Improve COVID-19 Outbreak Forecasting using Machine Learning Algorithms

Meaad Muhammad Alrehaili

1. Abstract:

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 applied ensemble models which are Gradient boosting and random forest to build a predictive model for COVID-19. The COVID-19 dataset is used from Johns Hopkins University (JHU). The ensemble models predicted a new confirmed case, recovered cases, and deaths cases of the upcoming 10 days. We used different evaluation to show the model performance in COVID-19 forecasting. The results of the ensemble methods show the effective of the models.

2. Design:

Forecasting the progression of such pandemic is indeed significant for planning the next step to be taken by organizations. We applied ensemble models for improving COVID-19 forecasting. We predicted a new confirmed cases, recovered cases, and deaths cases of the upcoming 10 days using the Covid-19 dataset.

3. Data:

The COVID-19 dataset is obtinaed from dataset was accessed from the COVID-19 Data Repository of the Center for Systems Science and Engineering at Johns Hopkins University (JHU). The data sourced from governments, national agencies across the world, and the WHO. 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+. The type of features includes object, date, floats, and integer.

4. Algorithms:

- *Feature Engineering*
  - Checking Null Values and Province/State has Nan values, we will fix it.
  - Deleitng the Province/State Country/Region Lat Long columns.
  - Date conversion and data preparation- splitting date into month, year and day of month.
  - Splitting data for training input and prediction output.

- *Models*

Ensemble of gradient boosting and random forest were used in this project to minimize the burden of data collection and to improve its efficiency.

- *Model Evaluation and Selection*

R^2-score, adjusted R^2, MAE, MSE, and RMSE are used to evaluate the model performance in COVID-19 forecasting.

- *Holdout*

GB-RF	***R^2-score***	***Adjusted R^2***	***MAE,***	***MSE***	***RMSE***
Confirmed cases	0.99	0.99	69138158.83	4092.88	8314.94
Recovered cases	0.99	0.99	1682739.6	711.0	1297.2
Death cases	0.99	0.99	200297.36	215.2	447.55

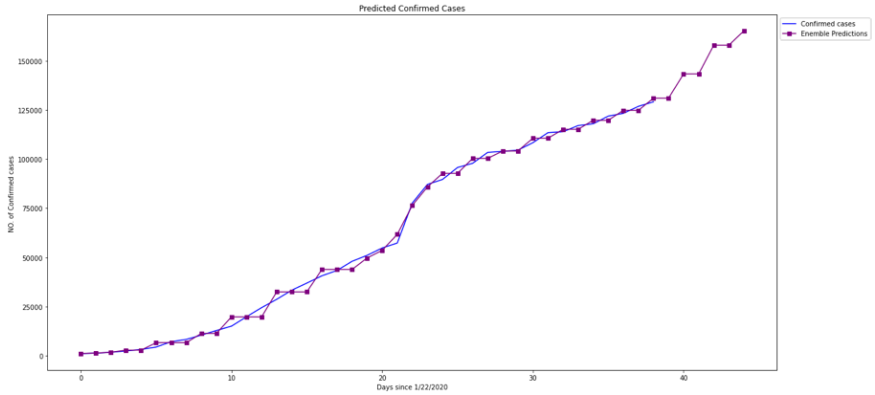
5. Tools:

- Jupyter notebook with Python 3
- Numpy and Pandas for data manipulation
- Scikit-learn for modeling
- Matplotlib and for plotting

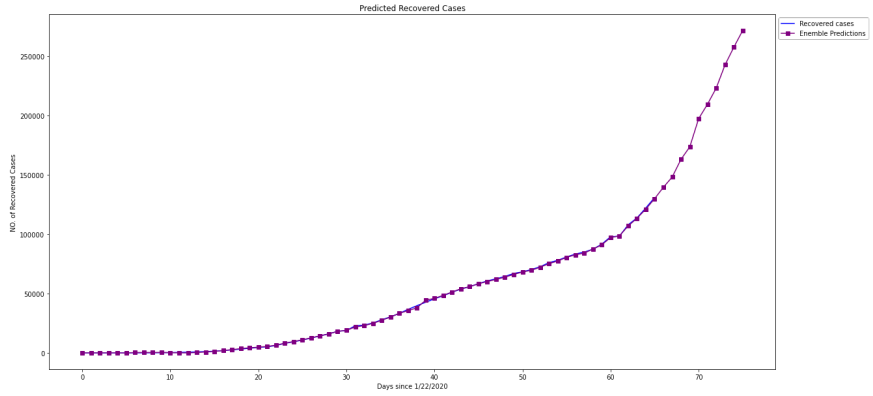
6. Communication:

Three numbers predicted numbers of confirmed, recovered, and deaths for a period of 10 days. See the code, slides, and writeup in repo: https://github.com/meaadalrehaili/Tuwaiq_Data_Science_Bootcamp

- *Confirmed cases*



- *Recovery cases*



- *Death cases*