**CMPT 459**

**Milestone 1: Report**

**Martin Ester**

**Mohammad Raad Sarar 301326232**

**Kazi Zubayer Kader**

**Ahmed Irtaza Haque**

* 1. **Exploratory Data Analysis**

All the attributes in *cases\_train.csv* and *location.csv* were read in by *eda.ipynb* and each attribute was analyzed independently of each other (except for latitude and longitude). Each attribute contains number of missing values. Additionally, for each numerical attribute, count, mean, standard deviation, min, max and the interquartile ranges were printed along with histograms to understand distribution, and box and whisker diagrams to understand outliers. For latitude and longitude values, scatterplots were printed and overlayed over an image of the world map to make sure latitude and longitude values are accurate. For each categorical attribute, histograms were created to understand the distribution of the unique values. Columns in location.csv (case fatality ratio, recovered, active, incidence rate, deaths) were very skewed and could probably contain outliers. The active column also contained negative numbers. Rows with negative numbers were removed first and then statistics and visualizations were created.

* 1. **Data cleaning and Imputing missing values**

We removed only 2 rows from the *cases\_train.csv* where latitude and longitude values were not present. Columns *additional\_information* and *source* were dropped as we believe the text data would not be of much help during the classification phase as a lot data was missing. Cleaning and imputing methods are described below:

**Age:** The values in the form of (*x-y*) were found by python regular expression and the middle value between x and y was assigned to the age. The values like *z-months* were found by python regular expression and replaced by *z*. The values like *a+* or *a-* were also found similarly and replaced by *a*. To **fill in missing values,** a normal distribution was created of the values that were not missing and age was assigned according to the probability determined by the mean and standard deviation of the normal distribution *(np.random.normal(age\_mean, age\_std, 1))*.

**Sex:** Before imputing the sex values, we figured out how many male and female values there were in the values that were not missing, we created a probability of male and female according to the counts and imputed the values according to that probability.

**Country:** We found that only a few rows (18) with attribute country was missing. So we printed them all. They all had province as “Taiwan”. So we assigned “Taiwan” as the country (not sure if Taiwan is the country or China, Google was a little unclear) of the rows that had country missing but had Taiwan as a province.

**Province:** For rows that did not have a province, we checked the country (as after the previous step all rows have a country value), and found the province with the highest occurrence (*mode*) of that country and assigned it as province. Some countries did not have any provinces listed, in that case finding the mode was not possible. So instead of leaving the province field empty or removing the entire row, we decided to replace the Null value with *“unknown”.* eg. (province: unknown, country: Togo).

**Date Confirmation:** We used a similar approach to the *province* *attribute* as we assumed countries have release their date of confirmation in a large batch. So, for missing date confirmation, we found the country first and then found the mode of the date confirmed of that country and imputed with that value.

* 1. **Dealing with outliers**
  2. **Transformation**
  3. **Joining the cases and location dataset**
  4. **Outcome labels**

The different outcomes and their meanings are:

*Recovered*: Individual contracted covid-19 but has tested negative afterwards.

*Hospitalized*: Individual has contracted covid-19 and been admitted to a hospital, but has neither died nor recovered yet.

Nonhospitalized: Individual contracted covid-19 but did not require medical attention. Nor recovered or dead

Deceased: Individual passed away (may they rest in peace) due to covid-19.

The data mining task that predicts outcome labels is **classification**