IS517 Project Proposal
World Health

By Abhinav Choudhry and Kerstin Wolf

Background and Motivation

This pandemic year has brought the topic of health to the world's forefront like nothing before, and it has also highlighted the differences among nations. European nations suffered more COVID-19 deaths in part because their high life expectancies meant that they had older people and more people suffering comorbidities. Most developing nations have had much fewer case and death counts, but it seems suspiciously correlated with their markedly lower life expectancy as well as fewer tests per capita and the paucity of reliable data. This made us wonder whether we could predict what caused these differences in key health outcomes. Were they related to health infrastructure or health behaviors or to a mixture of both? Are there particular parameters that might have greater predictive value for health outcomes than the income-based classifications typically employed? These questions underlie our motivation to explore the World Health Organization's 2020 statistics on global health.

Datasets

We will be utilizing two datasets for this project. The main dataset is World Health Statistics 2020 from Kaggle. This dataset contains 39 CSV files with a total of 186 columns. The data within these files range from air pollution death rate, road traffic death rate, and violence against women to alcohol consumption, number of dentists, and number of pharmacists. In most files, there are 184 unique countries; although, a handful of the CSV files are organized instead by region. Some files contain multiple years of data ranging from 1950 to 2019. A number of the files also break the data down by sex (male, female, both sexes) as well. An interesting quality of this dataset is that there are no null values. However, a number of values are listed as zero. For example, Afghanistan has it listed that there were no suicides in 2016. This seems very unlikely. We believe that a zero value in this dataset often signifies that the data wasn't collected. Another interesting quality is that some values list a range. This can be seen in the maternal mortality ratio CSV file where the ratio per 100,000 live births is listed as a single value followed by the range in brackets. For example, the maternal mortality ratio per 100,000 live births in Afghanistan in 2017 was 638 [427 - 1010].

The secondary dataset used in this project is the Gross Domestic Product 2019 data from The World Bank. This is a significantly smaller dataset that's only one CSV file at about 139 rows with 4 columns. The column attributes are country abbreviation, ranking, country name, and GDP (in millions of US dollars). Besides just countries, there are also GDP totals for each region as well. Since the attribute we will be using from this dataset is GDP, our data is numeric.

Research Questions

We have three research questions that will be addressed in this project:

1. Would clustering countries on the basis of healthcare factors produce different clusters than using economic factors (GDP)?
2. Are we able to successfully predict life expectancy at birth and healthy life expectancy at birth by a) health infrastructure and b) health behaviors?
3. Is there a significant difference between healthy life expectancy at birth and life expectancy at birth? What factors affect this difference the most?

Here, health infrastructure refers to attributes such as the number of doctors or dentists per unit population as well as drinking water or hand washing facilities available. Health behavior refers to population behaviors concerning health, such as incidences of substance abuse, violence against women, etc.

Data Analysis Plan

Our data analysis would broadly include the following stages: a.) select datasets, b.) evaluate datasets, c.) clean data, d.) combine datasets, e.) split data, f.) select features, g.) split model, h.) fit model, i.) refine model, j.) test model, and k.) interpret results.

Although the whole data set has 39 individual csv files, some of them are not relevant to our research plan. Therefore, the first step is to select the attributes relevant to our analysis from the files into a custom new file. This is also when we take from our time series data only the latest two years for which all the relevant csv files have data, probably 2017 and 2016. We would follow with data cleaning since not every nation might have reliable data on each attribute and this may cause outliers as well as well as faulty "zero" values. We would use outlier analysis and also check for each attribute the real world significance of zero values in order to understand which data points might be anomalous.

We intend to split our data equally between the training and test set using the penultimate year (2016) as the training set and latest year for testing. For RQ1, we would use clustering methods to compare clusters formed on health parameters versus those formed on the basis of GDP. For RQ2 and RQ3, we would use a number of methods that could be linear regression, LDA, QDA, KNN, and generalized additive models.

Expected Results

While health infrastructure might be related to national income or national priorities, health behaviors could be related to levels of education as well as cultural factors. The results of our analysis could potentially guide us towards a better model for predicting life expectancy and healthy life expectancy and also lend insights on which methods work better for such health datasets.

Rating of Partner Contribution

Abhinav and I both did an equal amount of work on this project. Abhinav really excels at thinking deeply about our data and our research questions, and I really organized and designed our slide presentation. We selected the datasets together, and we equally contributed and wrote this project proposal together. Our skills and interests really complement each other, so it’s been great working together on this project.