

Chris Weilacker, Kirk Kosinski, Patrick Cao, OJ Alcaraz
CST 383 Introduction to Data Science
26 February 2021

Covid Morbidity Factors Exploration Project

Since the beginning of 2020, the world has been under siege by the most severe pandemic in generations. With 2.5 million people succumbing to COVID-19 worldwide, and over 500 thousand in the US alone (Allen et al.), the pandemic is a topic of great concern. We decided to apply what we have learned about data science this semester to understanding COVID-19. Specifically, we attempted to create a model to predict the risk of mortality in a country based on features including the country's percentage of residents over age 65, percentage of residents suffering from obesity, and life expectancy.

The former two features were chosen based on the data we could find publicly available, and on information from the CDC, which has warned that older adults and people with medical conditions are “more likely than others to become severely ill” from COVID-19 (“COVID-19 and Your Health”). Additionally, news media including The New York Times have widely reported the extreme impact COVID-19 has had at nursing homes and long-term care facilities (Allen et al.), suggesting a likely link between age, other medical conditions, and the risk of mortality from COVID-19. The latter feature mentioned above, life expectancy in a country, was chosen because the quality of healthcare differs by country, and that might be reflected in the life expectancy numbers.

Our initial analysis used several datasets downloaded from Kaggle. Once the datasets were merged into one Pandas DataFrame, we had to remove a number of rows that were missing one or more features, and/or the target of COVID-19 mortality. The data also needed to be

scaled. The initial results seemed potentially promising but we were concerned about outdated COVID-19 data as the initial data we used was five months old.

After further discussion and research, we added more features by including a dataset provided by Our World in Data (“Owid/Covid-19-Data”). This dataset is updated very frequently, incorporating data from Johns Hopkins University, United Nations Development Programme, and other reputable sources. Utilizing a high risk classification target that was determined via those countries whose COVID-19 death per million population was greater than 0.65 std deviations representing about 25% of the countries, we evaluated over 12 features provided in the dataset using kNN and a range of k values. The following four features showed an accuracy of about 94% using kNN with k=7:

- human_development_index
 - Description: A composite index measuring average achievement in three basic dimensions of human development—a long and healthy life, knowledge and a decent standard of living.
 - Source: United Nations Development Programme (Our World in Data dataset)
- Female_smokers
 - Description: Share of women who smoke, most recent year available
 - Source: World Bank World Development Indicators, sourced from World Health Organization, Global Health Observatory Data Repository
- Obesity
 - Description: Prevalence of obesity among adults
 - Source: World Health Organization (Kaggle dataset)
- Life_Expectancy

- Description: Life expectancy at birth in 2019
- Source: James C. Riley, Clio Infra, United Nations Population Division

We then implemented a cross_validation scoring utilizing a cv of 8 to make sure the randomness of our train/test data was not being affected and got something similar, but with a best mean accuracy score of 89.22% with k=3 and 5 features including the following:

- Male_smokers
 - Description: Share of men who smoke, most recent year available
 - Source: World Bank World Development Indicators, sourced from World Health Organization, Global Health Observatory Data Repository
- Obesity
 - Description: Prevalence of obesity among adults
 - Source: World Health Organization (Kaggle dataset)
- Cardiovascular Death Rate
 - Description: Death rate from cardiovascular disease in 2017 (annual number of deaths per 100,000 people)
 - Source: Global Burden of Disease Collaborative Network, Global Burden of Disease Study 2017 Results
- aged_70_older
 - Description: Share of the population that is 70 years and older in 2015
 - Source: United Nations, Department of Economic and Social Affairs (Our World in Data dataset)
- Latitude
 - Description: GPS coordinates for every world country

- Source: Paul Mooney (Kaggle dataset)

Overall we are satisfied with the results we obtained from the relatively simple kNN algorithm, we were able to create several visualizations in the Jupyter Notebook with scatterplots comparing the different features to the high risk category that we found to produce the best model. As we hypothesized, age and obesity do seem to be factors in a country's mortality rate for Covid-19 with extra features such as smoking, cardiovascular disease, and life expectancy helping improve the results further. We were surprised, however, by the significance of latitude. Researching further we found a study that also noticed a correlation with latitude and COVID-19 mortality with it hypothesising having something to do with vitamin D levels as people closer to the equator would be getting more sun (Whittemore, 2020). It is worth further investigation if a national vitamin D supplement program can improve outcomes for possible future pandemics. It is also further evidence that something should be done to fix the obesity epidemic in the US as it is a significant contributing factor to our disproportionate COVID-19 death rate.

References

- Allen, Jordan, et al. "Coronavirus in the U.S.: Latest Map and Case Count." *The New York Times*, 25 Feb. 2021, www.nytimes.com/interactive/2020/us/coronavirus-us-cases.html.
- Arora, Aman. "Obesity among Adults by Country, 1975-2016." Kaggle, 8 May 2020, www.kaggle.com/amanarora/obesity-among-adults-by-country-19752016.
- "COVID-19 and Your Health." *Centers for Disease Control and Prevention*, 4 Jan. 2021, www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/index.html.
- Kruk, Matias. "DEMOGRAPHIC AND SOCIO-ECONOMIC (UNESCO)." Kaggle, 29 Nov. 2019, www.kaggle.com/krukmat/demographic-and-socioeconomic-unesco.
- Mooney, Paul. "COVID-19 Case Mortality Ratios by Country." Kaggle, 25 Sept. 2020, www.kaggle.com/paultimothymooney/coronavirus-covid19-mortality-rate-by-country.
- "Owid/Covid-19-Data." Our World in Data, 16 Feb. 2021, github.com/owid/covid-19-data/tree/master/public/data.
- Whittemore, P. B. (2020, June 26). COVID-19 fatalities, latitude, sunlight, and vitamin D. US National Library of Medicine, National Institute of Health. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7319635/>
- "World Health Statistics 2020|Complete|Geo-Analysis." Kaggle, 25 Jan. 2021, www.kaggle.com/utkarshxy/who-worldhealth-statistics-2020-complete.