Covid 19 : Study of factors explaining different case fatality rates among OECD countries

Larroque Baptiste

08 Mai 2020

On 1st April, many articles broadened the point of view that Artificial Intelligence (AI) had not been yet impactful against Covid-19 [1]. Indeed, its impact is hampered by a lack of data, too much noisy sometimes, and too many outliers generated by all social medias that deal with this topic and feed off this stressful and unusual worldwide situation. However, six areas where the AI can contribute to the fight against the virus have been pinpointed in the same article. The author explains that AI had been useful to develop prognostic predictions.

One interesting point for handling this sanitary crisis is to be able to determine countries which are at the higher risks for fatalities. Indeed, the data provided by Johns Hopkins University Center allow everybody to have a clear overall picture on the virus spread. It is blatantly obvious that the case fatality rate (CFR), which is the number of deaths on the number of confirmed cases, is not the same between countries which have been infected by the virus equally. This difference can be explained by many countries' features such as the quality of their health care system, or their demographic features. A Chinese study who took place in Wuhan's hospitals in January, achieved to identify the main comorbidities which can lead to severe complications putting the patient's life at stake [2]. The researchers used medical data from hospitals on people who had been put on life support. My first idea was to do the same with clinical data from French hospitals. However, I did not have access to this data, and I chose a broader scale. So, the idea is to implement an algorithm which can determine main factors that can explain a difference of mortality rates between countries. Thus, the WHO organization can foresee an outbreak which could end badly with many deaths. It turns out that this knowledge is capital for the WHO organization to be able to save lives as many as possible. For the data mining point of view, countries must be clustered and well classified according to their mortality rate.

To implement this algorithm, one need to discover patterns between countries regarding the evolution of their mortality rate, their recovery rate and their confirmed cases. After creating a file containing all needed data extracted from the John Hopkins' dataset, the first step will be to do a statistical analysis by plotting regressions. The scope will be focus on countries where the plague is already well settled and multiple data are available such as the OECD member countries. Then, a Principle Components Analysis (PCA) will be realized to cluster these countries regarding their mortality rate and their recovery rate. After discovering patterns and clusters, different classifications' methods will be built to portray the features of countries with a high mortality rate. These classifications will be analyzed with metrics and ROC curves.

Please notice that the morality rate and the death rate refer to the CFR as explained below. Also notice that the last day reported in my code is 17th April. My code and my proposed method have been built in such a way that they can be updated with new data.

1 Statistical Analysis

For this part, the aim is to have a quick overlook on the distribution of the mortality rate and the recovery rate between countries which are presented by medias to handle this sanitary crisis well and the other countries. This analysis will emphasize outliers and unveil different patterns between the two countries' groups. This part refers to the 'Statistical_Analysis.ipynb' file and more especially the library "seaborn" is used to obtain graphs.

With the help of the worldwide dashboard provided by John Hopkins University, it is obvious that main infected countries are from the northern part of the Earth. Indeed, the data unveils that more than 95% of the deaths come from the northern part of the hemisphere. Concerning the mortality rate and the recovery rate, the averages are respectively 0.03 and 0.10 if all daily reported rates are taken into account. However, these averages did not seem exactly accurate and advice researchers to be very aware and cautious. Indeed, the distribution evolves day after day on both rates.

On Figure 1, the box plot illustrates the distribution of the mortality rate on day 30 and day 86 after the first day[1] reported in the dataset. First, the average mortality rate is soaring up and there are more outliers which blur the real mortality rate of the virus at day 30. Few countries had even a mortality rate equal to one due to the lack of available tests at the beginning. As time goes by, countries equip themselves with necessary means to massively test the population and have a more precise idea. However interesting it might be, the death rate distribution is more widespread with less outliers. It inexorably entails that the death rate can differ from countries depending on politics put in place or countries' features. The same observation can be made for the recovery rate. Indeed, first patients generally recovered well. When more and more people are infected, this distribution shrank, and the average recovery rate fell due to the exposure of susceptible populations. These observations are illustrated by the evolution of both rates across the time in countries.

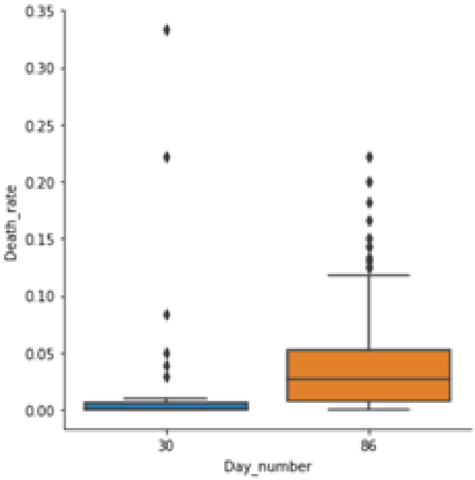

Figure 1: Boxplots of mortality rates

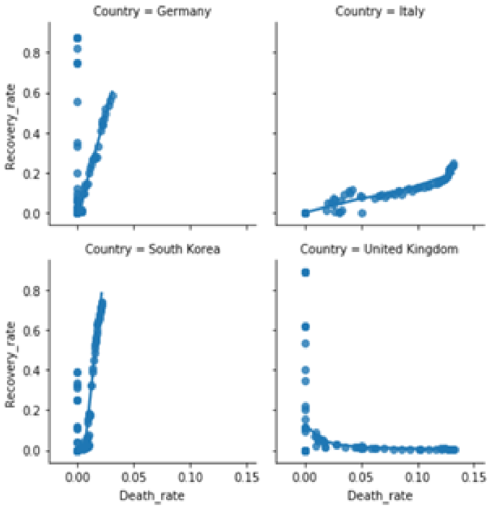

Figure 2: Regression between mortality and recovery rates

The curbs of the mortality rate and recovery rate seem to be similar across countries which are presented to handle well the crisis. On Figure 2, the correlation between the death rate and the recovery rate is indeed strong after a while despite outliers for particular cases. Before this moment, governments did not have a clear view on the progression of the epidemic. For countries which handle the crisis well, the stirring ratio is high. It means that the recovery rate is soaring up faster than the mortality rate which is bound to stagnate. Nevertheless, the others did not seem to have achieved this point. Indeed, the stirring ration is positive but not as high. It inexorably entails that the mortality rate soars up and the recovery rate barely skyrockets. For instance, in Great Britain, this coefficient is clearly below one which means that the death rate is growing faster than the recovery rate in the same time.

The statistical analysis unveils common patterns across countries regarding their mortality rate and recovery rate. Then, the idea is to cluster countries which have strong similarities regarding these rates and to understand the features that could explain theses clusters.

[1] 22nd January

2 Proposed method

Only European countries have been examined at a first glance, but they represent not enough data for the classification. Given that, all OECD member countries are added to the study. OECD member countries represent 33 samples instead of 14 for European countries.

Before getting into more details, a CSV file containing all daily data by country such as the recovery rate, the mortality rate, the number of confirmed cases, recovered cases and death cases is created. As already said, the last reported day was the 17$^{\text{th}}$ April. Nevertheless, the method will work and will adapt itself with an uploaded file with more recent data on Covid-19. Overall, the “pandas” library is used to make all SQL queries by applying the structure of Dataframe on data.

2.1 Clustering the data

This part will focus on clustering countries regarding their mortality rate and recovery rate because the statistics analysis unveils some common patterns between them. The library ”Sklearn” is mostly used in this part to manage a Principle Components Analysis. This part refers to ‘PCA_Analysis_.ipynb’ files.

2.1.1 Implementation of the Principle Components Analysis

Different methods can be used for clustering data. Instead of converting distance like the Multi-Dimension Scaling (MDS) algorithm does, the PCA algorithm converts correlations among samples into a two-dimension plot. The MDS needs information about the distance between variables which can be obtained by an Euclidian measure for example or by building a proximity matrix [3]. Principle components are linear combinations of tested variables, which are the mortality rate and the recovery rate in this case. Multiple principal components are calculated and the two components which are the more important when it comes to describe the spread out of the data are plotted on a two-dimension plot. Thus, clusters can be pinpointed by observing countries which are closer on the plot. As the former part highlights, governments obtain more and more visibility as time goes by. To assess this assumption, two two-dimension plots were processed. One had only the mortality rate and the recovery rate of the last day reported in the data by country as variables. The other one had all mortality rates and recovery rates by country from the day when all countries experienced at least one death due to Covid-19. This process is needed to avoid “Nan” values which are not compatible with this analysis.

The method will be assessed through “scree plots” and leading scores for each variable. A “scree plot” represents the percentage of variation that each principle component accounts for. The leading score for a principle component emphasizes what variables had the largest influence on separating the samples.

2.1.2 Results and interpretation

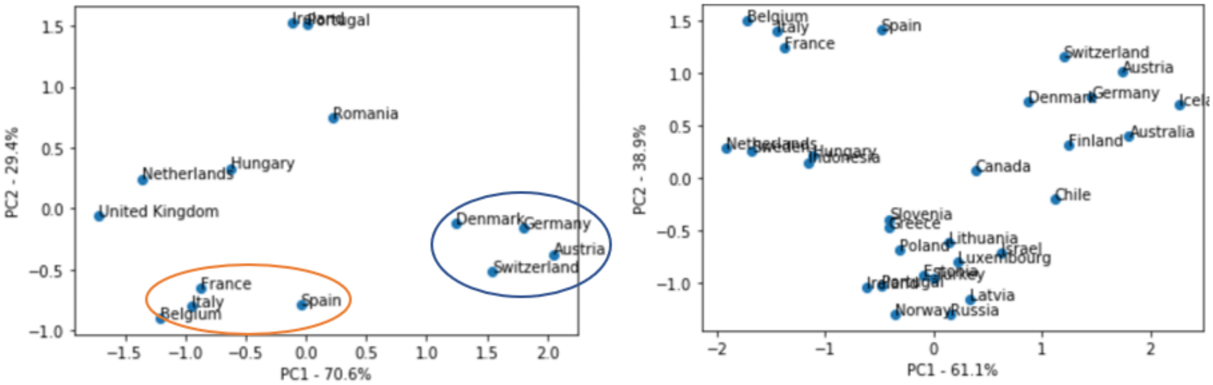

Figure 3: PCA graph with European and OECD member countries

On Figure 3, the left PCA graph represents the correlations between European countries regarding their last daily death rate and recovery rate. One cluster is clearly established on the right side with Germanic countries which apparently achieved to slow down the explosion of the mortality rate. Another group delimited by the red circle could be also described with countries which apparently struggled to maintain a low mortality rate. The spread out of countries is more important among the x-axis.

The PCA graph with OECD member countries is less representative than the one with European countries. Indeed, the identified clusters are also on this graph, but it is blurred by all the amount of data from other countries. One must notice that the contribution the y-axis and the x-axis is more equal in this case when it comes to describe the spread out of the data

2.2 Collect of data on countries' features

The next step is to collect data on countries' features to explain the clustering and differences in mortality rates between countries on the last reported day. Indeed, the assumption that as time goes by, this rate more accurately reflects the mortality of the virus, seems to be confirmed by both the statistical analysis and the PCA. OECD allows everybody to have access to countries' data through their website 'stats.oecd.org'. The proportion of overweight people and the proportion of people older than 65 years old are used to describe the tested countries' demographic and are expressed in percentage. These categories are often presented to be the more susceptible categories by doctors. Moreover, a country needs to have a good health care system before the crisis to be able to handle the amount of hospitalizations. That is why, the number of hospital beds per 1000 habitants and the number of healthcare workers per 1000 habitants are part of the analysis to assess the health care system. The death rate also depends on the visibility that each country has on the spread of the virus. To assess this spread, the number of tests per confirmed cases in the country[2] is added to the columns. This number illustrates the number of tests that each country can provide for one person who is infected [4]. The higher the number, the more the country is able to test the patient's entourage and to get a clear picture of the spread of the epidemic. A CSV file is created with the tested countries in index and with these features and the last reported daily death rate in columns. The death rate is also expressed in percentage with three significant digits.

2.3 Classification

This part will focus on the classification of the data. The aim is to be able to identify countries with a mortality rate higher than 8.0% among OECD member countries by regarding their features. For this, different classifiers are assessed with ROC curves and AUC scores. The best classifiers' performances will then be scrutinized with metrics such as the accuracy, the precision, or the recall. The library "Sklearn" is mostly used for the classification and this part refers to 'Classification_.ipynb' files.

2.3.1 Implementation of classifiers

31 OECD countries members are analysed. 7 countries are labeling "True" which means that their mortality rate is higher than 8.0% on the last reported day. The idea is to portray the main features of countries which mishandle this sanitary crisis. Four classifications' methods are tested: random forest, adaboost, gradientboosting, and decision tree. These are the steps for the implementation of classifiers:

- **Labeling** of each country which belongs of the Dataframe's index
- **Splitting** of the dataset into a training dataset and a test dataset which represents 25% of the data
- **Training** and **testing** of classifiers with the corresponding dataset
- **Evaluation** of best methods with metrics

Both decision tree classifier and random forest classifier are computed with specific arguments to best fit the data. For the decision tree, the split at each node is based on the Gini criterion instead of the entropy. For the random forest, 100 decision trees are made for deciding the label of each countries in the test dataset. Moreover, the maximum number of features to be used in trees of the forest is the root of the number of features to avoid overfitting. Concerning the adaboost and the gradientboosting, the parameters are set by

[2] This dataset is available on Github and is licensed under the Creative Commons BY license

default. It would be interesting to discover the effects of changing the learning rate on the accuracy of the two last classifiers.

ROC curves are plotted to compare classifiers through a cross validation process. Indeed, for each fold a ROC curve is plotted, and the result is the average ROC curve. It turns out that these models tend to overfit the dataset and that is why, comparing the methods on the classification of a single fold does not seem to be the appropriate way. Using the cross validation process is the silver bullet in this case for doing the comparison.

Regarding the AUC scores, the best methods are assessed though usual metrics. Particular attention will be paid to the precision[3] and the recall for categorizing countries at higher risks. One interesting to notice is that the dataset is imbalanced. So, it is easy to have a law false positive rate thanks to the large number of true negatives. Indeed, the number of countries having a mortality rate lower than 8.0% is fortunately lower than the number of countries labeled ”True”. In this case, the precision is more useful than the false positive rate because the precision does not take into account the number of true negatives and so is not infected by the imbalance. By the way, some test datasets do not include positive samples because of the randomly selection and the lack of positive samples in general, which leads to an undefined recall. In the same vein, some predictions made by the classifier do not include positive results which leads to an undefined precision. Therefore, the accuracy, the precision, and the recall are calculated by using a cross validation process. Undefined values are subtracted to calculate the average and the standard deviation of these metrics.

2.3.2 Results and interpretation

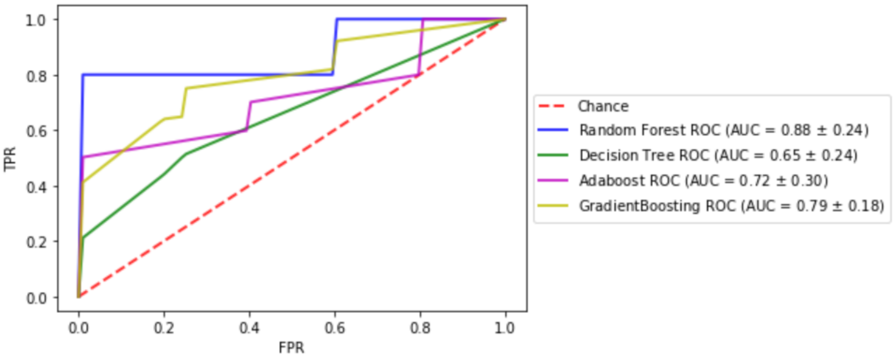

Figure 4: ROC curves and AUC scores for different classifiers

On Figure 4, the chance line means that the proportion of correctly classified samples is the same as the proportion of incorrectly classified samples. Besides, higher the AUC score, better the model is at distinguishing countries' mortality rates [5]. Thus, the random forest and the gradientboosting are the best classifiers. The AUC score is not only closer to one for the gradientboosting but it also more accurate as its lower standard variation hints at. Decision tree and adaboost can be removed from the analysis.

Classifier	Accuracy	Recall	Precision
Random Forest	0.90 ± 0.08	0.60 ± 0.37	1.00 ± 0.00
GradientBoosting	0.78 ± 0.15	0.60 ± 0.37	0.69 ± 0.32

Table 1: Results after the cross validation without regarding undefined values

[3] The precision is the number of positive results that were correctly classified. In term of conditional probability, it is the probability for the country to have a death rate higher that 8.0% knowing the fact that the country was classified in this category by the algorithm.

3 Results and Discussion

The Table 1 highlights that both differ by the precision and the accuracy. These metrics are better for the random forest especially with a precision equal to one. It means that this classifier is not mistaken when it classifies a country as a positive result. In other terms, if a country is classified as "True", the threat of a high mortality rate must be taken seriously. Even more so that the accuracy with the random forest is pretty good. However, both classifiers have an uncertain recall as the standard variations suggest. In other terms, both approximatively recognize a positive sample, but when it does, it is very sure than this country will strangle to deal with this sanitary issue. This bald result is particularly due to the overfitting as the perfect results for the classification of the training dataset unveil. So, the random forest seems to be better for the classification.

Concerning the features, both methods mainly use the number of tests per confirmed cases to classify. Indeed, an incorrect and bias view on the spread of the pandemic will automatically involve an overestimated mortality rate due to an underestimated confirmed cases' number. The contribution of this feature is higher with gradientboosting. It means that other features need to be considered to find the appropriate classification. Demographic features have also a greater contribution than health care system's features.

The main issue with the classification is the recall's uncertainty due to overfitting. The random forest is improved by testing randomly hyperparameters through a cross validation process in the 'Improvement_Random_Forest.ipynb' file based on this article [6]. By adjusting the number of trees and other parameters, the recall is improved by 10 points of percentage, but it is still fickle. The average recall reaches 70% which means that the classifier can recognize countries which will strangle to deal with Covid-19. However, the recall's improvement comes at the expense of the precision which still remains high (around 90%).

Nevertheless, one might be aware that the variations in mortality rates would be primary due to the reliability of data, especially at the beginning of the epidemic when it is practically impossible to establish an exhaustive count of the infected cases. "In any pandemic, there is a tendency to initially overestimate the case fatality rate. Then, as more people are infected and the management of severe cases improved, the fatality rate declined" [4]. So, countries do not have the same criterion to count infected people. Even if the number of tests per confirmed cases gives an idea about the capacity of country to have a clear view on the spread out, the main explanation of mortality rate's variations, namely the reliability of data, is not reflected in tested features. It will reluctantly involve misrecognitions by the classifier.

There are other limitations in this classification model. Indeed, the provided data do not convey information about different strategies put in place by countries to deal with this sanitary issue. For instance, Germany decides to put emphasis on testing the confirmed case's entourage and then quickly isolates these people. Other countries such as Portugal and Greece anticipate the spread of the pandemic by putting more money in research to have enough tests. South Korea chooses to track the patients with geolocation systems and Great Britain seems to follow this path. The strategy of "Isolation, testing and tracing" has been presented to be the silver bullet. It would also be interesting to compare and to discover patterns among countries which applied the same strategy.

The ultimate aim of this work is to be able to apply the classifier and to know if people's life in a specific country will be at higher risks regarding its main features. After observing that countries were unequally affected by the virus through the statistical analysis, the PCA enabled to cluster these countries. Different classification's methods were compared and the random forest was improved for classifying countries at higher risks regarding demographics features, their health care system, and their knowledge about the spread of the epidemic. The random forest did a good job to recognize positive samples among OECD member countries, but as the fickle recall hints at, the robustness and the scalability of this model are questionable. Overall, the contributions of the features for classifying unveil that all governments should massively invest in testing. Then, the protection of vulnerable people and the improvement of the health care system should be considered.

References

[1] Wim Naude. “Artificial Intelligence against COVID-19: An Early Review”. In: *Towards Data Science* (2020).

[2] Xiangao Jiang, Megan Coffee, Anasse Bari, et al. “Towards an Artificial Intelligence Framework for Data-Driven Prediction of Coronavirus Clinical Severity”. In: *CMC-Computers, Materials Continua* 63 (2020), pp. 537–551. DOI: `doi:10.32604/cmc.2020.010691`.

[3] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. 3rd ed. Springer, 2008. Chap. 14, pp. 501–534.

[4] Hannah Ritchie and Max Roser. “What do we know about the risk of dying from COVID-19 ?” In: *Our World Data* (2020).

[5] Jiaway Han, Micheline Camber, and Jan Pei. *Data Mining Concepts and Techniques*. 3rd ed. Morgan Kaufman, 2012. Chap. 8, pp. 327–385.

[6] Will Koerhsen. “Hyperparameter Tuning the Random Forest in Python”. In: *Towards Data Science* (2018).