**Rationale**

With the continued spread of the coronavirus around the world, the next few months will be critical to the United States. Upcoming events and holidays – national elections, Thanksgiving and Christmas holidays, Spring semester of classes, etc. – have historically urged human movement and social gatherings, which could lead to outbreaks without the proper policies in place. This project aims to analyze the onset of outbreaks in the past, conditions that led to these outbreaks, and the impact of the pandemic response and public adherence to policies (e.g. social distancing, health and safety protocols), to predict potential outbreaks in the United States and aid in policy development to contain the virus.

**Project Scope**

This project will analyze the spread of the coronavirus based on human behavior and policy implementation around the world but will focus outcome predictions on the United States given current conditions. Additionally, studies on the impact of the virus (e.g. economic) will be limited. Once the pilot has been proven successful, applications to other countries can be further developed in the future.

**Methodology**

Policy data will be reviewed in parallel with public opinion and behavior to determine which factors are the most significant in preventing or spreading the virus. An ensemble model between various types of regression, classification, and clustering will be used. All algorithms will be developed in Python, R, or Microsoft Excel. Visualization techniques will be applied using Tableau. All data will come from either provided data lake API’s or approved external data sources / APIs.

**Data Sources**

*C3.ai Data Lake*

* Policy (KFF, Oxford University) – social distancing policies; government response [James, Justin]
* Mobility data (Apple, Google, UC Berkeley & PlaceIQ) – where do people go; places of exposure [Chase, James]
* Public Surveys (Swayable) – symptom prevalence; demographics; political opinion [Hannah]
* Epidemiology data (IHME, MOBS Lab) – demographics of confirmed patients: race, age, gender, chronic diseases, travel history, etc. [Chase, Hannah]
* Demographic data (US Census, World Bank) – time-series population data and demographic characteristics [Justin]

*External Sources*

* Economic and other demographic data
* News reports and other qualitative information

**Project Timeline**

|  |  |
| --- | --- |
| **Date** | **Objective** |
| 9/28 – 10/04 | * Finalize project proposal |
| 10/05 – 10/18 | * Data collection * Data cleaning and exploratory data analysis |
| 10/19 – 11/01 | * Model development and evaluation |
| 11/02 – 11/08 | * Refresh data sources if applicable * Finalize findings |
| 11/09 – 11/15 | * Prepare submission requirements |
| 11/16 – 11/18 | * Review and submit requirements |