COVID19 PAKISTAN PROJECT

Group Members: Enayat Raza (DS-001), Sheraz Mohsin (DS-018), Shan Usmani (DS-005)

Instructor: Ma'am Murk Marvi | Course: DL - CT560 | NED UET - MS DS (E)

In the course project we are required to collect Pakistan's covid data and implement machine learning model on it to predict the cases and validate the model's performance. We have collected total 7 districts/provinces data which includes: **Sindh, Punjab, Balochistan, KPK, Islamabad, Gilgit-Baltistan and Azad-Jammu Kashmir**.

The first and important stage of project is to collect data. Since Pakistan's each district covid record/dataset is very scarce on open-web and very few sources has provided it available to fetch and download; yet after a lot of searching, we have found some authentic sources where our required data is available.

We used mainly three different sources to make up our final dataset csv files:

1. Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU).
2. Institute for Health Metrics and Evaluation (IHME).
3. Timeanddate.com (T&D).

While the features we got by using the above sources are: **Date, Province State, Confirmed Cases (n), Recovered Cases (n), Active Cases (n), Temp (DegC), Wind (Km/Hr), Hum (%), Mask Use (%), Change in Mobility / Soc Dis (%), ICU Beds Needed (n), Ventilators Needed (n), Test Conduct (n), Deaths (n).**

Since the data we collected are distinctively for individual districts/provinces that's why we have **total 7 CSVs (each csv for each state).** The data collected is for tenure **6 months (from start of July2020 till end of Dec2020).**

The features we got from these each of the three websites are:

1. **JHU: date, province, confirmedcases, deathcases, recoveredcase, activecases.**
2. **IHME: mask, mobility, beds, ventilators, testsconducted.**
3. **T&D: temp, wind, humidity.**

The way we extracted the data from these web sources are multiple i.e. like from JHU repository we got CSVs of whole world data in which our required Pakistan's 7 states records were present. We by using Python (using panda) read all original CSVs and filtered out only the Pakistan's 7 states rows from each of the original csv file, put it precisely date wise with the programming to make a data frame and finally exported/saved it to make a CSV file for each 7 states of Pakistan.

IHME has data in the form of fancy graphs in their website, we scrapped the graphical data using python (beautiful soup) and put it precisely date wise with the programming to make a data frame and finally exported/saved it to make a CSV file for each 7 states of Pakistan.

And lastly T&D has weather numeric data in their website, we extracted that data using scrapping and put it precisely date wise with the programming to make a data frame and finally exported/saved it to make a CSV file for each 7 states of Pakistan.

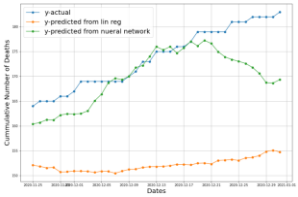
Now as our dataset (7 CSVs – each csv for each state) is finalized. The next stage now is to perform data preprocessing (EDA and feature engineering) before putting the data actually into the machine learning model. For it we initially check the null values and forward fill it, we then check heat map and with the help of correlation kept the necessary features and discarded the unnecessary ones. The target feature i.e. y-output we selected is number of deaths column, while the remaining columns we kept as x-input features.

After choosing the right features we now are ready to feed the processed data into our machine learning model. We treated the problem as continuous output problem. So we choose two machine learning models: **(1) linear regression model (2) deep neural network model.** After training the two models independently we then compare the accuracy output of the two.

We split the data into train and test samples. The last 20% data is kept for testing the model and starting 80% is for training the model. The complete solution code along with dataset has been provided with this report while the outputs obtained from both the models are pasted below:

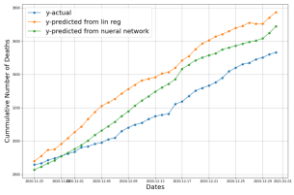
Results:

Balochistan



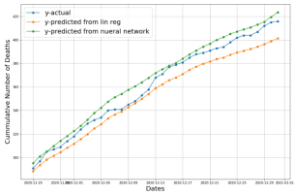
Accuracy lin reg: 88.0 %
Accuracy deep n/w: 98.0 %

Sindh



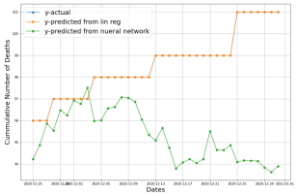
Accuracy lin reg: 94.0 %
Accuracy deep n/w: 97.0 %

Islamabad



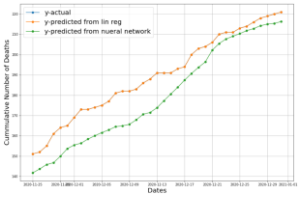
Accuracy lin reg: 98.0 %
Accuracy deep n/w: 98.0 %

Gilgit Baltistan



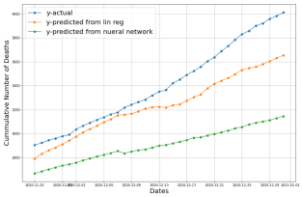
Accuracy lin reg: 100.0 %
Accuracy deep n/w: 97.0 %

Azad Jammu Kashmir



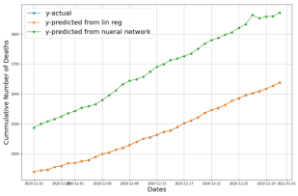
Accuracy lin reg: 100.0 %
Accuracy deep n/w: 94.0 %

Punjab



Accuracy lin reg: 95.0 %
Accuracy deep n/w: 86.0 %

KPK



Accuracy lin reg: 100.0 %
Accuracy deep n/w: 86.0 %

References:

Main features: date,pro,conf,death,rec,act

- https://github.com/CSSEGISandData/COVID-19

Weather features: temp.wind,hum

- https://www.timeanddate.com/weather/

Fancy features: mask,mobi,beds,vent,test

- https://covid19.healthdata.org/pakistan/islamabad-capital-territory?view=infections-testing&tab=trend&test=tests

Coding reference for implementing deepnetwork to predict continous traget varaibel at o/p:

- https://towardsdatascience.com/deep-neural-networks-for-regression-problems-81321897ca33
- https://machinelearningmastery.com/regression-tutorial-keras-deep-learning-library-python/ (Jason Brownlee)
- https://www.datacamp.com/community/tutorials/deep-learning-python

Coding reference for implementing lstm deepnetwork:

- https://machinelearningmastery.com/time-series-prediction-lstm-recurrent-neural-networks-python-keras/ (Jason Brownlee)
- https://machinelearningmastery.com/how-to-develop-lstm-models-for-time-series-forecasting/