Self-Study Session

Data Mining

Data

In this exercise, you are going to analyze data about the COVID-19 pandemic. The data is presented in the *Data_COVID_OWID_per_capita.csv* file. It contains information on 203 countries and regions.

All data used in this exercise was obtained from 'Our World in Data'. The website is a collaborative effort between the researchers of the Oxford Martin Programme on Global Development at the University of Oxford, and the non-profit organisation Global Change Data Lab. We have obtained data on a broad range of health, political, economic, pollution, and trust variables. A description of the variables is presented in Table 1.

The data contains several missing values that have to be accounted for in the analysis.

Analysis

The goal of the self-study is to use the algorithms and tools that we have discussed during the lectures to understand the effect of the pandemic better.

For that purpose, you should prepare a 5-pages summary with a summary of your analysis. The report could contain graphs, tables, and results from the methods you consider to explain the data well. You can think of the report as something you will present to the authorities responsible for controlling the pandemic.

You are free to use any method and variables that you consider pertinent for the analysis. Ideas could include:

- Explain the number of cases or deaths due to the pandemic using a regression algorithm considering several of the other variables.

  Note: There is an obvious, and so perhaps not that useful, correlation between the number of cases and deaths. That is, if we use cases to explain deaths, any other variable may no longer seem important.

- Classify the countries in worst and not-so-bad hit. You could create a variable labelling the countries with the top X% of deaths, and use classification algorithms to evaluate their properties against the bottom X% to explain what makes them different.

- Analyze how well does your models predict the effect of the pandemic in Denmark. That is, compare a predicted value given the explanatory variables for Denmark and compare against the true number of cases and/or deaths.

Data Description

Table 1 presents an overview of the data considered in this study.

Name of variable	Description
Pollution	Population-weighted average level of exposure to concentrations of suspended particles measuring less than 2.5 microns in diameter. ($\mu g/m^3$).
DeathsPollution	Number of deaths per 100,000 population from both outdoor and indoor air pollution. Age-standardized.
*PM*25	PM2.5 air pollution, mean annual exposure (micrograms per cubic meter).
OZONE	Ozone air pollution, mean annual exposure (particles per billion).
SmokeDaily	Estimates of the prevalence of daily smoking, defined as the percentage of men and women, of all ages, who smoke daily.
Drinking	Share of adults aged 15 and older who drank any form of alcohol within the previous 12 months.
UnsafeWater	Share of deaths from unsafe water sources.
Sanitation	Death rates from unsafe sanitation measured as the number of deaths per 100,000 individuals.
Overweigth	Share of adults that are overweight or obese.
Cardiovascular	Annual number of deaths per 100,000 people from cardiovascular disease.
Diabetes	Diabetes prevalence (% of population aged 20 to 79).
*Aged*65	Share of the population that is 65 years and older.
*Aged*70	Share of the population that is 70 years and older.
HospBeds	Hospital beds per 1,000 people (OECD, Eurostat, World Bank, national government records and other sources).
Corruption	Transparency International's Corruption Perception Index. Scores are on a scale of 0-100, where 0 means that a country is perceived as highly corrupt.
TrustShare	Share of respondents who answered 'a lot' or 'some' to the question: 'How much do you trust your national government?'
TrustMedics	Share of people who trust doctors and nurses in their country.
Literacy	Estimates of the share of the population older than 14 years that is able to read and write.
HumanRights	Degree to which governments protect and respect human rights. The values range from -3.8 to around 5.4 (the higher the better).
PoliticalRegime	The scale goes from -10 (full autocracy) to 10 (full democracy).
GiniIndex	Gini Index. World Bank inequality data. A higher Gini index indicates higher inequality.
EconomicFreedom	Calculated by the Fraser Institute. Measures the degree to which individuals are free to choose, trade, and cooperate with others. Scores are on a scale of 0-10, where 10 represents maximum economic freedom.
HealthShare	Public health expenditure (%GDP).
PopDensity	Number of people divided by land area, measured in square kilometers.

$GDPpcp$	Gross domestic product at purchasing power parity (constant 2011 international dollars).
$Poverty$	Share of the population living in extreme poverty, most recent year available since 2010.
$ReproductionRate$	Reproduction rate of the virus (R).
$TotalVaccinations$	Total number of people who received at least one vaccine dose per 100 people in the total population.
$FullVaccinations$	Total number of people who received all doses prescribed by the vaccination protocol per 100 people in the total population.
$TotalCases$	Total confirmed cases of COVID-19 per 1,000,000 people as of April 6, 2021.
$TotalDeaths$	Total deaths attributed to COVID-19 per 1,000,000 people as of April 6, 2021.

Table 1: Data considered. Source: Our World in Data.