CMPS 396X- Fall 20-21

Term project pre-stages

1. What is the problem you are trying to tackle?

Researchers are combining efforts in studying the COVID-19 virus and finding effective solutions in containing the virus. In our study, we aim to study the worldwide spread of COVID-19 and explore the different factors that might cause individuals to be more susceptible to contracting the virus. Some of the factors we are considering are:

- Age
- Health indicators
- Economic status
- Weather
- Population density
- Mobility/human interactions
- Government measures and policies

These factors will help us in predicting and forecasting the spread of COVID-19.

2. What would change five years down the road, if your solution were to scale?

COVID-19 is a new virus still in its early stages. Scientists and researchers are still learning the virus where a lot of information is still unknown and its future is still vague. With COVID-19 confirmed cases exceeding 9 million and continuing to increase, scientists are racing to develop vaccines and cures to control the virus. It is still uncertain whether this virus will control humanity or human beings will succeed to overpass it. Although many factors such as vaccine development, virus mutation, and effective treatments can change the scale of our study, it will help in giving insights in controlling COVID-19 and in preparing for future outbreaks and pandemics by applying accurate safety and governmental measures.

3. Who are your customers?

Several parties would be interested in knowing how the COVID-19 virus will spread and how that spread is affected by different factors. Ministries of Health for global governments would benefit from knowing how different preventative measures adopted in other countries can help curb the numbers of infected cases. Additionally, having an estimate of how the numbers will

grow will help governments prepare to accommodate these cases, in addition to hospitals and other health-care professionals. Moreover, knowing how individual factors, such as age and health, play a role in the spread of COVID-19 can help increase the understanding reasons making certain people at a higher risk of infection and can benefit these individuals, their caretakers, and their doctors curate preventative measures specific to their needs.

4. Who are your competitors? Explain whether this is applicable or not, and justify why in each case.

COVID-19 is a novel virus and has become a global issue. People worldwide are being proactive in their study and are competing to be the pioneers in this new field. We think our competitors include researchers who are trying to obtain the most precise norms and standards to contain the spread of the coronavirus pandemic and studying the factors of COVID-19 growth with more accurate results.

5. What are the data sources you are going to rely on? Describe the location and ownership of the data, whether scraping will be needed, its volume, the speed by which it is readily made available (is it daily data? Hourly data?), how structured (or not), it is. Whether it is textual data, numeric data, images, audios, …etc. Be explicit whether some of the data revolves around human subjects, in which case, explain issues related to the privacy and confidentiality imparted to the data itself.

We have collected different datasets that could be linked on date and the country name features. Most of the datasets we found are updated daily. The data sources are as follows:

- **Epidemiological data from the COVID-19 outbreak, real-time case information [2]**

  A real-time database of individual-level epidemiological data. It covers **daily** COVID-19 worldwide patient records anonymously having a unique ID that doesn't reflect case or patient details **[1]**. The dataset includes country, geolocation, gender, symptoms, dates (date of onset, admission, and confirmation), chronic disease, and travel history. This dataset contains date, textual, binary, and numeric data.

- **COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University [3]**

  The dataset covers daily COVID-19 worldwide reports collected by Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE) and used for Novel

Coronavirus Visual Dashboard. It includes **daily** confirmed cases, death records, and recovered patients. The dataset is separated into three CSV files containing numeric, dates, and textual data. Raw data is needed to be extracted from the files as the aforementioned sources are developed as reports having dates as columns and countries as rows. We aim to restructure the data into raw data showing the numbers per country and date.

- **Coronavirus (COVID-19) Testing [4]**

  The data is collected by the *Our World in Data* team from official reports. It includes **daily** COVID-19 tests per country, the cumulative tests number, and country's information (population, median age, 65 years and older,70 years and older, extreme poverty, female and male smokers, GDP per capita, Cardiovasc death rate, human development index, hospitals beds, and life expectancy). This dataset includes textual, date, and numeric data.

- **OXFORD COVID-19 Government Response Tracker(OxCGRT) [5]**

  The Oxford COVID-19 Government Response Tracker (OxCGRT) collected information on government responses to COVID-19 policies grouped into four different categories including containment and closure policies (school closure, workplace closure, public transport closure …), economic policies (income support, fiscal measures..), health system policies (testing policy, emergency investment in health care…), and miscellaneous policies (policies that don't fit). The dataset covers **daily** records per country representing an ordinal scale measurement based on the policy's severity/intensity [6]. Also, it provides a binary flag variable to denote whether the policy is applied on a limited or global scope. It includes textual, date, binary and numeric data.

- **Population by Country - 2020 [7]**

  Population by country dataset covers recent population features of 235 **countries**. It contains several important features such as population, population change, population density, land area, the fertility rate of individual countries, the median age of the county, and urban population. A very well structured data including numeric and textual data.

- **Global Historical Climatology Network-Daily (GHCN-D) dataset [8]**

  GHCN-D dataset covers the need for historical **daily** temperature, precipitation, and snow records for several global regions. It includes several meteorological elements as

daily minimum and maximum temperature, the temperature at the time of observation, precipitation which is a metric for rainfall and snow water equivalent, snowfall, snow depth. The data isn't structured properly, where each row contains a metric reported by a certain station on a specific date. It includes textual (decoded country and station), and numeric data (metrics: min/max temp, precipitation..).

- **Country Health Indicators dataset [9]**
  This dataset covers health indicators relevant to COVID-19 death and infection risks for 128 countries. It contains 70 features **per country** describing COVID-19 cases and deaths, death causes, food sources, health care system, TB vaccine status, school closures, and some facts on people/society. Our study will focus on the health features such as cardiovascular diabetes, blood, endocrines, respiratory, cancer, and other diseases percentages represented per country. It includes textual, numeric, and date data.

6. What is the amount of data preparation needed (missing data, erroneous, small data, insufficiently representative data)?

We represent below the summary of the available features of the aforementioned datasets to study its representation. It is good to mention that we may face missing data while joining the datasets per country and date.

- **Epidemiological data from the COVID-19 outbreak, real-time case information [2]**

  The dataset contains 2,676,311 records with 33 features, we have chosen some of the important features and summarized it. We will encounter missing values for age, sex, and symptoms features.

	age	sex	symptoms	country	date_confirmation	chronic_disease_binary	chronic_disease	travel_history_binary
Count	578018	580157	2052	2676196	2567822	2676311	215	2610732
%	21.60%	21.68%	0.08%	~100.00%	95.95%	100.00%	0.01%	97.55%

We studied more the missing values of the chronic disease feature, and we deduced that this feature is mainly filled out for patients who have chronic diseases.

		chronic_disease	
		count	unique
chronic_disease_binary	FALSE 2676124	69	7
	TRUE 187	146	77

We also noticed that the dataset needs cleaning where the age may be entered as weeks, ranges, and fractions, all the date features are represented as ranges, also chronic disease and symptoms are not cleaned and unique.

Sample features:

- Age: '11-80', '19-75', '21-61','22-60', '105', '14-60', '13-65', '4-64', '2-87', '20-57', '23-71', '6 weeks', '30-61', '0', '0.75', '34-44', '22-66', '6 months', '9 month', '5 month', '11 month'
- Symptoms: 'cough, difficulty breathing, fever', 'fever (38 ° C)', '37.1 ° C, mild coughing', 'cough', 'abdominal pain, pulmonary inflammation', 'Sore throat','feeling ill, coughing'
- Date confirmation: 06.03.2020 - 08.03.2020 , 05.03.2020 - 18.03.2020 …

- **COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University [3]**

The data is presented as reports showing the numbers of confirmed cases, deaths, and recovery grouped by date and country, we need to change it as a raw data with the following features: country, date, confirmed, deaths, and recovery.

- **Coronavirus (COVID-19) Testing [4]**

The dataset contains 45,850 rows and 41 columns, we did a summary over some features that we will use while conducting our study, we noticed that we will encounter missing data in the following features: total tests, test per case, female smokers, male smokers, handwashing facilities...

	total_test	tests_per_case	aged_65_older`	aged_70_older	extreme_poverty	cardiovasc_death_rate	diabetes_prevalence	female_smokers	male_smokers	handwashing_facilities	hospital_beds_per_thousand	life_expectancy
count	16608	16683	40285	40680	26935	40900	42341	32066	31661	19122	36964	45008
%	36%	36%	88%	89%	59%	89%	92%	70%	69%	42%	81%	98%

- **OXFORD COVID-19 Government Response Tracker(OxCGRT) [5]**

  The dataset consists of 64,589 records and 44 features. We investigated the dataset the results show that most all flags have approximately 40-50% missing data and most policy features have about 5% missing data. Below is a sample of policy/flag features:

	C1_School closing	C1_Flag	C2_Workplace closing	C2_Flag	C3_Cancel public events	C3_Flag	C4_Restrictions on gatherings	C4_Flag
count	62529	40549	62455	38658	62489	41677	62493	38668
%	96.81%	62.78%	96.70%	59.85%	96.75%	64.53%	96.75%	59.87%

- **Population by Country - 2020 [7]**

  The dataset contains 235 countries and 11 features, we studied the features and found out that we don't have missing data for this dataset only for Migrants (net) feature where 15% of the data is missing.

- **Global Historical Climatology Network-Daily (GHCN-D) dataset [8]**

  The data is represented as rows where each row contains the station(encoded string), date, and the metric (numeric). We will need to map each station to its specific country, and we will need to change the structure by transposing the rows into columns i.e to have the metrics(temperature, precipitation, etc..) as features. Some features such as snowfall and snow depth are missing for some dates.

- **Country Health Indicators dataset [9]**

  The dataset consists of 180 countries' records and 70 features. The below is a summary of the most important features we will use, the missing data ranges between 10% and 20% as shown below.

	Cardiovascular diseases (%)	Cancers (%)	Diabetes, blood, & endocrine diseases (%)	Respiratory diseases (%)	Liver disease (%)	Diarrhea & common infectious diseases (%)	Malaria & neglected tropical diseases (%)	Nutritional deficiencies (%)	Share of deaths from smoking (%)	alcoholic_beverages
count	165	165	165	165	165	165	165	165	165	151
%	91.67%	91.67%	91.67%	91.67%	91.67%	91.67%	91.67%	91.67%	91.67%	83.89%

7. What data bias issues you might need to mitigate? Explain whether this is applicable or not, and justify why in each case.

COVID-19 containment measures for halting the spread of the virus such as lockdowns, workplaces, and business closing have hurt the economy and unleashed a worldwide economic crisis. Most of the developed countries had the opportunity to be able to have more PCR tests than underdeveloped nations. We think we will face some bias in the number of PCR tests in our datasets depending on the country's economic state. Besides, many believe that the COVID-19 virus is only affecting the elderly and has no major effect on youth and males are more susceptible to the virus than females, thus we might face some data bias on age and gender.

8. For graduate students, which of the following techniques would benefit your ML approach? By this time, you must have read the abstracts of the papers posted under "Research Papers". Some form of probabilistic learning must be part of your choices. Justify all of your choices from the below.

- Meta-learning
- Probabilistic forecasting
- Uncertainty Quantification
- Utility Based Regression

Some of the above-mentioned techniques would be beneficial for optimizing our Machine Learning model. Predictions with deterministic or point forecasts of expected values given some conditional information are sufficient for decision making. However, these predictions don't say much about how much we can trust these results. We are going to handle a complex prediction i.e. the growth of COVID-19 cases in different countries throughout time, hence our predictions should take the form of probability distributions over future quantities or events. We will use Probabilistic Forecasting to express confidence in the predictions, it will significantly

increase the credibility of the results, and will allow us to quantify the distrust or uncertainty in a prediction as an essential component for ideal decision making. Uncertainty Quantification is also another technique that would be useful to implement in our project. It can be used to determine how likely certain outcomes are if some aspects of the system are not exactly known. As mentioned previously, we will use several datasets from different sources that contain a lot of features and variables all over the place with a lot of missing and unorganized entries. We will need Uncertainty Quantification to estimate the error in our data, also it will allow us to have an estimate for the error in our predictions, further improving the reliability of our results. Using both Probabilistic Forecasting and Uncertainty Quantification would allow us to more accurately form a judgment on whether or how much we can trust our results.

References

[1] Xu, B., Gutierrez, B., Mekaru, S., Sewalk, K., Goodwin, L., Loskill, A., ... & Zarebski, A. E. (2020). Epidemiological data from the COVID-19 outbreak, real-time case information. *Scientific data*, *7*(1), 1-6.

[2] https://github.com/beoutbreakprepared/nCoV2019/tree/master/latest_data

[3] https://github.com/CSSEGISandData/COVID-19

[4] https://ourworldindata.org/coronavirus-testing

[5] https://data.humdata.org/dataset/oxford-covid-19-government-response-tracker

[6] https://github.com/OxCGRT/covid-policy-tracker/blob/master/documentation/codebook.md

[7] https://www.kaggle.com/tanuprabhu/population-by-country-2020

[8] https://www1.ncdc.noaa.gov/pub/data/ghcn/daily/by_year/

[9] https://www.kaggle.com/nxpnsv/country-health-indicators