Identifying the Most Relevant Attributes to Explain Peaks of COVID-19 Infections and Deaths by Machine Learning Methods

Gabriel Pena, Juliana Gambini, and Nestor R. Barraza*

***Abstract*—One of the key factors related to assessing the spreading speed of a given disease is to determine the peak of infections, the point after which a wave starts to mitigate as the daily number of cases goes down. This issue has attracted the attention of scientists for the last two years in relation to the COVID-19 pandemic. At the present time, since several waves have affected most countries, there is plenty of information at our disposal: date and magnitude of contagion peaks; country-related data such as population density, gdp per capita, etc.; among other relevant status metrics at the dates of peaks, like vaccination, mobility, use of mask, occupied hospital beds, etc. Thus, finding which of those attributes are relevant and ranking them becomes an interesting field for research. In this work, we apply a filtering technique to identify peaks on the reported data and then perform feature selection algorithms with the peak magnitude as output. A comparative ranking of the attributes is thus obtained for several countries and for different waves in the same country. As pre-processing tasks, we performed a normalization and a conversion from numerical to categorical values on the output variable. As a result, a grouping of countries and waves is obtained, from where important information can be extracted. Our results contribute with knowledge for predicting and monitoring the spreading of diseases and become a relevant tool for health institutions and authorities.**

***Index Terms*—COVID-19, peak of infections, feature selection, clustering, machine learning, K-means, random forest, Boruta**

I. Introduction

In the recent years, many mathematical models have been proposed to monitor and predict the evolution of the COVID-19 outbreak (see for example [1] and references therein). One large class of those are the Machine Learning prediction models. In [2,3], surveys of Machine Learning techniques applied to COVID-19 pandemic are introduced. In [4] the author analyzed several Machine Learning methods that allow to predict risk factors such as age, social habits, location, and climate.

Among Machine Learning tools, feature selection algorithms are important in order to select the most relevant characteristics related to the prediction of a given variable, either to understand a given domain or to simplify the prediction procedure by allowing to work with less attributes (see for example [5, 6] and references therein). Random Forest [7] and Boruta [8] algorithms are examples of these. In epidemics, a feature selection process can be applied in order to determine which characteristics contribute the most to the virulence of a given disease. The aim of this work is to apply feature selection algorithms to find which country attributes are the most relevant in order to determine the magnitude of COVID-19 peaks of infections and deaths. Our proposal is to apply a classical feature selection procedure where the output variable is the magnitude of the peak of infections on the one hand, and the magnitude of the peak of deaths on the other. As input attributes, we take several characteristics of the analyzed countries, in order to detect which of those are most relevant to determine the magnitude of those peaks. It is clear that knowing the most important factors that affect the impact of the disease can be helpful to control outbreaks. We have an available dataset where the list of attributes involving both static variables such as life expectancy, median age, quantity of smokers, etc., and dynamic variables like mask use, vaccination, and mobility. Those dynamic variables are taken at the date of each identified peak. In this analysis, we have taken advantage of the fact that every country has gone through several waves and, consequently, there are various peaks of both infections and deaths in every country. We have analyzed the attributes of a total of 129 countries, though after adding several waves for each country (between 3 and 5) the dataset is composed by 423 records. Table I shows the complete list of input attributes with their descriptions.

We then chose an output variable, the peak of infections on the one hand and the peak of deaths on the other, and analyzed which attributes determine those outputs best. Since the positive rate was far away from the optimum values recommended by the World Health Organization (less than 5%), it is reasonable to consider

Manuscript received September 7, 2022; revised November 1, 2022; accepted December 29, 2022.

G. Pena is with the Technology and Science Department of the Universidad Nacional de Tres de Febrero, Argentina.

J. Gambini is with the Technological Institute of Buenos Aires (ITBA) and the Technology and Science Department of the Universidad Nacional de Tres de Febrero, Argentina.

N. R. Barraza is with the Technology and Science Department of the Universidad Nacional de Tres de Febrero and the School of Engineering of the University of Buenos Aires, Argentina.

*Correspondence: nbarraza@untref.edu.ar

DOI: 10.7763/IJCTE.2023.V15.1326

TABLE I
FULL LIST OF INPUT ATTRIBUTES

Attribute name	Description	Experiment
total_cases_norm	Total cases	Deaths
total_deaths_norm	Total deaths	Cases
icu_patients_norm	ICU patients	Deaths
hosp_patients_norm	Hospital patients	Both
total_tests_norm	Cumulative nº of tests	Both
new_tests_norm	Daily nº of tests	Both
mask_use	% of mask usage	Both
mobility	% of social mobility	Both
vaccination_1_dose	Nº of first doses applied	Both
vaccination_full	Nº of fully vacc. people	Both
reproduction_rate	Effective reproduction rate (R)	Both
positive_rate	COVID-19 tests that are positive, given as a rolling 7-day average (this is the inverse of tests_per_case)	Both
tests_per_case	Tests conducted per new confirmed case of COVID-19, given as a rolling 7-day average (this is the inverse of positive_rate)	Both
stringency_index	Government Response Stringency Index: composite measure based on 9 response indicators including school closures, workplace closures, and travel bans, rescaled to a value from 0 to 100 (100 = strictest response)	Both
population_density	Population density	Both
median_age	Median age	Both
aged_65_older	Share of the population that is 65 years and older, most recent year available	Both
aged_70_older	Share of the population that is 70 years and older in 2015	Both
gdp_per_capita	Gross domestic product at purchasing power parity (constant 2011 international dollars), most recent year available	Both
extreme_poverty	Share of the population living in extreme poverty, most recent year available since 2010	Both
cardiovasc_death_rate	Death rate from cardiovascular disease in 2017 (annual number of deaths per 100,000 people)	Both
diabetes_prevalence	Diabetes prevalence (% of population aged 20 to 79) in 2017	Both
female_smokers	Share of women who smoke, most recent year available	Both
male_smokers	Share of men who smoke, most recent year available	Both
handwashing_facilities	Share of the population with basic handwashing facilities on premises, most recent year available	Both
life_expectancy	Life expectancy at birth in 2019	Both
human_development_index	A composite index measuring average achievement in three basic dimensions of human development—a long and healthy life, knowledge and a decent standard of living. Values of 2019, imported from [9]	Both

that the reported data of COVID-19 cases (and, in a lower magnitude, deaths) are not exact, and are subject to additive noise. This assertion is supported by the inspection of the reported cases curves, where a sharp variation (characteristic of a high frequency additive noise) is clearly perceived, and more accurately, by frequency

domain analysis (see Fig. 2). In order to determine the location of peaks in each curve, a smoothing filtering process must be applied. Despite quite complex filtering techniques can be considered, since the spectrum shows mainly an additive noise component at the frequency of 1/7 a day; a straightforward low-pass filter is enough to smooth the curves in order to clearly identify the peaks. Another necessary task was to change the output variable from numerical to categorical, to improve the performance of the algorithms on the one hand, and to identify groups of countries and waves on the other. In order to achieve this, a clustering procedure was performed prior to the feature selection. The number of clusters was selected according to the Freedman-Diaconis criterion [10], and two grouping algorithms were considered: Jenks-Fisher [11] and K-Means [12]. This arrangement of countries and waves shows an interesting assembly of sets, where important information can be extracted from and is itself a relevant data analysis. Silhouette [13], Calinski-Harabasz [14] and Davies-Bouldin [15] coefficients were obtained to measure the efficiency of the clustering procedure. Fig. 1 depicts a graphical summary of the whole process 1.

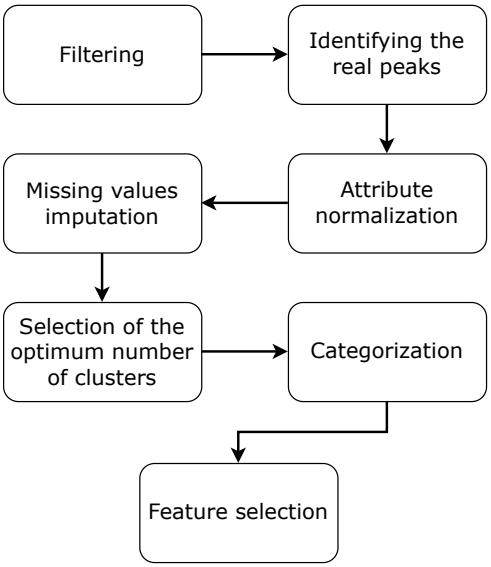


Fig. 1. Flowchart showing the complete feature selection procedure, including all preprocessing steps.

The results we obtained are aligned with those established by the scientific community for identifying the factors that contribute to the disease virulence. However, our work is useful since a measure of their relative importance and order is introduced. As an example of application of Machine Learning tools, this work is also of interest for the data analysis community.

This work is organized as follows. In Section II the data preprocessing, including the signal filtering, peak detection and data preparation, is described. In Section III we explain the feature selection process and show the main results. In Section IV, a discussion is presented. Finally, in Section V, some conclusions are drawn.

II. Data Preprocessing

A. Filtering and Peak Detection

Since the number of tests do not cover the total population, and the sampling is not performed uniformly over time, we can consider that the daily data curves are highly contaminated with noise, which must be suppressed in order to properly detect peaks. By observing the data spectrum (Fig. 2 shows the USA's; which exhibits the same shape in almost every country), significant components can be clearly distinguished at frequency $1/7$ day^{-1} and its harmonics. This is explained by the weekly periodic variations in the number of tests (as it is known, there are fewer COVID tests performed on weekends). An analogous behaviour can be seen in the deaths dataset, since a smaller number of deaths is reported during weekends. Since we are looking for a smooth curve of data, we just need to keep the low frequency components; hence, we preferred low-pass filtering instead of band-suppress or notch. Among these, we used moving-average (MA) filters, which are the simplest low-pass finite impulse response (FIR) filters (see [16]).

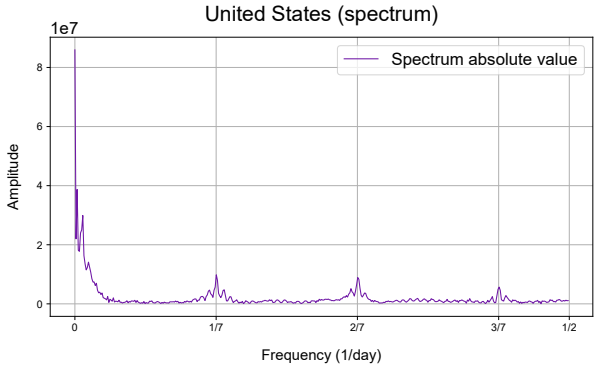
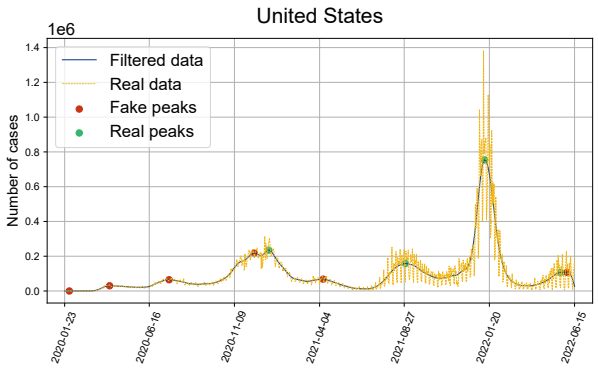


Fig. 2. USA data displayed in the frequency domain (spectrum). Vertical axis shows the amplitude of the Fourier transform, computed via the FFT algorithm. Since the data is reported in cases per days, the frequency axis has units of day^{-1}. Significant peaks can be appreciated at the harmonics of $1/7$ day^{-1}, as expected.

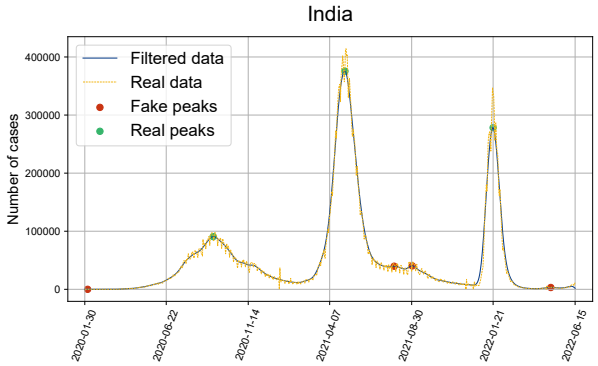
Since we need at least to remove the $1/7$ frequency component, L (the length of the filter kernel) must be chosen to be $L \geq 7$. In order to avoid harming the signal too much, we chose to fix $L = 7$ and apply the same filter n times in a sequence. As stated in [16], this kind of arrangement approximates a Gaussian filter, which is smoother than a single MA, and hence less likely to harm the signal.

Once we obtain a smooth curve, maximums and minimums detection can be done by straightforward methods. Hyper-parameters must be carefully chosen by the user (typically, values of $L = 7$ and $n \geq 7$ achieve good enough results, except in some pathological cases). Since

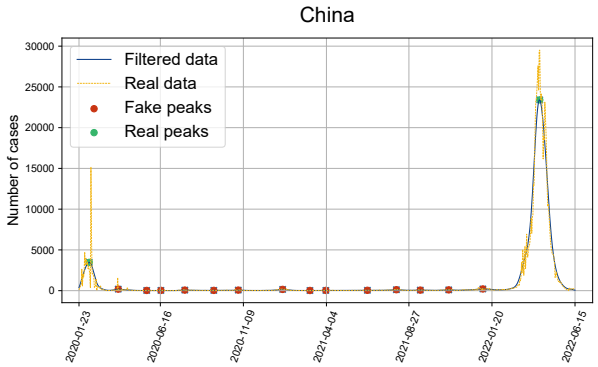
filtering is not perfect, user interpretation is necessary to decide which peaks are "false" (product of persistent noise) and which are the real ones. The result is shown in Fig. 3 for the United States (Fig. 3a), India (Fig. 3b) and China (Fig. 3c), where the fake peaks are shown with red dots and the real peaks with green dots. After deleting the fake peaks from the datasets, we kept a total of 422 peaks of cases and 464 peaks of deaths.



(a) Filtered and real data with fake and real peaks (USA).



(b) Filtered and real data with fake and real peaks (India).



(c) Filtered and real data with fake and real peaks (China).

Fig. 3. Filtered and real data curves and detected peaks taken from USA, India and China data.

B. Dataset Preparation

The next step in the process is the preparation of a full dataset in an adequate format. We obtained both the daily cases and daily deaths reports and the country attributes from Our World In Data [17], with a few exceptions: the social mobility, mask use, and vaccination data, which we took from the Institute of Health Metrics and Evaluation [18]. The data covers from January 2020, the beginning of the epidemic, to June 15, 2022. From these data we produced two different datasets, one to be used with the peaks of cases as the output variable, and the other to be used with the peaks of deaths as output. Then, some attributes needed to be normalized to take the population into account. Thus we divided their values by their country's population. The complete list of attributes taken for each experiment is shown in Tables V, VI (feature selection on the peak of cases), Tables VII and VIII (feature selection on the peak of deaths). Note in Tables V and VII that the input attributes are slightly different for each experiment: in Table V the peak of cases is taken as the output variable, and the total number of deaths appears as an attribute, whereas in Table VII the output variable is the peak of deaths and the total number of reported cases is an input variable. Also, ICU occupation was not considered as an input for the peak of cases, though it is certainly important when explaining the deaths, which is why it appears only in Table VII. Variables such as these, whose names end with "norm", take normalized values relative to the population. Finally, it was necessary to fill the blanks in the dataset (a 18.14% of the data were missing). Vaccination columns had a lot of blanks because some peaks happened before the vaccines were available. We filled these column with zeros. For the rest of the variables, the blank spaces were filled according to a register proximity rule by using the kNN imputation method [19].

C. Converting the Output Variable from Numerical to Categorical

Since exact numerical values are not of particular importance, but macro tendencies are, we aimed for a classification instead of a regression. Thus, the output variables (relative peak of cases in one experiment, relative peak of deaths in the other one) needed to be categorized into discrete classes.

Discretization can be done in several ways. We first chose an "optimal" number of clusters, or bins, according to some criterion. We considered three different estimates for the number of bins: Freedman-Diaconis [10], Sturges [20] and Scott [21], three well-known rules in the theory of histograms. Overall, Freedman-Diaconis performed better, so we discarded the rest. It must be remarked that these three methods are generally used to choose an optimal number of bins in a histogram, where each bin has an uniform length. However, for the purpose of discretization, this estimation can be greatly improved if one chooses non-uniform bins. Therefore, we just used the rules to determine the number and then assigned the data by other methods.

Once the number of bins is chosen, the next step is to place every peak into a class, which can be done by any grouping or clustering algorithm. We tested four methods: uniform binning ("base" Freedman-Diaconis estimate), agglomerative (hierarchical) clustering [22], [23], K-Means clustering [12] and Jenks-Fisher rule [11]. Clustering performance was evaluated by the Silhouette [13], Calinski-Harabasz [14] and Davies-Bouldin [15] metrics; results are shown in Tables II and III.

TABLE II
METRICS OF CLUSTERING PERFORMANCE FOR THE PEAKS OF CASES. NUMBER OF CLUSTERS: 39

Algorithm	Silhouette	C-H	D-B
Uniform binning	0.5738	5525	0.3295
Agglomerative clustering	0.5534	29345	0.3830
K-Means	0.5626	30771	0.3909
Jenks-Fisher	0.5779	33037	0.3912

TABLE III
METRICS OF CLUSTERING PERFORMANCE FOR THE PEAKS OF DEATHS. NUMBER OF CLUSTERS: 22

Algorithm	Silhouette	C-H	D-B
Uniform binning	0.5872	4231	0.4119
Agglomerative clustering	0.5708	10605	0.4144
K-Means	0.5744	11158	0.4087
Jenks-Fisher	0.5821	11640	0.4154

Overall, Jenks-Fisher and K-Means methods are the most consistent, so we chose these two sets of clusters as our categorical outputs for the feature selection experiments. Interesting information can be obtained by analyzing the output of the clustering procedure. For example, taking into account the normalization we performed, different waves from different countries are grouped together, the first peak of the United States is grouped together with peaks other than No. 1 of different countries. Clusters are of quite different sizes. Despite India had attracted the attention due to the number of infected people and deaths, it belongs to the group of not so noticeable countries and waves. All datasets are available for further use [24].

III. FEATURE SELECTION

We performed four experiments on both infected and deaths datasets, all of them based on Random Forest classifiers. Two standard Random Forest classifications were applied to the two chosen output variables (Jenks-Fisher and K-Means sets of clusters). Then, two experiments were made applying the Boruta algorithm to a base Random Forest estimator to obtain a first estimate of the attribute importance; then, attributes marked as non important were discarded and another Random Forest was trained with only the important variables. The Boruta experiment was also performed on the two output columns. Finally, the attribute importance in each experiment was estimated by the mean decrease in impurity (MDI) within all the trees in the forest. We also considered using the feature permutation method to estimate importance. However, since the results were quite similar, they are not shown here.

All the processing described in this work (preprocessing and feature selection) was done in the Python language; the source code is available at [25] for public use. Feature selection methods were performed using the algorithm implementations present in Scikit-Learn [26] and BorutaPy [27] libraries. The hyper-parameters were fixed in all the experiments as it is indicated in Table IV. Performance of each experiment (dataset + algorithm + output variable) was evaluated by four metrics: accuracy on the training set, accuracy on the testing set, out-of-bag error and mean cross-validation score. These experiments, applied to both the infections and deaths datasets, gave us eight sets of results, which are summarized on Tables V, VI, VII and VIII. The accuracy metrics listed on these tables show that the model exhibits overfitting (over 90% accuracy on the training set versus around $20-30\%$ accuracy on the test set). In our experiments, we found that different choices of hyperparameters produced a decrease in accuracy on the training set, while that of the test set remained constant, and without relevant changes in the order of importance of attributes. Since we are mainly concerned on fitting the present data instead of predicting new ones, in order to determine the best features for the actual data, we set the hyperparameters so as to obtain the best fitting on the training set. The choice of a small percentage as test set, was intended in order to quantify how such a predictive model could perform. Algorithm 1 summarizes the whole process.

Algorithm 1 Step by step description of our methodology.

1-Filtering
2-Actual peak detecting
3-Data preparation
4-Selection of the optimum number of clusters
5-Categorization
6-Feature Selection

TABLE IV
FEATURE SELECTION HYPER-PARAMETERS

Hyper-parameter	Value
Size of training set	90%
Number of trees	30
Impurity function	Shannon's entropy
Max. bootstrap samples per tree	80%
Max. distinct attributes per tree	70%
Min. Nº of samples to divide a node	5
Min. Nº of samples to mark a leaf	3
Boruta threshold	70% percentile

TABLE V
RESULT OF APPLYING RANDOM FOREST ALGORITHM FOR FEATURE SELECTION ON THE PEAK OF CASES. THE COLOURING IS A VISUAL INDICATOR OF THE IMPORTANCE; GREEN MEANS MORE IMPORTANT, RED MEANS LESS IMPORTANT

Variable	K-Means	Jenks-Fisher
hosp_patients_norm	2.35%	2.57%
total_tests_norm	8.88%	8.96%
new_tests_norm	7.81%	7.29%
total_deaths_norm	**23.60%**	**23.18%**
mask_use	2.92%	3.95%
mobility	2.57%	2.39%
vaccination_1_dose	2.26%	2.16%
vaccination_full	2.00%	1.73%
reproduction_rate	4.54%	4.89%
positive_rate	2.74%	2.15%
tests_per_case	3.25%	2.62%
stringency_index	2.86%	2.97%
population_density	2.02%	1.73%
median_age	7.61%	7.57%
aged_65_older	2.22%	2.12%
aged_70_older	1.79%	1.93%
gdp_per_capita	1.02%	1.02%
extreme_poverty	1.48%	2.04%
cardiovasc_death_rate	2.67%	3.10%
diabetes_prevalence	2.41%	2.52%
female_smokers	2.88%	3.43%
male_smokers	1.51%	2.32%
handwashing_facilities	1.14%	1.67%
life_expectancy	5.71%	3.99%
human_development_index	1.76%	1.71%
Performance metric		
Accuracy on the training set	93.14%	94.20%
Accuracy on the testing set	37.21%	20.93%
Out-of-bag error	0.2375	0.277
Mean cross-validation score	0.2321	0.2876

TABLE VI
RESULT OF APPLYING BORUTA ALGORITHM FOR FEATURE SELECTION ON THE PEAK OF CASES. THE COLOURING IS A VISUAL INDICATOR OF THE IMPORTANCE; GREEN MEANS MORE IMPORTANT, RED MEANS LESS I

Variable	K-Means	Jenks-Fisher
total_tests_norm	11.02%	9.72%
new_tests_norm	9.29%	9.56%
total_deaths_norm	**24.11%**	**26.16%**
mask_use	4.79%	6.26%
mobility	4.64%	-
vaccination_1_dose	5.55%	5.88%
stringency_index	5.84%	6.05%
median_age	12.57%	9.79%
aged_65_older	4.87%	-
cardiovasc_death_rate	-	5.31%
female_smokers	5.25%	5.89%
life_expectancy	6.26%	7.60%
Performance metric		
Accuracy on the training set	92.08%	92.35%
Accuracy on the testing set	30.23%	27.91%
Out-of-bag error	0.2375	0.285
Mean cross-validation score	0.2374	0.2717

TABLE VII
RESULT OF APPLYING RANDOM FOREST ALGORITHM FOR FEATURE SELECTION ON THE PEAK OF DEATHS. THE COLOURING IS A VISUAL INDICATOR OF THE IMPORTANCE; GREEN MEANS MORE IMPORTANT, RED MEANS LESS IMPORTANT

Variable	K-Means	Jenks-Fisher
hosp_patients_norm	4.58%	3.90%
icu_patients_norm	3.69%	7.02%
total_tests_norm	2.55%	2.32%
new_tests_norm	2.08%	1.78%
total_cases_norm	**12.39%**	**12.17%**
mask_use	2.94%	2.47%
mobility	3.31%	3.76%
vaccination_1_dose	1.81%	1.72%
vaccination_full	2.85%	1.54%
reproduction_rate	3.43%	3.20%
positive_rate	3.50%	3.45%
tests_per_case	3.79%	3.77%
stringency_index	3.15%	2.30%
population_density	3.58%	3.37%
median_age	3.65%	3.22%
aged_65_older	5.54%	7.14%
aged_70_older	6.82%	4.44%
gdp_per_capita	3.47%	4.28%
extreme_poverty	2.70%	2.75%
cardiovasc_death_rate	2.24%	2.71%
diabetes_prevalence	3.06%	4.12%
female_smokers	5.92%	5.18%
male_smokers	3.15%	2.42%
handwashing_facilities	2.66%	2.95%
life_expectancy	2.45%	2.86%
human_development_index	4.69%	5.18%
Performance metric		
Accuracy on the training set	94.96%	93.53%
Accuracy on the testing set	42.55%	31.91%
Out-of-bag error	0.3094	0.2686
Mean cross-validation score	0.3286	0.3068

TABLE VIII
RESULT OF APPLYING BORUTA ALGORITHM FOR FEATURE SELECTION ON THE PEAK OF DEATHS. THE COLOURING IS A VISUAL INDICATOR OF THE IMPORTANCE; GREEN MEANS MORE IMPORTANT, RED MEANS LESS IMPORTANT

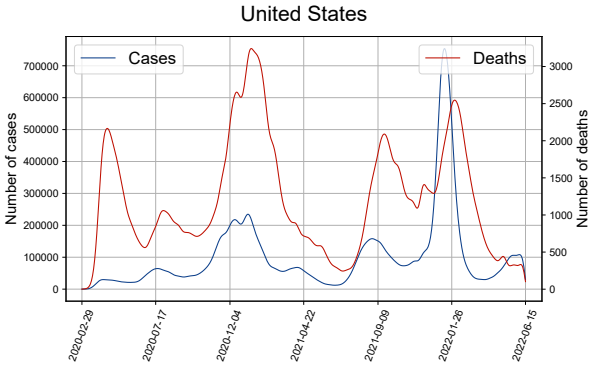
Variable	K-Means	Jenks-Fisher
hosp_patients_norm	6.46%	7.53%
icu_patients_norm	5.17%	6.98%
total_cases_norm	**14.99%**	**14.32%**
mobility	5.93%	7.38%
vaccination_full	5.11%	-
positive_rate	4.27%	4.95%
tests_per_case	6.46%	5.70%
population_density	5.42%	5.99%
median_age	4.45%	6.73%
aged_65_older	8.21%	7.75%
aged_70_older	5.65%	7.75%
gdp_per_capita	4.22%	5.37%
extreme_poverty	5.22%	5.92%
female_smokers	8.61%	7.70%
life_expectancy	5.58%	-
human_development_index	4.24%	5.94%
Performance metric		
Accuracy on the training set	92.33%	88.49%
Accuracy on the testing set	29.79%	27.66%
Out-of-bag error	0.307	0.2878
Mean cross-validation score	0.2973	0.295

IV. Discussion

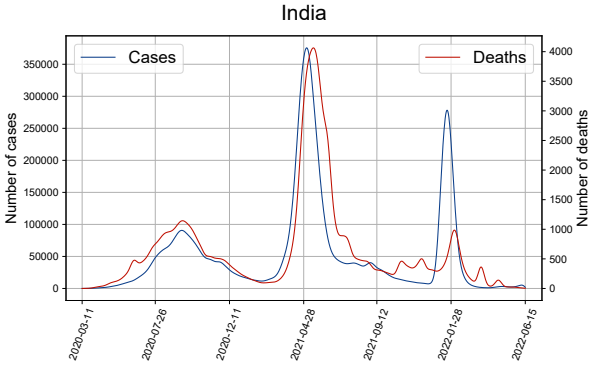
We have presented a workflow to find the most influencing attributes in order to explain the peak magnitude of infectious cases and deaths by COVID-19 disease. We can highlight that our results show that, by using any of the mentioned estimation methods, the variable describing the total deaths is by far the most influential one. This shows the existence of a strong correlation and justifies using the amount of deaths to predict the number of infections, as some models propose. This is obviously the result of a cause-effect relation: more cases imply more deaths. Fig. 4 shows the (filtered) daily cases and deaths curves of our three selected countries (United States, India and China). It is clear that both curves follow the same pattern, with the deaths curve delayed by around 14 days (the infectious period). This can be a useful tool to produce good estimates of the "real" number of cases when the testing procedures are weak or non existent, but the number of deaths is known. Correlation metrics are shown in Table IX; the correlation coefficients values along with the extremely small p-values, prove that this correlation, though not extremely strong, is statistically significant with a very high probability, as we expected. The other important variables, the second, third and fourth are not unanimous. The most voted variables are those related to median age, new tests and total amount of tests. Median age contributes more to the model than age groups over 65 and over 70. An interesting and quite surprising result is that the incidence of female smokers is superior to that of male smokers, which can be seen not only by their importance but also in the Boruta algorithm choosing to keep it (see Tables VI and VIII). Life expectancy is another attribute that contributes over most of them. To sum up, we conclude that static variables related to median age and life expectancy are more relevant than hand-washing facilities, mask use or mobility, which are generally used to control the disease. It can be observed that the variables median age and life expectancy have strong influence in the peak of cases, which in terms of public policies, shows the importance of protecting the senior population.

On the other hand, the variables age 65 older and female smokers have greater influence in the peak of death. This is in coincidence with the article [28], where the author shows that there are changes in the contribution of smoking to the sex differences in life expectancy in Europe, along 1950-2014, and that the life expectancy of female smokers is greater than male smokers. Then, the importance of this variable is related to the life expectancy.

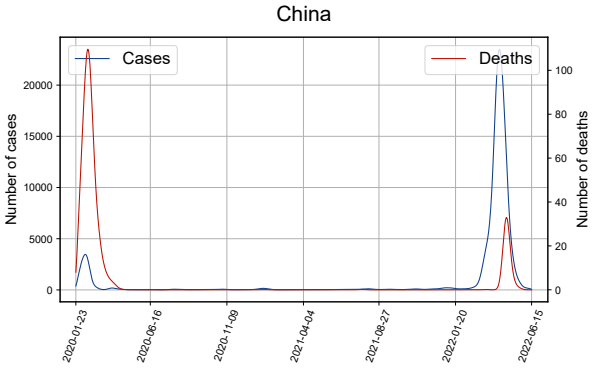
A remark on the contribution of the mobility factor must be pointed out: the effect of lockdown actions appears several days afterwards. These actions are difficult to be taken on time and were generally taken too late, like in European countries during the first wave, or too early, like the case of Argentina, where strong restrictions were decided when a small number of cases were being reported. As a consequence, a high and sharp peak is obtained in the first case as well as a low value in mobility, and a low and wide peak with a higher mobility



(a) Filtered daily cases and deaths curves (USA).



(b) filtered daily cases and deaths curves (India).



(c) Filtered daily cases and deaths curves (China).

Fig. 4. Filtered daily cases and deaths data from USA, India and China. The scale of the cases curves is shown on the left side axis, whereas the scale of the deaths curves is shown on the right side axis.

TABLE IX
Peak of Cases vs Total Deaths, Correlation Metrics

Metric	Coefficient	P-value
Pearson r	0.4632	7.88×10^{-24}
Spearman ρ	0.8089	6.26×10^{-99}
Kendall τ	0.6152	1.74×10^{-79}

value is obtained in the second. Consequently, at the date when the peak occurs, the mobility value does not correlate with the peak magnitude. The mobility should then be correlated to the slope of the curves, instead of a particular value at a given date.

The death cases variable shows similar results, though with more importance of the attributes related to aged population, as expected, and a low contribution of median age. In this case, female smokers are quite more important than male smokers, and contributes even more to deaths than to infections. The contribution of handwashing facilities, mask use and mobility are similar to that obtained in the dataset of infectious cases.

It is noticeable that all the important variables are extremely related to cause and effect, for example, if the population has a high rate of people older than 70, then it corresponds to a population with high life expectancy. In the same sense, if the total number of deaths from a disease is very large, there must have been many infected people. And therefore, the results are consistent.

Machine Learning methods provide us, not only with predictive model development, but also essential knowledge that defines the relationships between variables, which is the first step in very important tasks such as dimensional reduction. This specific knowledge can be used in a wide variety of applications as public policies design or establish strategic priorities.

V. Conclusions

A feature selection analysis on COVID-19 datasets was performed. We identified the most relevant attributes that determine the peak magnitude of both infections and deaths on several countries. From our procedure, we demonstrated that life expectancy and median age have more influence than other variables generally used to control the outbreak. A filtering procedure in order to smooth the reported curves and easily detect peaks and waves was introduced. Clustering algorithms were also performed in order to group countries and waves of infections and deaths, which also showed important information. This work contributes to identify the main factors that determine the COVID-19 disease virulence. A metric to measure the relative importance of the attributes was obtained.

As future work, we observed that the data sets related to COVID-19 contain information that should be further investigated, with the aim of designing better public policies, then we want to direct our efforts to research that point. In addition, we want to measure the impact of vaccination in the waves of the disease [29]. The two statistical methods popularly used prior to modern Machine Learning methods specified in this research would be:

I. Principal Components (by Hottelling's principal component technique).
II. Factor Analysis (by Galton and Pearson/Spearman) [30, 31].

One would expect the principal component to be the total_deaths_norm or the total_cases_norm, as evident from tabulated findings unanimously in this article. Last but not the least, the authors also anticipate an alternative non-analytical future method with which to conduct (as related to Fig. 1 and Fig. 2, and Table I to Table IX) certain selected Discrete Event Simulation (DES) analyses [32], where "the number of COVID-19 cases" could be the random variable of pivotal interest in order to compare and contrast with the proposed analytical findings to ensure flexibility and versatility.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

GP proposed, developed and implemented the filtering, clustering and Machine Learning algorithms and methods, contributed also to the research and data analysis; JG contributed to the discussions and data analysis; NRB introduced the main idea and conducted the research; all authors contributed to the draft and final versions writing and have approved the final version.

Funding

This work was supported by Universidad Nacional de Tres de Febrero, Pcia. de Buenos Aires, Argentina, under grants 32/19 80120190100010TF and 32/19 80120190100039TF, and PICT-2018-04485 from the Agencia Nacional de Promoción Científica y Tecnológica, Argentina.

References

[1] N. R. Barraza, G. Pena, and V. Moreno, "A non-homogeneous Markov early epidemic growth dynamics model. Application to the SARS-CoV-2 pandemic," Chaos, Solitons & Fractals, vol. 139, p. 110297, 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0960077920306937

[2] N. El-Rashidy, S. Abdelrazik, T. Abuhmed, E. Amer, F. Ali, J. Hu, and S. El-Sappagh, "Comprehensive Survey of Using Machine Learning in the COVID-19 Pandemic," Diagnostics, vol. 11, no. 7, 2021. [Online]. Available: https://www.mdpi.com/2075-4418/11/7/1155

[3] M. Rahman, K. Paul, M. Hossain, G. Ali, M. Rahman, and J. Thill, "Machine learning on the covid-19 pandemic, human mobility and air quality: A review," IEEE Access, vol. 11, no. 9, pp. 72 420–72 450, 2021.

[4] S. Kushwaha, S. Bahl, A. K. Bagha, K. S. Parmar, M. Javaid, A. Haleem, and R. P. Singh, "Significant Applications of Machine Learning for COVID-19 Pandemic," Journal of Industrial Integration and Management, vol. 5, no. 4, pp. 453–479, 2020.

[5] N. Barraza, S. Moro, M. Ferreyra, and A. de la Peña, "Mutual information and sensitivity analysis for feature selection in customer targeting: A comparative study," Journal of Information Science, vol. 45, no. 1, pp. 53–67, 2019. [Online]. Available: https://doi.org/10.1177/0165551518770967

[6] L. Molina, L. Belanche, and A. Nebot, "Feature selection algorithms: a survey and experimental evaluation," in IEEE International Conference on Data Mining, 2002, pp. 306–313.

[7] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 13, pp. 5–32, 2001.

[8] M. Kursa and W. Rudnicki, "Feature selection with the Boruta package," Journal of Statistical Software, vol. 36, no. 11, pp. 1–13, 2010.

[9] United Nations Development Programme, "Human Development Insights," 2019. [Online]. Available: http://hdr.undp.org/en/indicators/137506
[10] D. Freedman and P. Diaconis, "In the histogram as a density estimator: L2 theory," Probability Theory and Related Fields, vol. 57, no. 4, pp. 453–476, 1981.
[11] J. Hartigan, Clustering algorithms. John Wiley & Sons, Inc., 1975.
[12] E. Forgy, "luster analysis of multivariate data: efficiency versus interpretability of classifications," Biometrics, vol. 21, no. 4, pp. 768–769, 1965.
[13] P. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," Computational and Applied Mathematics, vol. 20, pp. 53–65, 1987. [Online]. Available: https://doi.org/10.1016/0377-0427(87)90125-7
[14] T. Calinski and J. Harabasz, "A dendrite method for cluster analysis," Communications in Statistics, vol. 3, no. 1, pp. 1–7, 1987.
[15] D. Davies and D. Bouldin, "A cluster separation measure," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 1, no. 2, pp. 224–227, 1979.
[16] S. W. Smith, "Moving average filters," in Digital Signal Processing, S. W. Smith, Ed. Boston: Newnes, 2003, ch. 15, pp. 277–284. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780750674447500522
[17] Our World In Data, "Data on COVID-19 (Coronavirus)," https://github.com/owid/covid-19-data/tree/master/public/data, 2022, Accessed: 2022-08-08.
[18] Institute of Health Metrics and Evaluation, "COVID-19 resources," https://www.healthdata.org/covid, 2022, Accessed: 2022-08-08.
[19] M. Kuhn and K. Johnson, Applied Predictive Modeling. Springer, 2nd ed. 2018.
[20] H. Sturges, "The choice of a class interval," Journal of the American Statistical Association, vol. 21, no. 153, pp. 65–66, 1926.
[21] D. Scott, "Sturges' rule," WIREs Computational Statistics, vol. 1, no. 3, pp. 303–306, 2009.
[22] R. Sibson, "Slink: an optimally efficient algorithm for the single-link cluster method," The Computer Journal, British Computer Society, vol. 16, no. 1, pp. 30–34, 1973.
[23] D. Defays, "An efficient algorithm for a complete link method," The Computer Journal, British Computer Society, vol. 20, no. 4, pp. 364–366, 1977.
[24] G. Pena, J. Gambini, and N. R. Barraza, "COVID-19 feature selection datasets," https://data.mendeley.com/datasets/wbjjz9bzxf/1, 2022.
[25] G. Pena, "Epy feature selection," 2022, implemented in Python. [Online]. Available: https://doi.org/10.5281/zenodo.6988290
[26] D. Cournapeau and project volunteers, "scikit-learn: Machine Learning in Python," https://scikit-learn.org/stable/, 2022, Accessed: 2022-08-12.
[27] D. Homola, "scikit-learn-contrib: BorutaPy," https://github.com/scikit-learn-contrib/boruta_py, 2022, Accessed: 2022-08-12.
[28] F. Jannssen, "Changing contribution of smoking to the sex differences in life expectancy in Europe, 1950–2014," European Journal of Epidemiology, vol. 35, no. 9, pp. 835–841, 2020.
[29] M. Sahinoglu and H. Sahinoglu, "Consequences and lessons from 2020 pandemic disaster: Game-theoretic recalibration of covid-19 to mobilize and vaccinate by rectifying false negatives and false positives," International Journal of Computer Theory and Engineering, vol. 14, pp. 109–125, 01 2022.
[30] D. F. Morrison, Multivariate Statistical Methods, ser. McGraw Hill Series in Probability and Statistics. McGraw Hill, 1976.
[31] R. A. Johnson and D. Wichern, Applied Multivariate Statistical Analysis. Prentice Hall, 1982.
[32] M. Sahinoglu, Cyber-Risk Informatics: Engineering Evaluation with Data Science. John Wiley & Sons, 2016.

Gabriel Pena (Ph.D. student) received the degree in computer engineering from Universidad Nacional de Tres de Febrero. He currently holds an assistant professor position at Universidad Nacional de Tres de Febrero. He has participated in several research projects and received teaching and research scholarships. He has published articles and participated in Conferences. His research interests involve stochastic processes, probability and statistics, machine learning and signal processing.

Juliana Gambini received the B.Sc. degree in mathematics and the Ph.D. degree in computer science both from Universidad de Buenos Aires (UBA), Argentina in 1996 and 2006, respectively. She is currently a titular professor at the Instituto Tecnológico de Buenos Aires (ITBA), Buenos Aires, Argentina, a member of the Center for Computational Intelligence – ITBA and a Titular Professor at Universidad Nacional de Tres de Febrero, Pcia. de Buenos Aires, Argentina. She leads research projects and doctoral theses related to SAR image processing, machine learning and computational statistic.

Néstor R. Barraza (Ph.D. '1996) received his electronics engineering, and Ph.D degrees from the University of Buenos Aires, Argentina in 1993 and 1996 respectively. He holds a full professor position at Universidad Nacional de Tres de Febrero, and a part time professor at University of Buenos Aires, both in Argentina. His research interests involve stochastic modeling, software reliability, information theory and coding, machine learning and applied statistics. He has published many papers in important international journals and attended important Conferences, some of them as invited speaker. He was also invited to give courses and talks. He received many grants and financial support for his research projects, both, in the Academy as well as in the Industry, where he worked as software and communications engineer leading important technological projects. He is a Senior Member of the IEEE.