COMP2200 – Group 105

*Nhu Mai Nguyen, Dinh Nghi Dung Le, Tran Chau Ngoc, Thi Kieu Ngan Nguyen*

**Project Proposal**

***COVID-19 social distancing method***

***Summary***

COVID-19 is a global pandemic with a crisis in people and economy when the trend of two objectives is in opposition. The number of deaths and infected persons increased gradually day by day. The government of each nation introduced the policy about social distancing to reduce the increasing cases. The popular methods that many countries apply so as to decrease positive cases are partial or complete lockdown and banned overseas travel. This project will discover the number of cases in different countries and make comparisons between methods such as partial and complete lockdown to figure out the effects of those on the number COVID-19 cases. Additionally, this project considers community transition and overseas travel can be one of the reasons for the cases rising.

***Goals***

The project goal is to analyse different rises of COVID-19 cases among an abundance of nations. The effects of an eclectic mix of factors that may influence COVID-19 case quantity such as lockdown situations, community transmission, and overseas travel would be taken into analysis.

Data selection and cleaning is an essential part of the project since there are heterogeneous lockdown periods for each nation and among different nations. Furthermore, deciding whether to drop some nations with huge missing values which might be potential outliers could also reduce bias.

For data analysis, K-means clustering method is used to cluster countries for different lockdown situations. Regression model and correlation matrix could be applied so as to figure out whether lockdown situation, community transmission and oversea travel has possible relationship on the decrease of COVID-19 cases. Recursive Feature Elimination (RFE) method might be implemented to keep most useful factors for covid-19 growth rate restrain. As the COVID-19 dataset in different nations is high-dimensional and quite large in mean of each variable. Principal Component Analysis (PCA) can be utilized for extracting the dataset from high-dimensional space and subtracting the mean of those and the dataset will be centred on the origin.

The project outcome could play a pivotal role to decipher which factor most affect the number of covid-19 cases, then appropriate methods could be applied in order to reduce covid-19 cases among nations with different characteristics.

***Datasets***

To give the overview of a project, obtaining a dataset is an important process. Dataset was obtained via Johns Hopkins University and University of Oxford which collected data related to COVID-19 precisely from the day the outbreak started until now.

The data are .csv datasets with all countries around the world. Because this project will focus on the effect of partial/complete lockdown and oversea travel, so that the sample of appropriate countries using one of these lockdowns will be used for data analysis. The COVID-19 remains until now but has a slightly positive effect on the number of deaths, hence 2020 is the milestone this project uses to analyse.

***Method***

The analysed techniques intended to use in the project to measure and evaluate the dataset may include PCA, K-mean, Naïve Bayes, Logistic Regression and RFE. The main reasons for using those methods are to classify of each factor such as the number of people getting Covid-19 in different countries and the number of Covid-19 cases decreasing after executing a partial or full lockdown in specific countries and to predict the effectiveness after executing restriction lockdowns as well as whether the number of Covid-19 cases in the future in different nations will decrease after implementing some solution for preventing Covid-19 contagious. Furthermore, we also clarify whether community transmission or oversea travel could be potential reasons for the increase of COVID-19 cases.

***Milestones***

**Week 8**:  Discussing project ideas and deciding our topic. Also searching for some sources of data is necessary to draw general vision for what works need to be done for analysis. The proposal of the project is planned during our discussion and will be done this week.

**Week 10**:  All the preparations for data analysis is expected to be done including cleaning, data selecting and formatting.  This is where K-means analysis will be applied

By this week, we are expected to finish building the project algorithms such as regression, correlation matrix and running test on the model to determine model performance.

**Week 13**: All tasks should be finished by this week and accomplished the video presentation.