CADERNOS DO IME – Série Estatística

Universidade do Estado do Rio de Janeiro - UERJ
ISSN on-line 2317-4536 / ISSN impresso 1413-9022 - v.56, p.16-38, 2024
DOI: 10.12957/cadest.2024.84973

ASSESSING THE CONTRIBUTION OF CONTAINMENT MEASURES, VACCINATION COVERAGE AND MOBILITY IN THE EVOLUTION OF DEATHS FROM COVID-19 IN BRAZIL

Paulo Henrique Couto Simões
Universidade do Estado do Rio de Janeiro
ph.simoes@gmail.com

Abstract

This work proposes a method to rank the contribution of containment measures, vaccination coverage and mobility to contain the evolution of the COVID-19 pandemic in different states of Brazil. The proposed method applies the automatic learning of regression models using decision trees through the XGBoost algorithm. To interpret it globally, the SHapley Additive exPlanations (SHAP) was used, which is an algorithm based on Shapley's cooperative game theory to quantify the contribution of the analyzed characteristics to the evolution of the target variable (deaths). The evaluation results of the method indicated its efficiency to quantify the contribution of each variable in a robust way. It reveals that the percentages of vaccine coverage of the first and second dose, in addition to the closure of schools, were the measures that had the greatest contribution to the evolution of the number of deaths from COVID-19. The weighting of the variables can help the responsible actors in the elaboration of public policies to minimize the socioeconomic effects in their regions. Since, in countries with great social inequality, the use of only a few more efficient measures would be less harmful than a lockdown, as it would be extremely harmful to the quality of life of these poorer populations.

Keywords*: COVID-19, XGBoost, Shapley Values, Interpretability.*

1. Introduction

In December 2019, several cases of pneumonia have been associated with COVID-19, a disease caused by a new human coronavirus SARS-CoV-2 (ZHU *et al*., 2020). The new virus has a high capacity for inter-human transmission, so it has become imperative to implement methods to control its spread, such as social isolation, reduced mobility, information campaigns, and vaccination.

Due to its specific characteristics of contamination, it quickly spread around the world, and is classified by the World Health Organization as a pandemic in March 2020. COVID-19 has represented a considerable challenge for managers of public health policies, researchers, and doctors.

Several studies have been published revealing that experts were correct in warning that non-pharmacological measures (social distancing, interruption of non-essential activities, and the use of masks) were the best options to contain the spread of the virus (FLAXMAN *et al*., 2020; SHARMA *et al*., 2021). The containment measures were implemented sparsely and located in Brazil, and there was no centralization of these measures by the federal government, since the containment measures were defined and adopted by local authorities. (ANTUNES *et al*., 2020).

In recent years, several works and research have been prepared related to COVID-19. The pandemic crisis generated an unprecedented need and required the rapid learning of new skills (SOHRABI *et al*., 2020). In addition, containment measures such as social distancing, greater care with personal hygiene, lockdown, among others, were taken to contain the spread of the virus (CUCINOTTA & VANELLI, 2020).

COVID-19 has reached the planet's five continents, becoming a pandemic with millions of deaths. Vaccines were rapidly developed in just over a year, such as AstraZeneca (Oxford), CoronaVac (Sinovar), Pfizer (BioNTech), Janssen (Johnson & Johnson), Moderna, and Sputnik V (Russian vaccine). Even so, the necessary care continues to prevent the emergence of new variants of the virus, we do not know for sure when the pandemic will end, given that it has already reached several countries and infected millions of people around the world (SOHRABI *et al*., 2020).

COVID-19 has a mortality rate of approximately 3.7% which, in comparison, with influenza, which has a rate of less than 1%, is much more lethal (MEHTA *et al*., 2020). The risk of exposure to respiratory viruses such as SARS-COV-2 increases in crowded

and closed environments, with little air circulation because its main form of contagion is through the air (CUCINOTTA & VANELLI, 2020). The probability of a primary case transmitting COVID-19 in an indoor environment was 18.7 times higher compared to an outdoor environment (NISHIURA *et al*., 2020).

A series of machine learning models were used for the dataset of the mortality of patients with COVID-19. The dataset consisted of blood samples from 375 patients admitted to a hospital in the Wuhan region, China, where 201 survived hospitalization and 174 died. The focus of this study was not just on seeing which model can achieve the highest absolute accuracy, but on interpreting what the models provide. The variables age, days in hospital, lymphocytes, and neutrophils were found to be important and robust predictors in predicting a patient's mortality. After the start of vaccination coverage, which was an important turning point in the characteristics of the pandemic, several studies emerged around the world on the effectiveness of vaccines (DAGAN *et al*., 2021; CHEMAITELLY *et al*., 2021; LOPEZ BERNAL *et al*., 2021; JARA *et al*., 2021). Pfizer's effectiveness was 72% for days between 14 and 20 after the first dose. However, after seven or more days of the second dose, this efficacy reached 92% for severe disease development (DAGAN *et al*., 2021). Moderna, against critical or fatal cases, was approximately 82% effective after the first dose and 96% after the second dose (CHEMAITELLY *et al*., 2021). Regarding CoronaVac, the efficacy of two doses is approximately 75% among people with the Alpha variant and 67% for the Delta variant (LOPEZ BERNA *et al*., 2021). In complete immunization, the estimated vaccine efficacy was approximately 66% for prevention, 88% for prevention of hospitalization, 90% for prevention of ICU admissions, and 86% for prevention of death (JARA *et al*., 2021).

Knowledge about the socioeconomic spread of COVID-19 infections in Germany was assessed in different studies. In one of them, we sought to find out what the incidence rates of COVID-19 were and whether they would be different between municipalities according to their socioeconomic characteristics using a wide variety of indicators. A total of 204,217 COVID-19 diagnoses were used in the total German population, distinguishing five distinct periods between 1 January and 23 July 2020. For each period, age-standardized incidence rates of COVID-19 diagnoses were calculated at the level of County. Gradient-increasing models were trained to predict age-standardized incidence rates with the macrostructures of the municipalities and Shapley Additive Explanations

(SHAP) values were used to represent the 20 most prominent characteristics in terms of negative/positive correlations with the outcome variable. (DONG, DU, GARDNER, 2022).

To quantitatively identify the optimal control measures for regulators to minimize the growth and death rates of COVID-19, a multi-scale approach (global, continental, and national levels) was developed, which encompasses a series of systematic analyses. Predictive modeling of growth and mortality rates was developed, followed by the determination of the most effective control factors that can best minimize both parameters over time through explainable Artificial Intelligence with the SHAP method. Average MAPE scores for predicting COVID-19 growth and death rates were below 10% on both global and continental scales. It was demonstrated in the study that in a quantitative way the top three most effective control measures for regulators to minimize the growth rate were COVID-CONTACTTRACING, PUBLIC-GATHERING-RULES, and COVID-STRINGENCY-INDEX. The control factors related to death depend specifically on the modeling scenario (CHEW & ZHANG, 2022).

In a regional program for Brazil, a Scientific Committee for the Northeast region was created with the objective of proposing and articulating strategies to combat and mitigate COVID-19, as a great diversity of socioeconomic and human development contexts was observed at the regional level during the pandemic. It is possible to relate social isolation and living conditions in Brazilian states. The regions considered to have the highest poverty rate are those with the lowest percentage of Social Isolation (NATIVIDADE *et al*., 2020).

The present study proposes to rank the measures and present those that have the greatest impact on the evolution of the number of cases and deaths in Brazilian states and motivate the use of these more efficient measures, thus avoiding the application of a lockdown as much as possible. In this way, we would achieve a lower socioeconomic impact and better quality of life for these poorer populations.

Thus, it became opportune to carry out an analysis including containment measures, vaccination coverage, and population mobility. Therefore, 10 containment measures collected by the Oxford COVID-19 Government Response Tracker (OxCGRT) project (HALE *et al*., 2021), two vaccine coverage measures, and six mobility measures were analyzed in terms of their contribution to the evolution of the number of deaths

related to the pandemic in Brazil, in all 26 states and the Federal District. In this case, data on vaccination were extracted from the Information System of the National Immunization Program (SI-PNI) (MINISTÉRIO DA SAÚDE, 2022), and mobility data provided by Google (COVID-19 Community Mobility Reports) (AKTAY *et al*., 2022).

The structure of this study consists of an introduction, where the general plot about COVID-19 is presented, followed by a general theoretical review. Soon after, the method proposed in the work is presented, which brings information about the databases, the XGBoost algorithm, and the Shapley value. After presenting the method used, we will bring the contributions that were implemented in the academic context. In the results, we will describe the study that was carried out for Brazil and its states; where the objective was to classify and rank the contribution of containment measures, vaccine coverage, and population mobility on the evolution of deaths caused by COVID-19 after the start of vaccination. And finally, in the conclusions, we will discuss the results found in this study and other works related to COVID-19.

2. Data base

The study was done by analyzing ten indicators that make up the OxCGRT Stringency Index which is part of the Oxford COVID-19 Government Response Tracker initiative. The University of Oxford that is responsible for monitoring and reporting the different government responses to COVID-19, which are coded into 23 indicators such as school closures, travel restrictions, and vaccination policy. The Stringency index is calculated daily, and for a given day is the average score of ten metrics, each having a value between 0 and 100, being the same already provided by the OxCGRT dataset. These containment policies are scaled to reflect the extent of government action, and the scores are aggregated into a set of policy indices. The indicators used are: C1-school closing, C2-workplace closing, C3- cancel public events, C4- restrictions on gatherings, C5- close public transport, C6- stay at home requirements, C7- restrictions on internal movement, C8-international travel controls, H1-public info campaigns and H6-facial Coverings. Two measures of vaccination coverage were also used: D1-vaccinal coverage of the first dose and D2-vaccinal coverage of the second and single dose. And six mobility measures: M1-retail and recreation, M2-grocery and pharmacy, M3-parks, M4-transit stations, M5-workplaces, and M6-residential. Each of these 3 dimensions of analysis (containment

measures, vaccination coverage, and mobility) plays an important role in understanding it as a whole:

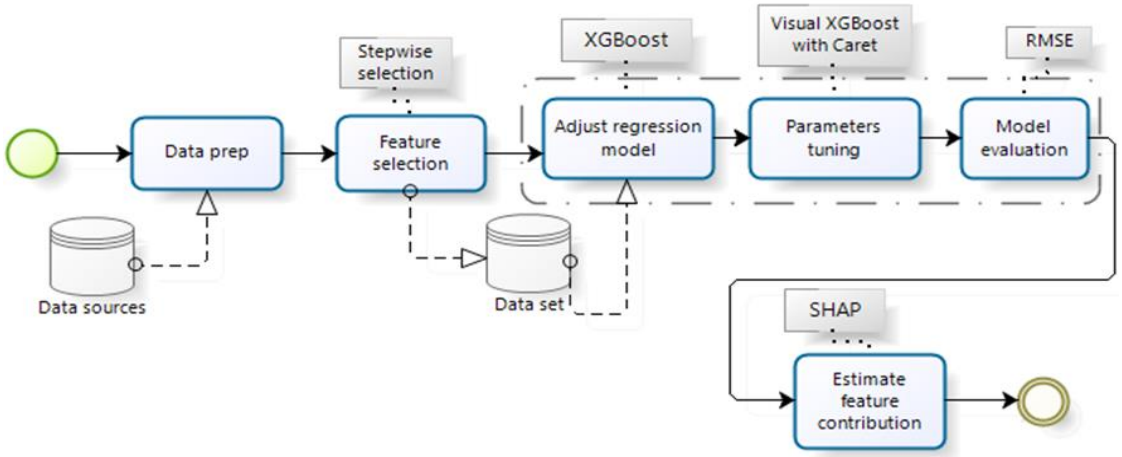
- Containment measures: it has the task of measuring the political responses that each government has taken. Records these policies on a scale to reflect the extent of government action to combat the spread of COVID-19.
- Vaccination coverage: it is the percentage measure of the number of people vaccinated in each region (country or state).
- Mobility: Mobility reports are intended to provide information on how the population movement pattern has changed in each geographic region (which can be a city, a state, or a country) each day in relation to the average population movement measured before the beginning of the pandemic.

These indicators were related to the evolution of the number of deaths from the pandemic in Brazil and all its 26 states plus the Federal District. In this case, data on vaccination were taken from the Information System of the National Immunization Program (SI-PNI) (MINISTÉRIO DA SAÚDE, 2022), and the data on mobility taken from Google COVID-19 Community Mobility Reports (AKTAY *et al.*, 2022).

3. Methodology

The construction procedure and the steps relevant to the method proposed by this work can be seen in Figure 1. The flowchart presents the steps used from the source and preparation of the data to the elaboration of the final models that estimate the contribution of the variables.

Figure 1: Flowchart of the ranking process of containment measures, vaccination coverage, and mobility using the XGBoost algorithm and the Shapley Additive Explanations (SHAP).



Source: Author (2022)

3.1 Data prep

In the first step of data preparation, the transformation of containment measures, vaccine coverage, and mobility measures were normalized using the MinMaxScalers technique to have all features between 0 and 1. This procedure was applied to prevent the algorithm as a whole from being biased by the variables with a higher order of magnitude. Each daily Xi value of each of the variables used in the study was divided by its maximum value.

$$\mathrm{Xp}_i = \frac{\mathrm{Xi - MIN(X)}}{\mathrm{MA(X) - MIN(X)}} \ , \ \mathrm{i} = 1,2,3, \dots , \mathrm{n} \qquad (1)$$

The software used for the analysis of the study was R (R CORE TEAM, 2019).

3.2. Feature selection

The stepwise variable selection method was also used. This method consists of a technique used to select statistically significant features from a set of predictor variables, inserting and removing these features in the model until there is no statistically valid reason to add or remove any more features. The stepwise methodology works through a series of Student's T tests or F tests (THOMPSON, 1995).

Together, the lags in days of each of the models and the response variable of the respective states were tested so that the lags combined with each of the models were also validated. In this way, it was possible to select the models with the lowest information criteria (BIC) and with all the statistically significant features for each of the states and Brazil.

3.3. Adjust regression model

To model the impact of indicators on the evolution of the number of daily deaths from COVID-19, a tree model was built for multiple regression using Extreme Gradient Boosting (XGBoost). The design of the XGBoost algorithm is based on an implementation of the Gradient Boosted Trees algorithm, which is a supervised learning method that uses function approximation, optimizing specific loss functions, as well as the application of regularization techniques (LUNDBERG *et al.* , 2019; MOLNAR, 2020). Where the objective function (loss and regularization function) at iteration t should be minimized.

$$L^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t), \tag{2}$$

such that $f(x + \Delta x)$ where $x = \hat{y}_i^{(t-1)}$.

The purpose of XGBoost is to build a function of functions, that is, l is a function of learning a sum of current and previous additive trees that cannot be optimized using traditional optimization methods in Euclidean space (CHEN & GUESTRIN, 2016). We need to transform the original objective function into a function of the Euclidean domain so that we can use traditional optimization techniques. The simplest linear approximation to a function f(x).

$$f(x) \approx f(a) + f'(a)(x - a), where\ \Delta x = f_t(x_i) \tag{3}$$

We can transform a function f(x) into a simpler function of Δx around a specific point using Taylor's theorem. Before the Taylor approximation, the x in the objective function f(x) was the sum of t trees and after that, it becomes a function only of the current tree at step t. In this case, f(x) is the loss function l, while a is the predicted value from the previous step (t-1) and Δx is the new learner we need to add at step t. Using the previous step in each iteration t we can write the objective function (loss) as a simple function of the new learner added and thus apply Euclidean space optimization techniques. The prediction at step (t-1) while (x-a) is the new learner we need to add at step (t) to minimize the goal. In a second-order Taylor approximation.

$$f(x) \approx f(a) + f'(a)(x - a) + \frac{1}{2} f''(a)(x - a)^2$$

$$\mathcal{L}^{(t)} \cong \sum_{i=1}^{n} \left[l\left(y_i, \hat{y}_i^{(t-1)}\right) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \tag{4}$$

where: $g_i = \partial_{\hat{y}^{(t-1)}} l\left(y_i, \hat{y}^{(t-1)}\right)$ and $h_i = \partial^2_{\hat{y}^{(t-1)}} l\left(y_i, \hat{y}^{(t-1)}\right)$.

By removing the constant parts, we have the simplified minimization equation at step t:

$$\tilde{\mathcal{L}}^{(t)} \cong \sum_{\mathrm{i}=1}^{\mathrm{n}} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \tag{5}$$

Since this is a sum of simple quadratic functions of one variable and can be minimized using known techniques, then our next objective is to find a learner that minimizes the loss function at iteration t.

$$argmin_x \, G_x + \frac{1}{2} H x^2 = -\frac{G}{H}, H > 0, min_x \, G_x + \frac{1}{2} H x^2 = -\frac{\frac{1}{2} G^2}{H} \tag{6}$$

$$\tilde{\mathcal{L}}^{(t)} = -\frac{1}{2} \sum_{\mathrm{J}=1}^{\mathrm{T}} \frac{\left(\sum_{\mathrm{i} \in \mathrm{I_j}} \mathrm{g_i} \right)^2}{\sum_{\mathrm{i} \in \mathrm{I_j}} \mathrm{h_i} + \lambda} + \gamma \mathrm{T} \tag{7}$$

where I am the set of indices of the samples contained in the leaf of the decision tree, g_i is the gradient (first-order derivative) and h_i is the Hessian (second-order derivative).

The qualification function above returns the minimum loss value for a given tree structure, which means that the original loss function is evaluated using the ideal weight values. So, for any tree structure, we have a way to calculate the ideal weights of the leaves. What we do for learning construction is to start with a single root, iterate over all resources and values per resource and evaluate each possible loss reduction and gain for the best split should be positive (CHEN *et al*., 2015).

3.4. Parameters tuning

To optimize the hyperparameters of the XGBoost model, a visual implementation of the Caret package (KUHN *et al*., 2008) from R version 4.1.2 was used. This is a process of optimizing the hyperparameters of XGBoost. It focuses on visually evaluating the steps in the process while building a simple, stepwise logic to fine-tune the model. This uses five steps to optimize the hyperparameters: in the first step, the eta learning rate and the number of iterations are established; in the second step, the maximum depth of a tree and the minimum sum of the instance weight is defined. In the third step of the process, the subsampling ratio of the columns and the subsampling ratio of the rows are optimized. In the fourth step, experiments are performed with different gamma values. And finally, in

the fifth step, the learning rate is optimized (HASTIE, TIBSHIRANI, FRIEDMAN, 2009).

The idea is to initially use a higher learning rate to adjust the hyperparameters. These parameters will be used to fit the final model with a higher number of trees and a lower learning rate. Another must at this stage of the process is to use cross-validation, as a simple random resampling of the time series is not the best way to resample the data. Block techniques were applied to divide the original dataset into training and test sets along the time series, using the timeslices function of the Caret package (KUHN *et al*., 2008).

3.5. Model evaluation

The mean square error or mean square deviation is one of the most commonly used measures to assess the quality of forecasts. It shows how far predictions fall from true values measured using Euclidean distance. To calculate the RMSE, calculate the residual (difference between prediction and truth) for each data point, calculate the norm of the residue for each data point, and calculate the mean of the residuals as the square root of that mean. RMSE is commonly applied to supervised learning models (CHAI *et al*., 2014).

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} \left|\left|y(i) - \hat{y}(i)\right|\right|^2}{N}}, \tag{8}$$

where N is the number of data points, $y(i)$ is theth measurement, and $\hat{y}(i)$ is its corresponding prediction.

The importance of the RMSE is to have a single number to evaluate the performance of a model, whether during training, cross-validation, or post-deployment monitoring. The mean square error is one of the most used measures for this. It's a proper scoring rule, intuitive to understand, and compatible with some of the most common statistical assumptions.

3.6. Estimate feature contribution

The purpose of SHapley Additive exPlanations (SHAP) is to explain the prediction of an instance X by calculating the contribution of each feature to the prediction. The SHAP explanation method calculates Shapley values from coalition game

theory. A dice instance's resource values act like players in a coalition. Shapley values tell us how to fairly distribute the "payout" (= the forecast) across resources. A player can be an individual resource value, for example for tabular data. A player can also be a group of resource values. For example, to explain an image, pixels can be grouped into superpixels and the prediction is distributed between them. An innovation that SHAP brings is that the Shapley value explanation is represented as an additive resource assignment method, a linear model (MOLNAR, 2020). SHAP specifies the explanation as:

$$(z') = \emptyset_0 + \sum_{j=1}^{M} \emptyset_0 z'_j \tag{9}$$

Of which g where g is the model of explanation $z' \in \{0,1\}^M$, is the coalition vector, M is the maximum size of the coalition and is the feature assignment for a feature j, the Shapley values. What I call the "coalition vector" is called "simplified resources" in SHAP. It is important to think of the z's as a description of coalitions: In the coalition vector, an entry of 1 means that the corresponding resource value is "present" and 0 that it is "absent". To calculate Shapley values, we simulate that only some resource values are being reproduced ("present") and some not ("missing"). The representation as a linear model of coalitions is a trick for calculating ∅'s. For x, the instance of interest, the coalition vector X is a vector of all 1's, that is, all characteristic values are "present". The formula simplifies to:

$$g(x') = \emptyset_0 + \sum_{j=1}^{M} \emptyset_j \tag{10}$$

The Shapley value is a solution that satisfies the properties of Efficiency, Symmetry, Dummy (Shapley axiom of Dummy, which says that a feature that does not contribute to the result must have a Shapley value of zero), and Additivity. SHAP has the following three desirable properties:

3.6.1. Local accuracy

$$f(x) = g(x') = \emptyset_0 + \sum_{j=1}^{M} \emptyset_j \, x'_j \tag{11}$$

If you set $\emptyset_0 = E_X(\hat{f}(x))$ and set all x'_j to 1, this is Shapley's efficiency property. Just with a different name and using the coalition vector.

$$f(x) = g(x') = \emptyset_0 + \sum_{j=1}^{M} \emptyset_j \, x'_j = E_X\left(\hat{f}(x)\right) + \sum_{j=1}^{M} \emptyset_j \, x'_j \tag{12}$$

3.6.2. Missingness

Missingness says that a missing resource is assigned an assignment of zero. Note that it refers to coalitions, where a value of 0 represents the absence of a characteristic value. In coalition notation, all characteristic values of the instance to be explained must be '1'. The presence of a 0 would mean that the resource value is missing for the instance of interest. This property is not among the properties of "normal" Shapley values. So why do we need this for SHAP. Lundberg calls this "secondary accounting ownership". A missing feature could - in theory - have an arbitrary Shapley value without harming the local precision property as it is multiplied by. The Missingness property enforces that missing features get a Shapley value of 0. In practice, this is only relevant for features that are constant.

3.6.3. Consistency

Be $f_x(z') = f\left(h_x(z')\right)$ e $z'_{\setminus j}$ indicates that $z'_j = 0$. For all models f and f" that satisfy:

$$f'_x(z') - f'_x\left(z'_{\setminus j}\right) \geq f_x(z') - f_x\left(z'_{\setminus j}\right) \tag{13}$$

for all entries z'$\in \{0,1\}^M$ that: $\emptyset_j(f', x) \geq \emptyset_j(f, x)$

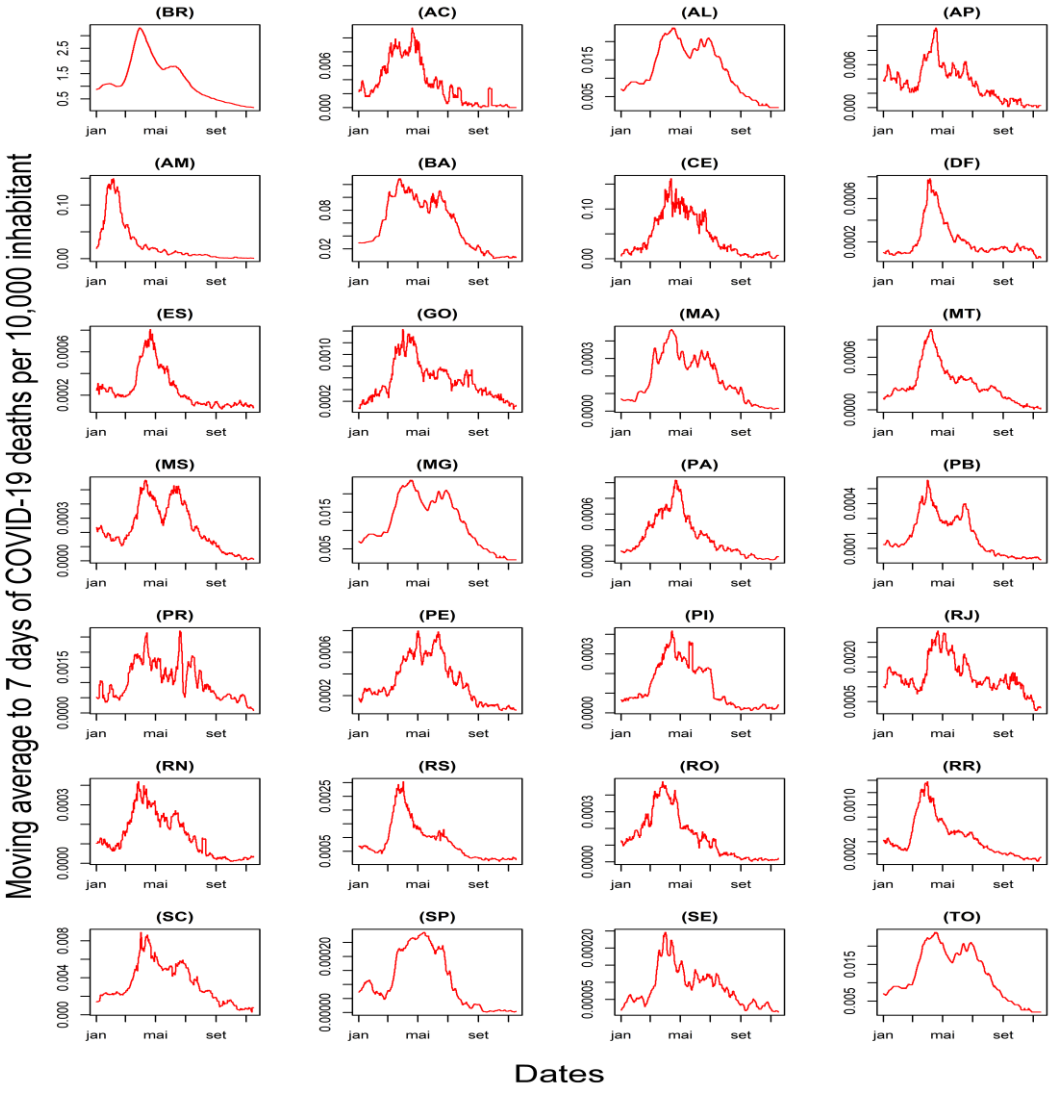
The consistency property says that if a model change such that the marginal contribution of a resource value increases or remains the same (independently of other resources), the Shapley value also increases or remains the same. From Consistency,

Shapley's properties follow Linearity, Dummy, and Symmetry (ALGABA, FRAGNELLI, SÁNCHEZ-SORIANO, 2019).

4. Results

The evolution of deaths from COVID-19 in 2021 in Brazil and its states had very different characteristics. Figure 2 illustrates 28-time series profiles in the different scenarios of the pandemic, where peaks and valleys are presented at different times for different states. The country's peak occurred in April/2021 when the highest rate of deaths was obtained, then declined until the end of May/2021. However, in June/2021, deaths started to grow again, having, from then on, a gradual reduction until the end of 2021.

Figure 2: Time series chart of seven-day moving averages for deaths on the national (BR) and state (AC-TO) scales for the year 2021.

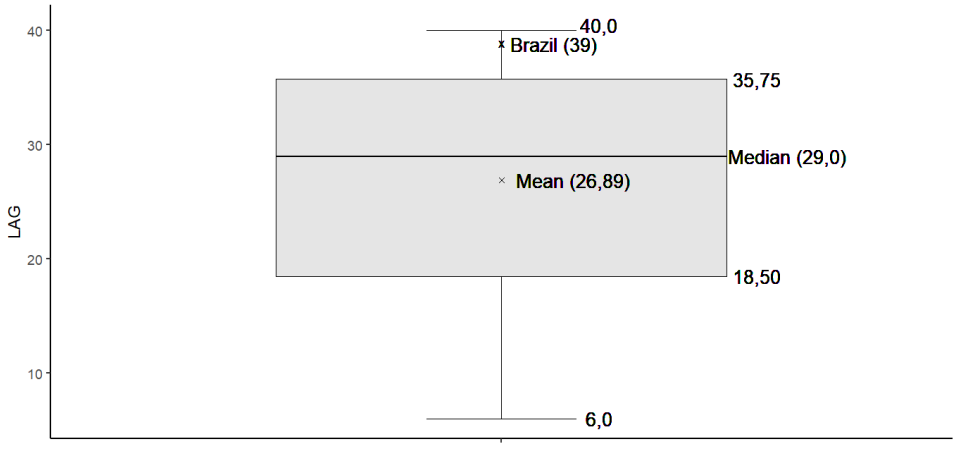


Source: Author (2022)

We can see that the time series patterns for each state were not similar. However, all states showed an increase at the beginning of the year, especially Amazonas, which had the highest peak among all Brazilian states in this period. The states of Amapá, Paraná, and Rio de Janeiro showed three waves over this period. Other states had two waves (as well as Brazil, a bigger peak and a second smaller one): Alagoas, Bahia, Goiás, Maranhão, Mato Grosso do Sul, Minas Gerais, Paraíba, Pernambuco, Piauí, Rio Grande do Norte, Rio Grande do Sul, Santa Catarina, São Paulo, Sergipe, and Tocantins. The others, including the Federal District, showed only one peak in 2021, which occurred in the first half of the year. In absolute numbers, São Paulo was the record holder. The state recorded 104,632 deaths caused by the virus.

For Brazil and each of the 27 states, the stepwise method was applied to select the statistically significant features from a set of predictor variables (18 features). Together, the lags in days of each of the models and the response variable (deaths) of the respective states were tested so that the lags combined with each of the models were also validated. Among the lag values that we can highlight in the boxplot of Figure 3, there is Brazil which presented 39 days of lag between explanatory variables and deaths. The states of Rio Grande do Sul, Rondônia, and Santa Catarina also stand out with 40 days of lag, which is the highest value presented in this study.

Figure 2: Boxplot of the lag in days of the 26 Brazilian states, the Federal District and Brazil between predictor variables of your model and deaths.



Source: Source: Author (2022)

It was also relevant to observe the median of the boxplot in Figure 3, which was 29 days lag, referring to the states of Amapá, Piauí, and Roraima. And with the smallest lag obtained in this study, we have the state of Pará with only 6 days. In general, we have

a boxplot with negative asymmetry that points to a higher concentration of values above the median (29 days) and without the presence of outliers. This difference between the lags of the selected state features and the deaths in each state, respectively, may have resulted from the different implementations of the restriction measures. In Brazil, the Federal Supreme Court gave the states, the Federal District, and municipalities the competence to decide to implement social distancing measures to mitigate and suppress COVID-19 (SUPREMO TRIBUNAL FEDERAL, 2020).

In this sense, few measures were implemented at the federal level, limited to restricting the entry of foreigners into the country and determining that people over sixty years of age observe social distancing (MINISTÉRIO DA SÁUDE, 2020). The suspension of events and/or the quarantine of risk groups were the first measures to be adopted, with the exception of Tocantins, which first implemented the suspension of classes. Espírito Santo, Distrito Federal, Mato Grosso do Sul, Tocantins, São Paulo and Rio Grande do Norte did not restrict intercity and/or interstate passenger transport, and 20 states did not implement quarantine for the entire population (SUMMAN & NANDI, 2022). The first social distancing measures implemented in Brazil took place in the Federal District on March 11, 2020. In the other states, most measures were implemented in the second half of March 2020 (MINISTÉRIO DA SAUDE, 2020). Mato Grosso do Sul, Santa Catarina and Rio Grande do Sul were the states that adopted these groups of measures in a shorter period of time, with a difference of one to two days. At the other extreme, in Pará, the time between the implementation of the first measure and the economic stoppage was 50 days. In 74% of the states, the time between the implementation of the first measure and the stoppage (full or partial) was equal to or less than one week (SUMMAN & NANDI, 2022).

Most states implemented the measures before the first death, especially the Tocantins, which began its implementation thirty days before the first reported death. Regarding the implementation of the partial stoppage category, Pará, Rio Grande do Norte, and São Paulo were the only ones that adopted this measure after the first death, with an interval of 34, 5, and 7 days, respectively (SUMMAN & NANDI, 2022).

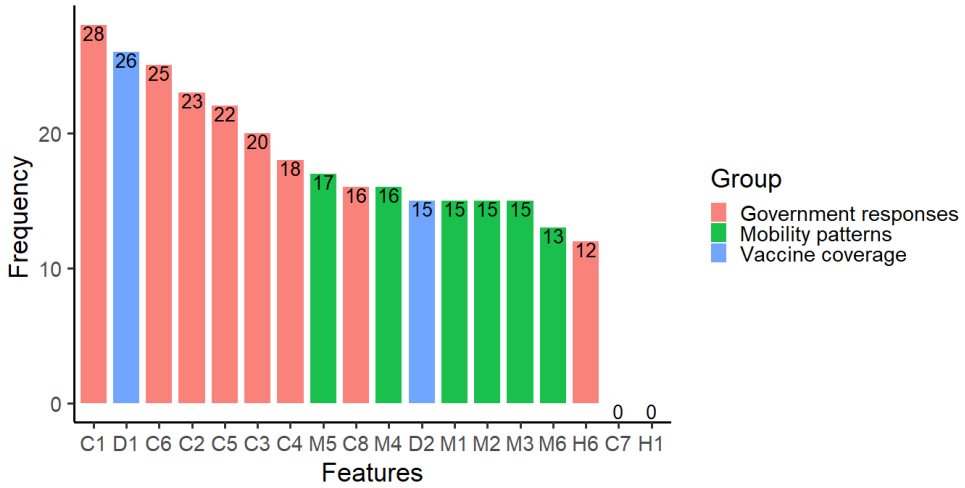
When assessing the timing of the implementation of social distancing measures in several countries around the world, they identified that the suspension of classes occurred, on average, 13 days after the first detected case of COVID-19, followed by restrictions

on international air travel with an average of 18 days and national lockdowns averaging 21 days after the first case of the disease (SUMMAN & NANDI, 2022).

The authors also identified that poorer countries, with a greater number of notifications after two weeks of the first case, less democratic and less populated systems, and with a younger and less dense population implemented these measures earlier, while countries with higher incomes, larger populations and more prepared in relation to the health system response adopted measures later in relation to the occurrence of COVID-19 cases locally (SUMMAN & NANDI, 2022). Therefore, it is possible to attribute the finding in the present study in relation to the differences in state lags in relation to the number of deaths compatible with the literature.

In Figure 4, we have the frequency in which each feature appears in the models selected for Brazil and each of the 27 Brazilian Federative Units. Feature C1 (school closing) was selected as significant in all scenarios (Brazil and states). The feature D1 (vaccinal coverage of the first dose) was not significant only in the states of Alagoas (AL) and Bahia (BA). The D2 feature (vaccinal coverage of the second and single dose) in the states of Amazonas (AM), Roraima (RR), Sergipe (SE), Piauí (PI), Mato Grosso do Sul (MS), Rondônia (RO), Ceará (CE), Acre (AC) were also not selected because they were not statistically significant in the selection of variables for the best model. Features C7 (restrictions on internal movement) and H1 (public info campaigns) were not selected in any scenario (at the national and state levels) because they were not statistically significant in the selection of variables applied to the best models.

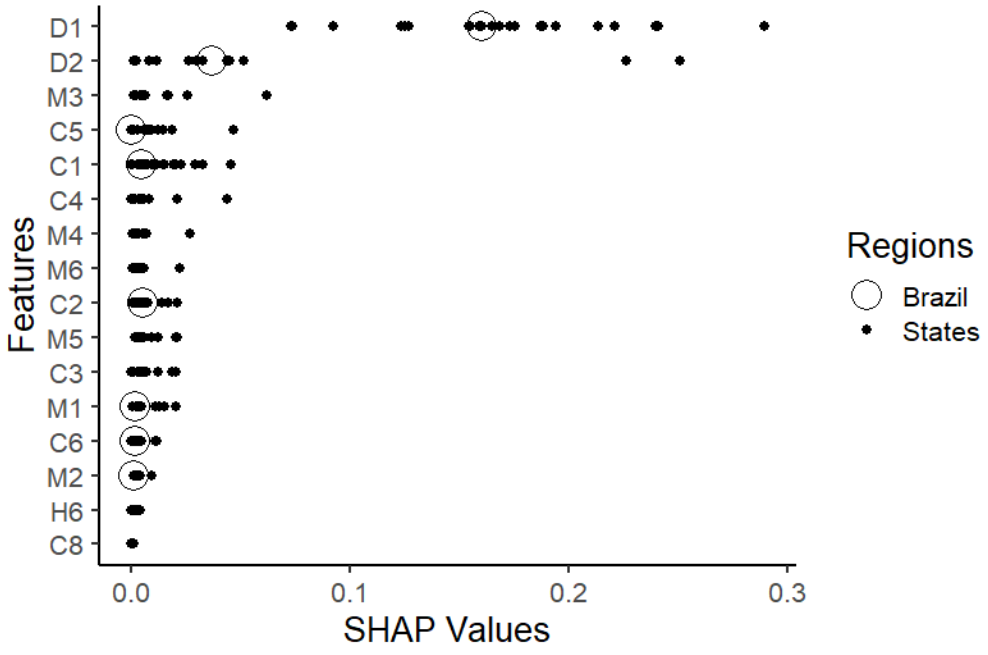
Figure 4: Features related to deaths from COVID-19 in order of frequencies selected by the 28 study models and separated by colors related to the respective dimensions.



Source: Author (2022)

The quantitative contribution of each feature is measured by its absolute contribution from the Shapley value. To quantify the contributions of each feature processed by the XGBoost algorithm in the analysis, SHAP values were calculated. For the analysis of the importance of the features via SHAP, Figure 5 illustrates the classification of the different measures (Vaccine coverage, Containment measures, and Mobility patterns) in terms of their absolute average impacts in Brazil and in the states, respectively. The analysis focuses on determining the most influential resources that can best interfere with the evolution of the target parameter (deaths). In relation to Brazil, Figure 5 presents the features D1 (vaccinal coverage of the first dose) and D2 (vaccinal coverage of the second and single dose) with a higher absolute contribution of SHAP value. Thus, they are the two that have the greatest influence on the evolution of death. On the other hand, features C6 (stay at home requirements) and M2 (grocery store and pharmacy) have the lowest absolute contributions of SHAP values, thus having the least influence on the evolution of deaths in Brazil. Regarding Brazilian states, Figure 5 indicates that, in general, D1 (vaccinal coverage of the first dose) and D2 (vaccinal coverage of the second and single dose) are the two features that have the highest absolute contributions in the states.

Figure 5: Importance of features via Shapley value illustrates the classification of different measures (vaccination coverage, restriction measures and mobility patterns) in terms of their absolute average impacts in Brazil and in the states.



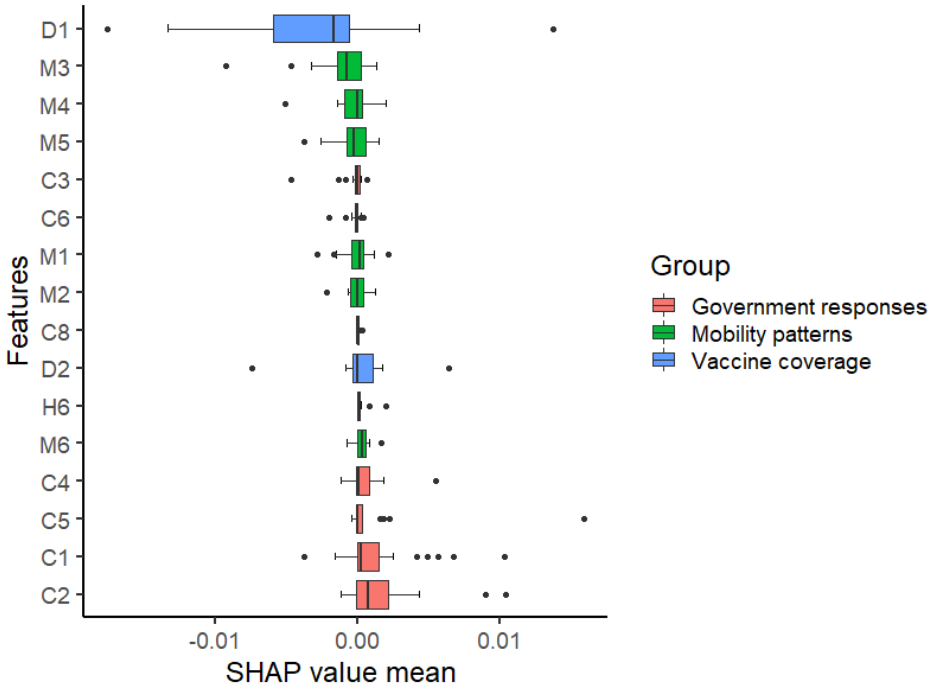
Source: Source: Author (2022)

We can still mention the features C8 (international travel controls) and H6 (facial Coverings) with the lowest absolute contributions of SHAP values for all 27 states.

For the analysis of the importance of the average values (positive or negative) of SHAP at the national and state level, Figure 6 illustrates the ranking of the boxplots of the different features in descending order of contribution in terms of their median impacts on the evolution of the number of deaths per COVID-19. It is observed that the negative SHAP value indicates that the specific resource has an inverse correlation with the target parameter (deaths), whereas the positive SHAP values have a direct correlation with the parameter. With that, we are here in Figure 6 evaluating the direction that each specific feature contributes to the evolution of deaths, where negative feature values contribute to the reduction of deaths and vice versa.

In Figure 6, feature D1 (vaccinal coverage of the first dose) has the highest negative median contribution, while feature C2 (closing of workplaces) has the highest positive median contribution. The greater the length of the box, the greater the distance between the first and third quartiles, which implies that the range of variation of the distributions is greater. Thus, we can compare the length of the boxplots of the features to determine which of the distributions has a greater variation.

Figure 6: The ranking of the boxplots of the different features in descending order of contribution in terms of their median impacts on the evolution of the number of deaths from COVID-19.



Source: Source: Author (2022)

5. Final remarks

The study aimed to analyze the evolution of deaths by COVID-19, applying SHAP values to interpret Poisson Regressions build through XGBoost algorithm. In this sense, the study worked with data on containment measures, vaccination coverage, and variation of population mobility that impacted evolution of the number of deaths by COVID-19 in Brazil and in all its 26 states plus the Federal District, totaling 26 states.

A regression model was built from a set of predictors that were previously selected by the stepwise method. Considering all the models (one for each state plus Brazil), 18 variables were used, ten of containment measures, two of vaccination coverage, and six of mobility. The lags (in days) were also validated for each state and, in the end, the models with the lowest Bayesian Information Criteria (BIC) were obtained.

When studying the lags obtained between the selected variables and the evolution of deaths in each state it is emphasized that the states of Rio Grande do Sul, Rondônia, and Santa Catarina (40 days) and Brazil itself (39 days) obtained the highest values. Thesmallest lag occurred in Pará (6 days). Other studies shows that the expected is a lag of 30 to 40 days after the application of the measures for an observable effect of the same on the evolution of the number of deaths studies (MOURA, 2021; PEIXOTO, 2020). One of the possible reasons for a 34-day span for the state of Pará may come from the different scenarios of the evolution of the number of deaths by COVID-19, reflecting the behavior of the population of each of the states and of Brazil in relation to the containment measures implemented by the competent authorities. However, the speed with which vaccination coverage evolved in each of the 28 regions does not seem to have influenced the difference in lags between Brazilian states (DE CASTRO-NUNES & DA ROCHA RIBEIRO, 2023).

In all model, the most frequent variable was school closure, present in all 28 states (plus Brazil) studied, followed by first-dose vaccination coverage, present in 26 states First-dose vaccination coverage was not statistically significant in the construction of the model in Alagoas and Bahia However, in six states, vaccination coverage stood out with a Shapley value greater than 0.2: Sergipe (0.290), Piauí (0.241), Maranhão (0.240), Mato Grosso do Sul (0.221), Rondônia (0.221) and Ceará (0.213).

Second-dose and single-dose vaccine coverage were not statistically significant in the construction of the model in the states of Amazonas, Roraima, Sergipe, Piauí, Mato

Grosso do Sul, Rondônia, Ceará, and Acre Two states stood out with a Shapley value of second-dose vaccine coverage greater than 0.2: Bahia (0.251) and Alagoas (0.226). It is possible that the proximity between the dates of the first and second vaccination doses have influenced these values for these two states. In Bahia, the start of the first dose was on 01/19/2021, and the second dose was on 02/16/2021. In Alagoas, the start of the first dose was on 01/19/2021, and the second dose was on 02/11/2021.

Restrictions on internal movement and public information campaigns were not selected in any scenario. Specifically, these two measures were the only ones that remained constant throughout 2021 in all states, so the models did not capture any contribution from them to the evolution of the number of deaths.

In general, due to the complexity of the evolution of deaths at the national and state level, each state adopted containment measures with different rigors and, additionally, with different population mobility indices for the same measures due to the heterogeneous behavior of the Brazilian population (GALINDO, SILVA, PEDREIRA JUNIOR, 2022). Another relevant measure was the vaccination coverage that, despite having started almost at the same time in all states, its evolution was quite uneven in the comparison between states (FLEURY & FAVA, 2022). Due to this variability of scenarios, the models proposed for Brazil and states were well diversified both in terms of measurements and lag in days.

The quantitative contribution of each measure was evaluated by its absolute contribution of the Shapley value. In relation to Brazil and the states, it was vaccination coverage that most contributed to the evolution of deaths (ORELLANA *et al*., 2022). On the other hand, for Brazil, the stay-at-home requirements and the displacement of people to grocery stores and pharmacies were the measures that were least influenced. And for the states, it was the control of international travel and the use of masks that had the least influence.

The study selected 28 regions (26 Brazilian states, the Federal District, and Brazil) analyzing the contribution of vaccine coverage (which is a pharmacological measure), the population mobility variation index provided by Google (AKTAY *et al*., 2022), and measures containment system (Oxford) to better understand the evolution of the number of deaths in the post-vaccination period, until December/2021.

COVID-19 has resulted in high mortality worldwide. The researchers have built several pandemic models based on data limited by the difficulty of collection and in most cases, they focus on analyzing the path of spread and the impact of the pandemic. However, many states geographic factors were also crucial for the predictions of confirmed COVID-19 cases in Brazilian states. In this study, we considered data from daily time series of confirmed deaths and attributes of the 27 federative units in Brazil. To contribute to the growing body of knowledge of the global community related to the control and management of COVID-19, this article develops a data-driven approach that uses an explainable algorithm, through the SHAP method (SHapley Additive exPlanations), to explain the effect of containment, vaccination, and mobility measures in the evolution of deaths by COVID -19.

Since its inception in late 2019, COVID-19 has greatly disrupted the daily life of the global community due to its highly infectious nature, where more than 349 million individuals have been infected by the virus. Due to the disparate socio-environmental characteristics in different countries on various continents, it has been difficult for regulators to identify a universal set of optimal control measures that can best control the behaviors of COVID-19 cases and death growth at various spatial scales (global, continental, and state scales). The complexity reinforces the difficulty faced by public managers in implementing the most appropriate control measures to better combat the spread of COVID-19. Therefore, the question remains about how different localities can implement more appropriate control policies, such as restriction measures, vaccine distribution strategies, vaccination prioritization criteria, face mask policies, and others.

With this intention, it is possible to expand the methodology of the present work to the global and continental scale. The current literature lacks a generic model structure that can systematically and effectively investigate the impacts of the multiplicity of control measures on the growth of cases and deaths, bringing recommendations of the most effective that can better minimize the two parameters related to COVID-19. 19 over time.

References

ALGABA, E.; FRAGNELLI, V.; SÁNCHEZ-SORIANO, J. (Ed.). **Handbook of the Shapley Value.** CRC Press, 2019.

AKTAY, A. et al. **Google COVID-19 community mobility reports**: anonymization process description, 2022. Disponível em: https://arxiv.org/abs/2004.04145. Acessado em: 20/05/24.

ANTUNES, B. B. de P. et al. Progressão dos casos confirmados de COVID-19 após implantação de medidas de controle. **Revista Brasileira de Terapia Intensiva**, v. 32, p. 213-223, 2020.

CHAI, T.; DRAXLER, R. R. Root mean square error (RMSE) or mean absolute error (MAE)–Arguments against avoiding RMSE in the literature. **Geoscientific model development**, v. 7, n. 3, p. 1247-1250, 2014.

CHEMAITELLY, H. et al. Eficácia da vacina mRNA-1273 COVID-19 contra B. 1.1. 7 e B. 1.351 variantes e doença grave de COVID-19 no Catar. **Medicina da Natureza**, v. 27, n. 9, pág. 1614-1621, 2021.

CHEN, T. et al. Xgboost: extreme gradient boosting. **R Package version 0.4-2**, v. 1, n. 4, p. 1-4, 2015.

CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system**. In**: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. p. 785-794, 2016.

CHEW, A. W. Z.; ZHANG, L. Data-driven multiscale modelling and analysis of COVID-19 spatiotemporal evolution using explainable AI. **Sustainable Cities and Society**, v. 80, p. 103772, 2022.

CUCINOTTA, D.; VANELLI, M. WHO declares COVID-19 a pandemic. **Acta Bio Medica: Atenei Parmensis**, v. 91, n. 1, p. 157, 2020.

DAGAN, N. et al. BNT162b2 mRNA Covid-19 vaccine in a nationwide mass vaccination setting. **New England Journal of Medicine**, v. 384, n. 15, p. 1412-1423, 2021.

DE CASTRO-NUNES, P.; DA ROCHA RIBEIRO, G. Equidade e vulnerabilidade em saúde no acesso às vacinas contra a COVID-19. **Revista Panamericana de Salud Pública**, v. 46, p. e31, 2023.

DONG, E.; DU, H.; GARDNER, L. An interactive web-based dashboard to track COVID-19 in real time. **The Lancet infectious Diseases**, v. 20, n. 5, p. 533-534, 2020.

FLAXMAN, S. et al. Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. **Nature**, v. 584, n. 7820, p. 257-261, 2020.

FLEURY, S.; FAVA, V. M. D. Vacina contra Covid-19: arena da disputa federativa brasileira. **Saúde em Debate**, v. 46, p. 248-264, 2022.

GALINDO, E. P.; SILVA, S. P.; PEDREIRA JUNIOR, J. U. Impactos fatais da COVID-19 nos trabalhadores brasileiros. **Repositório do Conhecimento do IPEA**. v. 27. 2022. Disponível em: https://repositorio.ipea.gov.br/bitstream/11058/11084/2/NT_27_Dirur_Impactos_fatais.pdf. Acessado em: 25/05/24.

HALE, T. et al. **Oxford COVID-19 government response tracker** (OxCGRT). Last updated, v. 8, p. 30, 2021.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **Boosting and Additive Trees.** In: The elements of statistical learning. Springer, New York, NY, p.337-384, 2009.

JARA, A. et al. Effectiveness of an inactivated SARS-CoV-2 vaccine in Chile. **New England Journal of Medicine**, v. 385, n. 10, p. 875-884, 2021.

KUHN, M. Building predictive models in R using the caret package. **Journal of Statistical Software**, v. 28, p. 1-26, 2008.

LOPEZ BERNAL, J. et al. Effectiveness of Covid-19 vaccines against the B. 1.617. 2 (Delta) variant. **New England Journal of Medicine**, v. 385, n. 7, p. 585-594, 2021.

LUNDBERG, S. M. et al. **Explainable AI for trees**: From local explanations to global understanding. 2019. Disponível em: https://arxiv.org/abs/1905.04610. Acessado em: 18/05/24.

MEHTA, P. et al. COVID-19: consider cytokine storm syndromes and immunosuppression. **Lancet** (London, England), v. 395, n. 10229, p. 1033, 2020.

MINISTÉRIO DA SAÚDE (BR). **Sistema de Informação do Programa Nacional de Imunizações SI-PNI**. 2022. Disponível em: http://pni.datasus.gov.br. Acessado em: 08/05/24

MINISTÉRIO DA SÁUDE. **Portaria nº 454 de 20 de março de 2020**. Declara, em todo o território nacional, o estado de transmissão comunitária do coronavírus (covid-19). Diário Oficial da União 2020; 20 mar. Disponível em: https://www.planalto.gov.br/ccivil_03/portaria/prt454-20-ms.htm#:~:text=Portaria%20n%C2%BA%20454%2D20%2Dms&text=Declara%2C%20em%20todo%20o%20territ%C3%B3rio,coronav%C3%ADrus%20(covid%2D19).&text=Considerando%20a%20necessidade%20de%20dar,Art. Acessado em 16/05/24.

MOLNAR, C. **Interpretable machine learning: a guide for making black box models explainable.** 2 ed. Leanpub Publisher, 2020. Disponível em: https://christophm.github.io/interpretable-ml-book. Acessado em: 24/04/24.

NATIVIDADE, M. dos S. et al. Distanciamento social e condições de vida na pandemia COVID-19 em Salvador-Bahia, Brasil. **Ciência & Saúde Coletiva**, v. 25, p. 3385-3392, 2020.

NISHIURA, H. et al. Closed environments facilitate secondary transmission of coronavirus disease 2019 (COVID-19). **MedRxiv**, p. 2020.02. 28.20029272, 2020.

ORELLANA, J. D. Y. et al. Mudanças no padrão de internações e óbitos por COVID-19 após substancial vacinação de idosos em Manaus, Amazonas, Brasil. **Cadernos de Saúde Pública**, v. 38, p. PT192321, 2022.

PEIXOTO, V. R.; VIEIRA, A.; AGUIAR, P.; SOUSA, P.; ABRANTES, A. Timing, Adesão e Impacto das Medidas de Contenção da COVID-19 em Portugal. **Escola Nacional de Saúde Pública**, 2020. Disponível em: https://www.unl.pt/sites/default/files/impacto_das_medidas_de_contencao_da_covid-19_em_portugal_3_maio_final.pdf. Acesso em: 12/05/24.

R CORE TEAM. **R: A language and environment for statistical computing**. 2019.

SHARMA, M. et al. Understanding the effectiveness of government interventions against the resurgence of COVID-19 in Europe. **Nature communications**, v. 12, n. 1, p. 1-13, 2021.

SOHRABI, C. et al. World Health Organization declares global emergency: A review of the 2019 novel coronavirus (COVID-19). **International journal of surgery**, v. 76, p. 71-76, 2020.

SUMMAN, A.; NANDI, A. Timing of non-pharmaceutical interventions to mitigate COVID-19 transmission and their effects on mobility: a cross-country analysis. **The European Journal of Health Economics**, v. 23, n. 1, p. 105-117, 2022.

THOMPSON, B. Stepwise regression and stepwise discriminant analysis need not apply here: A guidelines editorial. **Educational and psychological measurement**, v. 55, n. 4, p. 525-534, 1995.

ZHU, N. et al. A novel coronavirus from patients with pneumonia in China, 2019. **New England Journal of Medicine**, v. 382, n. 8, p. 727-733, 2020.