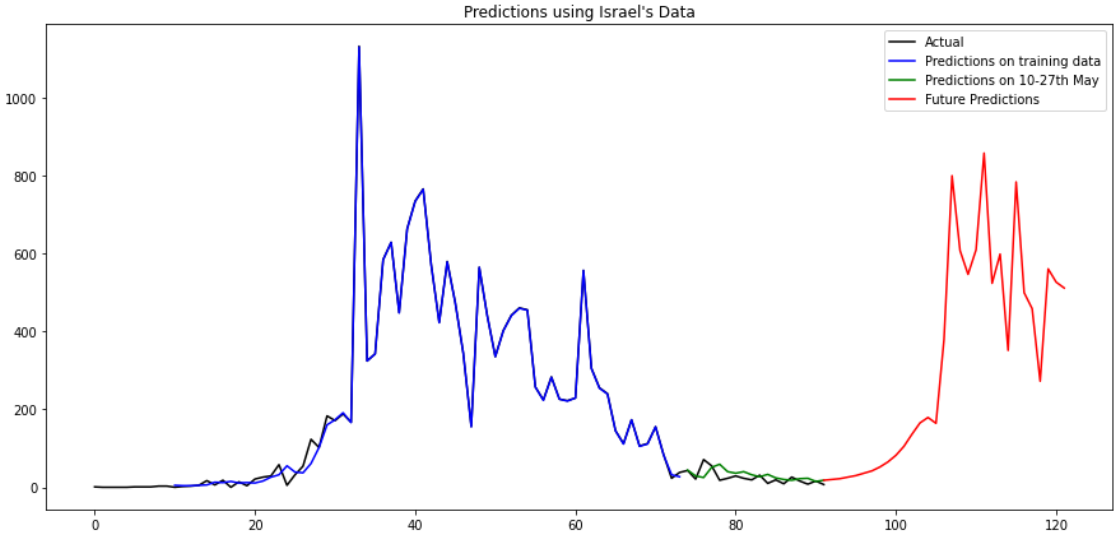

```
Prediction for  Canada :
Epoch: 0 Loss: tensor(0.0978)
Epoch: 400 Loss: tensor(4.2900e-05)
Epoch: 800 Loss: tensor(9.8174e-05)
Epoch: 1200 Loss: tensor(1.4306e-05)
Epoch: 1500 Loss: tensor(4.9451e-05)
Each day prediction from 28May to 27June for Canada: [1082  970 1634  473 1454 1497 1066 2477 1696 1237 2669 1869  799 3677
 1643 1376 2199 1893 1228 1673  677 1369 1368 1360 2014 1703 2078 1606
 2287 1617 1964]
```



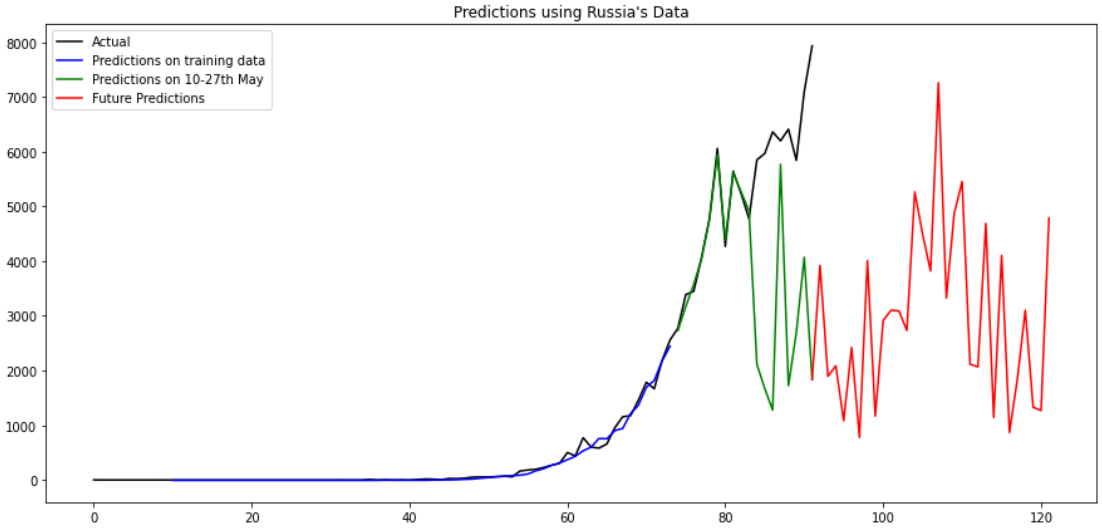
```
Prediction for  Romania :
Epoch: 0 Loss: tensor(0.2784)
Epoch: 400 Loss: tensor(0.0009)
Epoch: 800 Loss: tensor(0.0002)
Epoch: 1200 Loss: tensor(2.5684e-05)
Epoch: 1500 Loss: tensor(4.1018e-05)
Each day prediction from 28May to 27June for Romania: [199 142 288 359 347 265 237 311 404 456 295 248 304 305 403 312 192 231
 247 235 301 242 169 178 219 228 325 295 150 313 239]
```



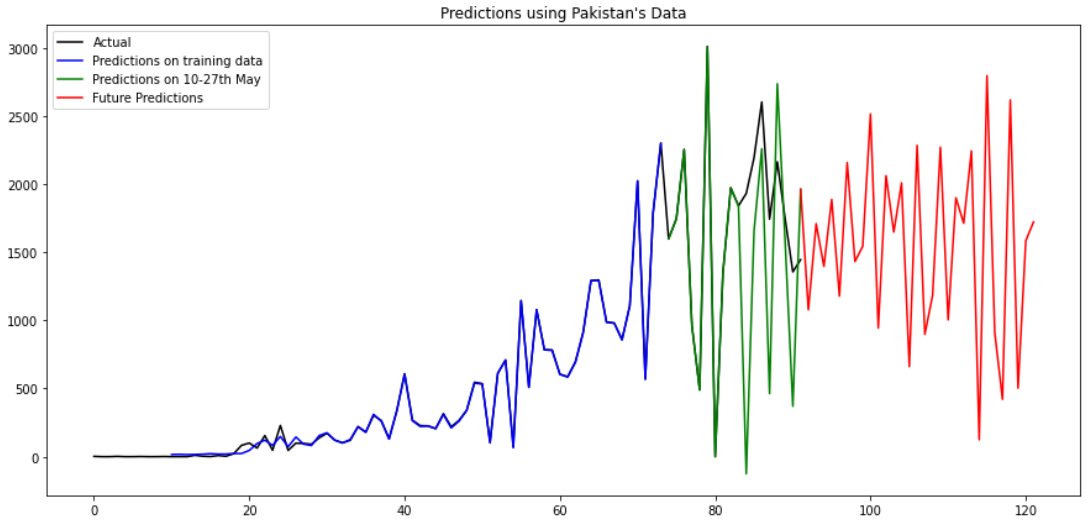

```
Prediction for  Israel :
Epoch: 0 Loss: tensor(0.0972)
Epoch: 400 Loss: tensor(3.4036e-05)
Epoch: 800 Loss: tensor(9.8112e-06)
Epoch: 1200 Loss: tensor(0.0002)
Epoch: 1500 Loss: tensor(0.0001)
Each day prediction from 28May to 27June for Israel: [ 18  20  22  26  30  36  42  52  65  82 105 136 165 179 164 379 799
 546 609 857 523 598 351 783 499 458 272 560 526 511]
```



```
[ ] Prediction for  Russia :
    Epoch: 0 Loss: tensor(0.0614)
    Epoch: 400 Loss: tensor(7.7610e-05)
    Epoch: 800 Loss: tensor(6.5713e-05)
    Epoch: 1200 Loss: tensor(4.7045e-05)
    Epoch: 1500 Loss: tensor(6.5565e-05)
    Each day prediction from 28May to 27June for Russia: [1837 3919 1893 2086 1082 2421  775 4010 1166 2913 3102 3090 2733 5265
     4489 3819 7261 3326 4870 5454 2115 2067 4689 1143 4103  869 1847 3102
     1329 1269 4788]
```




```
[ ] Prediction for  Pakistan :
    Epoch: 0 Loss: tensor(0.1013)
    Epoch: 400 Loss: tensor(0.0002)
    Epoch: 800 Loss: tensor(4.7683e-05)
    Epoch: 1200 Loss: tensor(4.7657e-05)
    Epoch: 1500 Loss: tensor(3.2315e-05)
    Each day prediction from 28May to 27June for Pakistan: [1962 1078 1710 1396 1887 1178 2158 1432 1544 2515  944 2062 1648 2009
      661 2284  898 1185 2271 1004 1900 1713 2244  124 2796  911  421 2618
      503 1585 1723]
```



Findings:

My model takes the past data of each country and predicts it future on that basis. Since in the past

data the number of cases in the initial days were much low. And the number of cases sometimes drops and sometimes get higher in day to day purpose. But when number of cases gets much higher (in validation set) the actual curve keeps on increasing (Russia) and the countries where they control , it keeps on decreasing (Israel and Belgium). But my prediction works on the fact that it sometimes decreases and sometime increases. So it's not that much better technique to predict but for the countries like (Romania and Canada), It fits better. And on the basis of that model we can predict the future values of Pakistan.