How our health makes us susceptible to COVID-19
Springboard Capstone 2 Project

LISA HAHN-WOERNLE
Data Scientist Bootcamp @ Springboard
August 13, 2020

1 Introduction

COVID-19 has changed the world as we know it. Airports, schools and offices closed. What seemed impossible is the new normal. But why are some countries hit more seriously than others? Culture, climate, social structures, health, ... these are just a few factors that can influence the spread of the disease. In this project, the susceptibility to COVID-19 and especially the fatalities due to COVID-19, is analyzed according to a country's health and demography.

2 Data

	count	mean	std	min	25%	50%	75%	max
Confirmed	147.0	0.073584	0.118237	0.000117	0.003848	0.016204	0.101505	6.296591e-01
Deaths	147.0	0.004182	0.011031	0.000000	0.000105	0.000417	0.002382	7.436310e-02
Cardio Death Rate	147.0	0.256981	0.119594	0.079370	0.162807	0.243811	0.324205	7.244167e-01
Diabetes Percentage	147.0	7.303469	3.760702	0.990000	4.800000	6.930000	9.030000	2.202000e+01
Obesity	147.0	18.534694	9.416538	2.100000	8.200000	21.900000	25.700000	3.730000e+01
Undernourished	147.0	10.653061	12.095308	1.000000	1.000000	6.500000	14.600000	5.960000e+01
PopMale	147.0	1.552883	1.258690	0.185446	0.461343	1.098134	2.546442	5.512896e+00
PopFemale	147.0	2.386105	2.062953	0.095829	0.663079	1.577732	3.994760	8.528817e+00
PopTotal	147.0	3.938988	3.281700	0.301110	1.126508	2.678089	6.688357	1.404171e+01
Total Population	147.0	49803.233163	167285.832961	110.593000	4226.758500	11263.079000	35370.147500	1.433784e+06

Figure 1: Properties of the numerical features in the COVID-19 data set. All features are relative to the Total Population (in thousand people).

2.1 COVID-19 records

The COVID-19 data is based on the COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University on github (`https://github.com/CSSEGISandData/COVID-19`). The study is based on the global COVID-19 of May 9, 2020. While the notebooks could be reused for any other date, May 9, 2020 is the date this capstone project was first implemented.

The focus of this study is the mortality rate of the virus for a given country and its potential relationship to the national demography and health statistics. Therefore, only the number of confirmed cases and of deaths are of interest to this study.

2.2 Health records

Here, the health of a nation is analyzed according to its most recent data of cardiovascular disease death rates, diabetes prevalence, and obesity and undernourishment rates.

Cardiovascular disease death rates and **diabetes prevalence** are sourced from `ourworldindata.org` and are based on the 2017 statistics. **Obesity** and **undernourishment rates** are sourced from the Kaggle data set by Maria Ren (`https://www.kaggle.com/mariaren/covid19-healthy-diet-dataset`).

2.3 Demography

Population size and gender-specific population percentages in the age group 75 and older are sourced from the UN population data set (`https://population.un.org/wpp`) for the year 2019.

3 Method

Data processing, analysis and modeling are all done with python in Jupyter Notebooks. If not stated differently, methods of the sklearn module are used.

First, the data set was reduced to the countries with COVID-19 cases. The resulting data set had no missing values and the only outlier was the US with its high number of confirmed COVID-19 cases. Second, the data set was scaled with the StandardScaler method and three clusters were derived with the KMeans method. The number of clusters was chosen based on the Elbow method. Third, the data was inspected with the Principal Component Analysis (PCA) method to verify the correctness of the clustering and visualize the data set in 2D. Fourth, the target, COVID-19 death rate,

was separated from the data set, and both were split in a training (75 %) and testing (25 %) set. In the fifth and final step, the split data set was used to train and test different models: Linear Regression Model, Decision Tree Regressor, Random Forest Regressor, and Gradient Boosting Regressor. As performance measure the R^2 value and the mean squared error (MSE) are used.

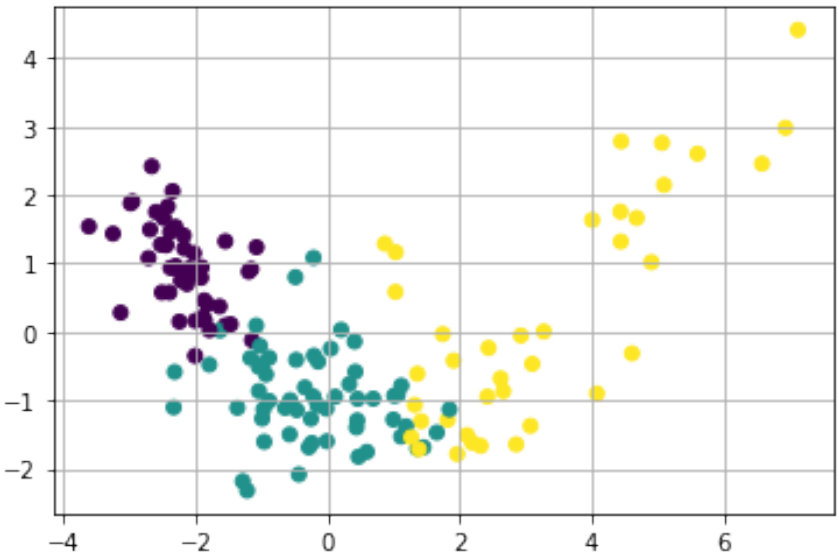

Figure 2: Projection of the 11-dimensional dataset onto the 2 main components of the principal component analysis with the color coding according to the 3 clusters determined with the KMeans method.

3.1 Linear Regression Model (LRM)

For the linear regression modeling the data is preprocessed with the PolynomialFeatures method to describe the data as a second order polynomial function (without bias). No further parameter tuning was performed.

3.2 Decision Tree Regressor (DTR)

Tuning with the GridSearchCV method lead to the following parameters for the Decision Tree Regressor: max_depth = 3, max_features = 6, min_samples_leaf = 3. All other parameters remained at their default value.

3.3 Random Forest Regressor (RFR)

The RandomForestRegressor method was used with 250 estimators and all other parameters set to their default values.

3.4 Gradient Boosting Regressor (GBR)

Based on a GridSearchCV tuning analysis, the following parameters for the GradientBoostingRegressor were set: max_depth = 3, max_features = 3, and n_estimators = 150.

4 Results

The goal of the modeling is to identify i) whether a relationship between the COVID-19 fatalities and the health statistics and demography data exists, ii) how strong this relationship is, and iii) which features are most relevant for the prediction of COVID-19 fatalities.

4.1 Model Performance

Table 1 lists the performance for all regression models as described in the method section. The Random Forest Regressor performs best in R^2 and MSE, but if computation time is a critical condition, the Gradient Boosting Regressor might be preferred despite its slightly weaker performance.

Model	R^2	MSE	cost [ms]
Linear Regression	0.6030	0.7643	3.8
Decision Tree	0.6088	1.001	9.4
Random Forest	0.7173	0.7233	26.4
Gradient Boosting	0.6961	0.7778	5.0

Table 1: R^2, mean squared error (MSE) and computation costs for the different models used in the analysis.

4.2 Feature Importance

Both the Random Forest and the Gradient Boosting Regressor provide the feature importance as a percentage in relationship to all features. As could have been expected, for both models the number of confirmed COVID-19 cases is the most important feature. The RFR is dominated by this feature (74.5 % for the confirmed cases), followed by Diabetes (9.1 %) and male

population older than 75 (7.1 %). Obesity (3.1 %) and cardiovascular death (2.4 %) rank 5th and 6th after the total population size (3.2 %).

In the GBR, the confirmed cases make up only 41.2 %. The male population older than 75 (20.5 %) and Diabetes (13.4 %) are again in the top 3, but in reverse order than in the RFR. Rank 4 is again the total population size (8.5 %), followed by the cardiovascular death (6.2 %) and obesity (5.4 %). The latter are again in the reverse order compared to the RFR. Undernourishment appears to be not suited as indicator for the fatality of COVID-19 (RFR: 0.4 %, GBR: 1.8 %).

Figure 3 shows the relationship between the death rate and the top 4 important features according to RFR and GBR. Generally speaking, there is no strong trend in the data but higher death rates occur more likely in nations with more confirmed cases, more males older than 75, and a larger population. Cluster #2 appears to contain countries with high fatalities, many confirmed cases, and the largest male population older than 75. Cluster #0 contains countries with Diabetes cases and a small male population older than 75+.

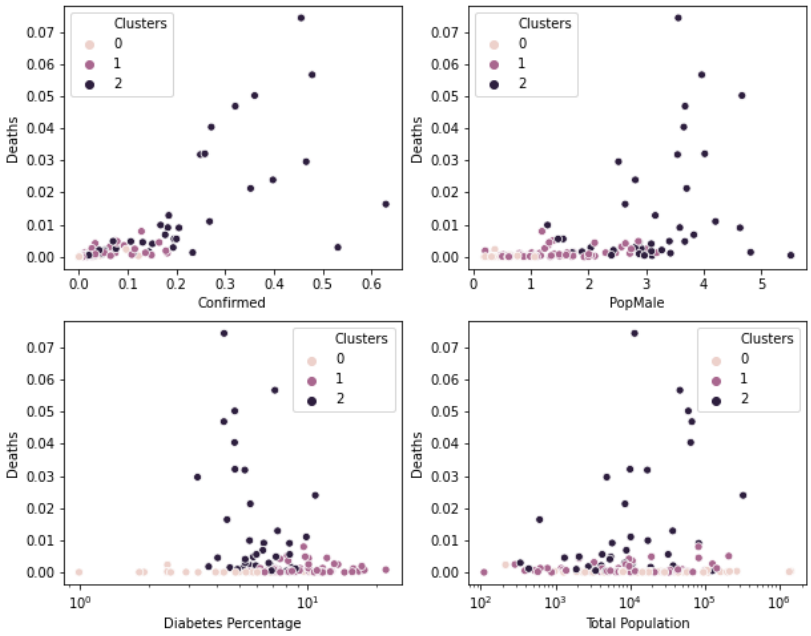

Figure 3: Relationship between the death cases and i) confirmed cases, ii) male population (>75), iii) Diabetes (logarithmic x-scale), and iv) total population (logarithmic x-scale).

5 Conclusions and Recommendations

The data analysis and modeling has revealed that next to the number of confirmed cases the percentage of the population that is male and older than 75, and the diabetes rate of a nation correlates to the fatality of COVID-19. Whether this is due to the effect of COVID-19 on elderly males and diabetes patients, or rather due to the fact that wealthier countries have higher diabetes rates and an older population, cannot be determined from this data set. But since obesity is also a sickness of the wealthy people, and since obesity ranks low, it is still very likely that elder males and diabetes patients are more susceptible to COVID-19 compared to the rest of the population.

Based on this study, it is recommended to particularly protect the male elderlies and diabetes patients from contact with the COVID-19 virus until a cure or vaccination has been found. Further analysis is needed to determine to which degree these two features represent wealthy countries and to which degree the correlation is purely health related. Additionally, more recent data should be analyzed since it will extend the data set to more nations, unfortunately, including more developing countries. This shift in the distribution could lead to different modeling results and hence different feature importances.