Technical Project Proposal

Title: COVID-19 Time Series Analysis

Author: Priyanka Sahoo

Supervisor: Shagufta Henna, PhD

Degree: MSc in Big Data Analytics

Module: Big Data Analytics

Problem Description

The intended research would focus on data analytics and visualize predictions using various Machine Learning techniques against COVID-19 data sets. This 2019 corona virus pandemic globally affected every corner of the world. With the implementation of regression, clustering and classification models using python libraries greater percentage of accuracy could be achieved while forecasting the future effects of COVID-19 comprehensively [7]. This analysis may prove its potential to forecast the possibilities with unique and random occurrence of spread across countries [1]. Moreover, there will be scope for graphical representation of COVID-19 projections in sample countries based on available data set.

Data Description

The data set related to COVID-19 will be mainly extracted from government sites [5] and Kaggle [6]. Preferences will be given to .csv and json file as it has well defined structured format. There will be data related to countries, population, approximate test done on people, number of confirmed cases, death, recoveries with respect to dates, age, gender and country etc [2]. Various data acquisition process like cleansing, scrubbing, aggregation, and merging multiple raw data source will be the first step. Then the model will be trained with training data set and executed with test data to perform accurate predictions. The size of the data set is not completely agreed at this moment.

Proposed Methodology

Initially, raw data set from various source will be gathered and the prepossessed. The Source data could be .csv, json files or HTML. Using the various python libraries, the raw data will be staged in form of data frames in Pandas. Further, the exercise of data wrangling will be performed on the data set. Actual data analytics implementation on the test data will done in cloud environment e.g. AWS, Google cloud etc. using Hadoop cluster for storage medium and Apache Spark, well equipped with machine learning library. Finally, Plotting and visualization will be used on top of aggregated data to observe the predictions. Linear, polynomial, support vector regression, K-means clustering, decision tree classification will be used mostly to get the predictions with the least root mean squared error. The more we dive into the data there may be possibilities of enhancement with algorithms used [3] [4].

There will be application of Jupyter notebook for data analytic, as its an open-source web application, where we could build and execute the step-by-step prediction model using python. Basically, datetime library will be applied for time-series analysis and the visualizations will be developed using the Matplotlib, Seaborn, Plotly libraries in Python.

Goals

The primary objective of this research is to analyze the COVID-19 related data with respect to time and then predicting the number of confirmed cases (Daily and cumulative), total number of deaths and recovered using Machine Learning Regression Methods (Linear, Polynomial, Support Vector Regression etc) [3].

As part of this venture, we would be able to appease further down queries (subject to change as data exploration proceeds):

1. Percentage of total deaths and recovered daily/weekly/monthly across all over the world.
2. Effect of Corona virus on different age and gender.
3. COVID-19 variations and trends over time.
4. COVID-19 cases spread Visualization across countries in 2019-2020.
5. Predictions for the new confirmed cases, death and recovered numbers.

References

[1] "Time Series Analysis of the Covid-19 Datasets - IEEE Conference Publication." https://ieeexplore.ieee.org/document/9298390 (accessed Jan. 18, 2021).

[2] A. Bansal, A. Bhardwaj, and A. Sharma, "Forecasting the Trend of Covid-19 Epidemic," in 2020 Sixth International Conference on Parallel,

Distributed and Grid Computing (PDGC), Nov. 2020, pp. 406–409, doi: 10.1109/PDGC50313.2020.9315795,2020.

[3] S. Roy, M. N. Pal, S. Bhattacharyya, and S. Lahiri, "Implementation of an Informative Website – 'Covid19 Predictor', Highlighting COVID-19 Pandemic Situation in India," in 2020 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS), Sep. 2020, pp. 1–6, doi: 10.1109/IEMTRONICS51293.2020.9216352, 2020.

[4] S.Bodapati, H. Bandarupally, and M. Trupthi, "COVID-19 Time Series Forecasting of Daily Cases, Deaths Caused and Recovered Cases using Long Short Term Memory Networks," in 2020 IEEE 5th International Conference on Computing Communication and Automation (ICCCA), Oct. 2020, pp. 525–530, doi: 10.1109/ICCCA49541.2020.9250863, 2020

[5] https://data.gov.ie/dataset/covidstatisticsprofilehpscirelandopendata1/resource/388b6f93-8514-4426-bfd9-b41fb46efe33, (2019-2020)

[6] https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset, (2019-2020)

[7] https://www.worldometers.info/coronavirus/, (2019-2020)