Data Science Project Proposal

Project Group:
NO Work In Main Branch

Team members:
Danling Zhou, Mingyi Gong, Yiqi Zhu

Project Title:
Daily Covid Cases Prediction

Project Overview:
Our target variable is daily confirmed new cases. We will use variables including but not limited to demographics, economic info, policy, and vaccination data to predict our target variable.

As now we've entered the post-covid era, we would like to take a retrospective perspective to reflect on the pandemic and provide insights on future infectious diseases.

Objectives:
Prediction: To make a machine learning model (or an ensemble model) to predict how the pandemic propagated.

Inference: To interpret how economic conditions, government policies, and access to COVID vaccine impacted the spread of COVID.

Stakeholders:
Epidemiologists, public policy researchers, the general public who wanted to have a sense of how the pandemic went.

Methodology:
Data Visualization: We will use matplotlib and/or seaborn to visualize the distribution of the data, especially with respect to time.

Missing data: we will select columns with <20% of missing values. For the missing values, we will decide whether to use KNN to impute or drop the rows with too many missing values. We recognize that imputing may lead to extra bias and dropping rows means letting go (possibly important) information.

Feature selection: we expect to have a final table with 40-50 columns, and since we will be likely to use tree-based models, feature selection is not a priority. However, if needed, we will use stepwise selection methods and information gain.

Building models: We plan to use the algorithms we learned before, like linear and logistic regression, decision trees and random forests, boosting. We will also try to explore new algorithms like neural networks. To improve the accuracy of our models, we will use cross validation and ensemble models as well.

Data Sources:

The Google Health COVID-19 Open Data Repository is one of the most comprehensive collections of up-to-date COVID-19-related information. Comprising data from more than 20,000 locations worldwide, it contains a rich variety of data types to help public health professionals, researchers, policymakers and others in understanding and managing the virus.

https://health.google.com/covid-19/open-data/raw-data

Dataset:

We plan to use the following tables in the datasets mentioned above.

After preliminarily assessing the missing values in each table, we expect that the final aggregated dataset will contain at least 100000 rows and 50 columns by joining these tables together using common location and/or date.

Name	Rows	Columns
epidemiology	12525825	10
demographics	21689	19
economy	404	4
emergency_declaration	8364	104
geography	22130	8
health	3504	14
hospitalization	1768485	11
mobility	6321226	8

search_trends	2713929	424
vaccination	2545118	32
government_response	303969	22
world_bank	215	1405
epi_by_age	3822577	152
epi_by_sex	3848340	30
key	22963	15

Conclusion:

By analyzing the dataset derived from the Covid 19 Open Data, the project will provide us with a model (or an ensemble of models) that can predict the daily confirmed cases and insights on what impacted the transmission of COVID-19.