Improve COVID-19 Outbreak Forecasting using Machine Learning Algorithms

Meaad Alrehaili

Prediction is crucial for clinical purposes, particularly for health-based systems. Computer-aided clinical predictive models have been used in various areas, including predicting the progression of different diseases. In this project, we will apply prediction models to build a predictive model for COVID-19.

Proposed Work:

We proposed ensemble models for improving COVID-19 forecasting which is the gradient boosting and random forest ensemble model. Gradient boosting is a machine learning model that generates a forecasting model in the form of an ensemble of weak prediction models to increase the prediction performance. Random forest is a common machine learning method for developing prediction models in many research settings. To minimize the burden of data collection and to improve its efficiency; the random forest can be used as a prediction model to decrease the number of variables required to achieve a prediction. We will predict a new confirmed cases, recovered cases, and deaths cases of the upcoming 10 days. We will use the evaluation methods R^2-score, adjusted R^2, MAE, MSE, and RMSE to show the model performance in COVID-19 forecasting.

Question/need:

- How will the prediction of COVID-19 help in controlling the situation by assisting the authorities in taking the right actions at the right time?
- What are the results of performance of the ensemble models?

Data Description:

The COVID-19 dataset that will be used in this project and it obtinaed from Johns Hopkins University (JHU). It contains data from January 22, 2020, and is updated daily. The data sourced from governments, national agencies across the world, and the WHO. The number of global confirmed cases of the time that this project will conducted is 108,535,89, and the number of global deaths is 2,393,707. The aforementioned dataset was accessed from the COVID-19 Data Repository of the Center for Systems Science and Engineering at JHU. The data features include the state, region, date for , number of confirmed, deaths, and number of recovered cases. Total features are 49 and the total samples around 7000+.

Tools:

- Jupyter notebook, Python 3 with some libraries (math, NumPy, pandas, sklearn, matplotlib).
- Ensemble technique to combine Gradient Boosting Regressor with random Forest Regressor.

MVP Goal:

Three numbers predicted (the numbers of confirmed cases, recovered cases, and deaths) for a period of 10 days.