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

LISA HAHN-WOERNLE
Data Scientist Bootcamp @ Springboard
August 21, 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 demographics.

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 demographics 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 Demographics

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.

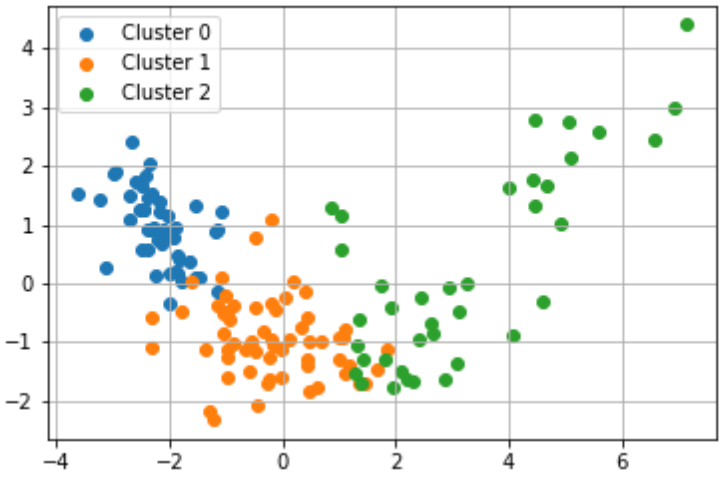


Figure 2: Projection of the 11-dimensional dataset onto the 2 main components of the principal component analysis with the color coding according to the 3 clusters determined with the KMeans method.

3.1 Motivation of Model Choice

The key question of this project focusses rather on the feature importance than on the predictive power of a model. Still, the predictive power is seen as a measure on how reliable the choice of features and their importance are. The **2nd degree polynomial regression model** (2dPRM) is the result of a *GridSearchCV* hyperparameter tuning step which was motivated by the polynomial shape of the PCA (Fig. 2). While the 2dPRM is cheap and easy to understand, it is clearly limited to the polynomial relation of the underlying data and is also prone for overfitting. Therefore, the other three models are based on the decision tree principle, which does not rely on linear or polynomial relations among features. To start off, the simple **Decision Tree Regressor** is used since it is easier to understand than the other, more complex models like the **Random Forest Regressor**. Random Forests, as the name says, rely on an ensemble of randomly selected decision trees. This

makes the model more accurate and less prone to overfitting. The **Gradient Boosting Regressor** is similarly structured as the Random Forest, but learns from each decision tree to improve the next. While this can lead to more accurate results, overfitting might again become an issue.

For all models, but especially the latter two ones, the small data set, with its 147 countries and 9 features, might limit the predictive power. Still, the choice of models provides insight from two different angles: one with a classical regression model and three different approaches using more abstract and modern machine learning techniques.

3.2 2nd Degree Polynomial Regression Model (2dPRM)

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). Based on the *GridSearchCV* method, the 2nd degree polynomial was found to have the best predictive power. A 2nd degree polynomial is described as follows:

$$f(\vec{x}) = \sum_{i,j}^{N(i\leq j)} [a_{i,j} \cdot x_i \cdot x_j] + \sum_{i}^{N} [b_i \cdot x_i] + c \ , \tag{1}$$

where N is the number of features x and the coefficients a ($\frac{1}{2}N[N+1]$), b (N) and c (1) are determined during the fitting procedure.

3.3 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.

3.4 Random Forest Regressor (RFR)

The RandomForestRegressor was tuned to 250 estimators and all other parameters were left at their default values.

3.5 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 demographics data exists, and ii) 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 Polynomial Regression Model has the best predictive power as measured with R^2 and MSE. While the Gradient Boosting Regressor is faster in prediction, the difference is so too small to make an argument for the weaker performing model. The Decision Tree performs the worst which stresses that the ensemble method improves the predictability and reduces the risk of overfitting.

Model	R^2	MSE	time fit [ms]	time predict [ms]
Polynomial Regression	0.7643	0.6030	14.1	3.1
Decision Tree	0.6088	1.001	34.6	9.4
Random Forest	0.7173	0.7233	520	26.4
Gradient Boosting	0.6961	0.7778	87.8	1.7

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

4.2 Feature Importance

The six largest coefficients (positive and negative) of the trained Polynomial Regressor based on Equ. 1 are given as:

$$f(x) = \text{Cluster}_2 \cdot \begin{pmatrix} +1.28\ \text{Undernourished} \\ -1.15\ \text{Confirmed} \\ -0.79\ \text{Male}_{75+} \\ -0.62\ \text{Diabetes} \end{pmatrix} + \text{Confirmed} \cdot \begin{pmatrix} +0.72\ \text{Male}_{75+} \\ +0.79 \end{pmatrix} + \dots . \quad (2)$$

This result clearly identifies the number of confirmed COVID-19 cases and of males of age 75 and older as a key feature for the modeling. Further more, $Cluster2$ appears to bear relevant information.

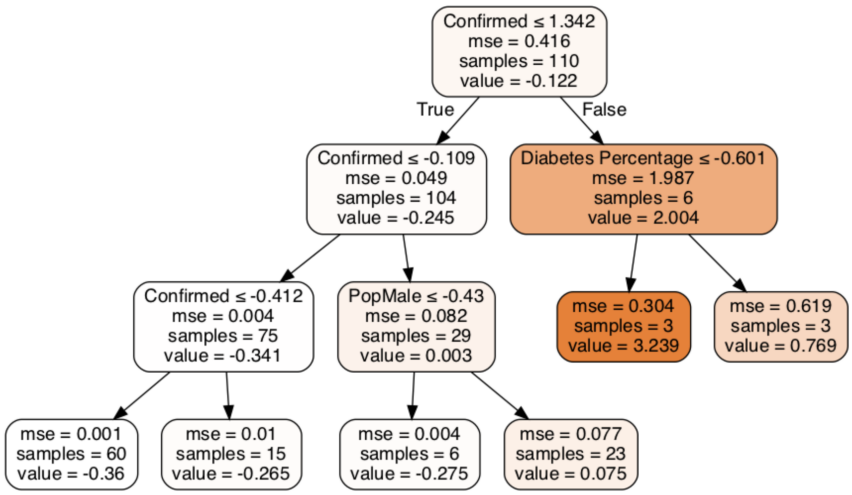


Figure 3: Setup of the trained decision tree with a maximum depth of 3 and a minimum leaf size of 3.

Figure 3 shows the trained setup of the Decision Tree. The setup already visualizes the dominant importance of the confirmed COVID-19 cases (76.3 %), and the lower importance of diabetes (22.3 %) and the male population older than 75 (1.4 %). Even though the feature number was only limited to 6, the model actually relies on just 3.

Both the Random Forest and the Gradient Boosting Regressor provide the feature importance as a percentage in relationship to all features (Fig. 5). 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 %) (Fig. 5)a.).

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 %) (Fig. 5)b.).

Figure 4 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+.

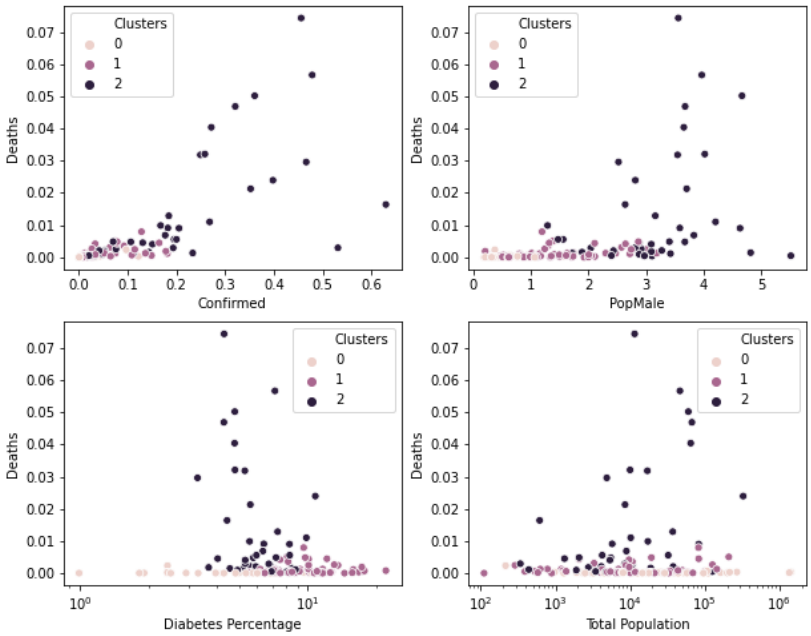
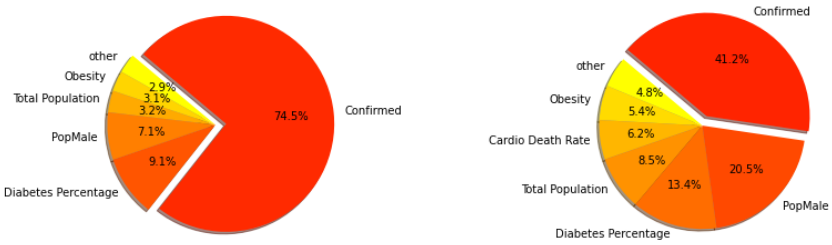


Figure 4: Relationship between the death cases and i) confirmed cases, ii) male population (>75), iii) Diabetes (logarithmic x-scale), and iv) total population (logarithmic x-scale).

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



a. Random Forest

b. Gradient Boosting

Figure 5: Feature importance based on the training of the two regression models.

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.