Predictive Analytics for Covid-19 (PAC)

Kim Barrios, Krutal Patel, Michael Puiefoy, Jonas Simmons
Johns Hopkins University Applied Physics Laboratory

Brief Summary/ Abstract

- The Covid-19 Pandemic has directly affected over 596 million people worldwide. However, the response from each country has had varying outcomes on both the number of covid cases and the number of covid deaths. This project aims to explore a large dataset collected by the CDC, in collaboration with the World Health Organization, and OurWorldinData editors and researchers.
- This dataset contains critical information on population demographics, health history, and socioeconomic metrics for each country.
- This project aims to analyze this data and extract critical patterns and relationships between the various factors and the risk for covid in each country

Goal/ Objective

- Apply both unsupervised and supervised machine learning models to classify countries on a covid severity index based on their socioeconomic and health factors.
- This model and analysis will allow Researchers to predict the impact of a pandemic in a country given its current standard of living, economic and health resources, and current health statistics.

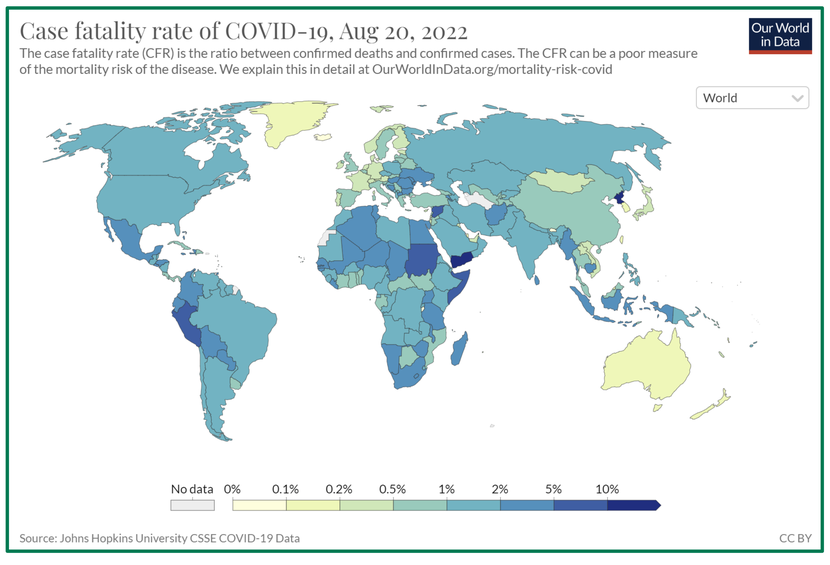


FIGURE 1: Case fatality rate of Covid-19 to Aug 20, 2022 by country. This map highlights the various Covid severity index groups created in Data Cleaning and Data extraction steps.

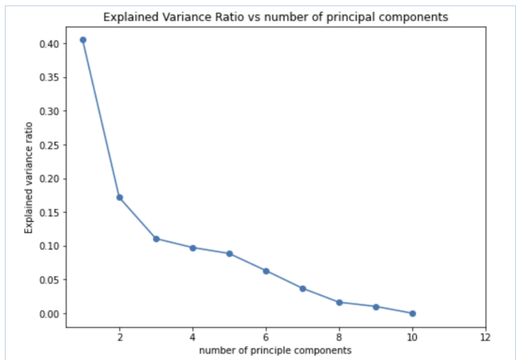
Methods / Process Pipeline

Data Cleaning

Organize data from time series to static data for each country.
Handle missing/null data
Compute statistics and visualize distributions of the features

Data Classification

Generate a severity index classification group based on number of covid cases and number of covid deaths in a country



Applied Principle Component Analysis

Determined n=8 number of principle components best describe the variance in the data

K-means Clustering

Unsupervised classification approach to extract number of groups best explained by the variance in the data

K-Nearest Neighbors

Supervised classification approach to predict covid severity level from features, including gdp, population, and human development index

Model Evaluation and Metrics

Applied k number of nearest neighbors test to determine most optimal value for k. Tested for model accuracy.

Feature Analysis / Visualization

- Some of Features from the Dataset include Human Development Index, Diabetes Index, Population Density, and GDP.
- We studied the distributions of each feature to extract any outliers in the data as this may skew the model and cause the model to be an overfit.

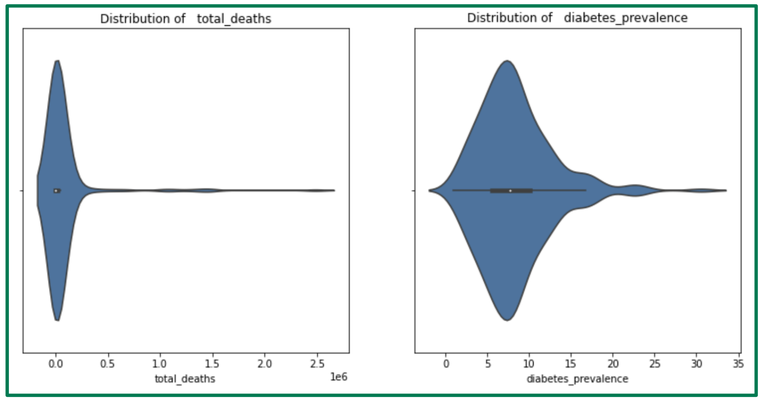


FIGURE 2: Two violin plots of Total Deaths by country and Diabetes prevalence by country.

Results

- Applied K-means clustering to determine number of optimal groups or clusters formed.
- Applied k-Nearest Neighbors, a supervised machine learning algorithm, to predict a country's covid severity index based on its features, as described earlier → Received an avg accuracy of 0.6.

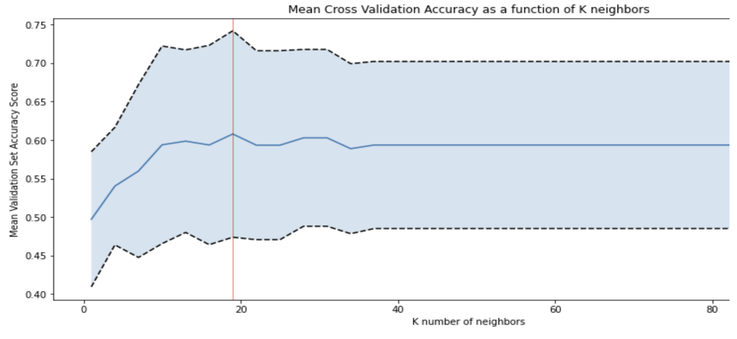


FIGURE 3: Plot of Mean validation set accuracy score using k fold cross validation as a function of k number of neighbors.