SEP 24, 2022

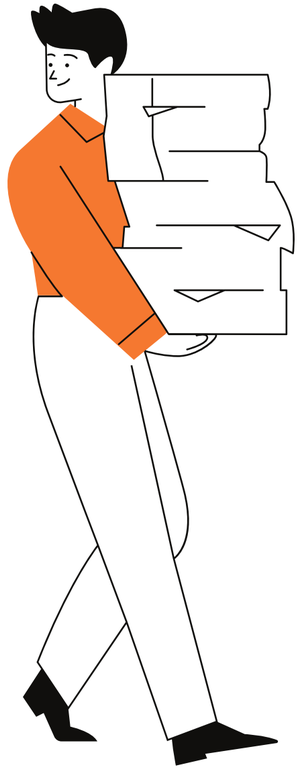
Research Methodology

Fish Out the Root of the Problem

Back to Agenda Page

Data Ingestion

- Download 3 datasets from its source and directly load (no conversion) into the development environment.



Back to Agenda Page

ETL (Extract, Transform, Load)

- Data Imputation => Clean each file separately
  - **Clean** :
    - **Missing Analysis** => Using dtale lib to analyze the whole dataset. Resulted in a form of graph showing how much are the missing and valid (%)
    - **Columns Dropped =>** Drop columns that has 80% of its values are NA
  - **Join :**
    - **NA Filling for another columns =>**
      - Joining owid.csv with vaccination.csv by using date as key for columns named "people_vaccinated", "total_vaccinated", "total_boosters", and "new_vaccinated".
      - Extract "new_cases", "total_vaccinated", "total_boosters", "new_vaccinated", and 'people_vaccinated' from images from ddc.moph.go.th

Back to Agenda Page

Modeling

- Visualize most of the features in a form of scatter plot which **y = new_deaths** to see how distribute is the features to the target.
  - Most of the features aren't in a linear form, but some
- Decided to use **Ensemble** and **Linear** models.
- Split data by **datetime split technique.**
  - **25%** for test size.
- Data Transformation:
  - Base Line - No Scaled data; R2 Score of Model = 0.92
  - **Improved-v1 - Scaled Both X and Y; R2 Score of Model improved to = 0.99**
    - ;because range of value of features aren't the same. Scaling by using **sklearn.StandardScaler** (Standardized scaling).
  - **Improved-v2 - Scaled X but not Y; R2 Score of Model = 0.99**
- Train every models which are in the mentioned fields.
- Model assumption: due to Spread of data (From scatter plot), Most of the features are non-linear but some of numerical data such as people_vaccinated is linear, so I decided to use both **Ensemble** and **Linear** models to see how it's gonna be and tracking their metrics later.

Back to Agenda Page

Models Evalutation

- Benchmark models =>
  - Track models' result by using MLFlow in metrics; R2, MAE, MSE, and RMSE.
- Model selection => Each metric presents different result; R2 presents accuracy, MAE, MSE and RMSE presents errors. Because this dataset is about medical information, error must be an important factor to concern, by this reason, metrics that should be used for this problem is **MSE.** for error, MSE must be low due to concerned error.

Back to Agenda Page

Deployment

- Deployed to production by using FastAPI and containerize it as a docker image.
- Testing it with 2 test cases using Postman to send POST request to the API.