ISCTE IUL
REPOSITÓRIO
INSTITUTO UNIVERSITÁRIO DE LISBOA

Repositório ISCTE-IUL

Deposited in *Repositório ISCTE-IUL*:
2025-05-19

Deposited version:
Accepted Version

Peer-review status of attached file:
Peer-reviewed

Citation for published item:
Duarte, M., Ferreira da Silva, C. & Moro, S. (2025). Machine learning models to predict the COVID-19 reproduction rate: Combining non-pharmaceutical interventions with sociodemographic and cultural characteristics. Informatics for Health and Social Care. 50 (2), 81-99

Further information on publisher's website:
10.1080/17538157.2025.2491517

Publisher's copyright statement:
This is the peer reviewed version of the following article: Duarte, M., Ferreira da Silva, C. & Moro, S. (2025). Machine learning models to predict the COVID-19 reproduction rate: Combining non-pharmaceutical interventions with sociodemographic and cultural characteristics. Informatics for Health and Social Care. 50 (2), 81-99, which has been published in final form at https://dx.doi.org/10.1080/17538157.2025.2491517. This article may be used for non-commercial purposes in accordance with the Publisher's Terms and Conditions for self-archiving.

Use policy

Creative Commons CC BY 4.0
The full-text may be used and/or reproduced, and given to third parties in any format or medium, without prior permission or charge, for personal research or study, educational, or not-for-profit purposes provided that:

- a full bibliographic reference is made to the original source
- a link is made to the metadata record in the Repository
- the full-text is not changed in any way

The full-text must not be sold in any format or medium without the formal permission of the copyright holders.

Serviços de Informação e Documentação, Instituto Universitário de Lisboa (ISCTE-IUL)
Av. das Forças Armadas, Edifício II, 1649-026 Lisboa Portugal
Phone: +(351) 217 903 024 | e-mail: administrador.repositorio@iscte-iul.pt
https://repositorio.iscte-iul.pt

Machine learning models to predict the COVID-19 reproduction rate: combining non-pharmaceutical interventions with sociodemographic and cultural characteristics

Margarida Duarte[1], Catarina Ferreira da Silva[2,3], Sérgio Moro[2,4]

[1] Instituto Universitário de Lisboa (ISCTE-IUL), 40 Avenida das Forças Armadas, 1649-026 Lisboa, Portugal

[2] Instituto Universitário de Lisboa (ISCTE-IUL), ISTAR, Lisboa, Portugal

[3] Center for Informatics and Systems of the University of Coimbra (CISUC), Portugal

[4] University of Jordan

Abstract

Since the beginning of the COVID-19 pandemic, countries worldwide have implemented a set of Non-Pharmaceutical Interventions (NPIs) to prevent the dissemination of the pandemic. Few studies applied machine learning models to compare the use of NPIs, socioeconomic and demographic characteristics, and cultural dimensions in predicting the reproduction rate R_t. We adopted the CRISP-DM methodology using as data sources the "Our World in Data COVID-19", the "Oxford COVID-19 Government Response Tracker" and the Hofstede Insights data. We analysed the impact that the Hofstede's cultural dimensions, the implementation of various degrees of restriction of NPIs and the sociodemographic variables may have in the reproduction rate by applying machine learning models to understand whether cultural characteristics are useful information to improve reproduction rate predictions.

We included data from 101 countries to train several machine learning models to compare the results between the models with and without the Hofstede's cultural dimensions. Our results show the use of cultural dimensions helps to improve the models, and that the ones that obtained a better prediction of the R_t were the ensemble models, especially the Random Forest.

Keywords: COVID-19, NPIs, Non-Pharmaceutical Interventions, Hofstede's cultural dimensions, Machine Learning, reproduction rate R_t

1. Introduction

At the end of 2019 in China, the first cases of infection by coronavirus 2019 (COVID-19) (1,2) were identified, generating unprecedent socioeconomic disturbances on a global scale throughout the following months (3,4). On the 30th of January of 2020, the World Health Organization (WHO) declared the COVID-19 as a Public Health Emergency of International Concern (1,3,5). Then, the world faced a global health crisis in which the development of new variants of the coronavirus that cause COVID-19 further aggravates the infection rate (6).

In this context, mathematicians developed a measure to calculate the number of effective reproduction also known as the R_t which "is defined by the expected number of secondary cases arising from a primary case infected at time t" (7), i.e., to measure the rate of transmission of an infectious COVID virus in a population (8). This new method estimates the R_t in real time, which constitutes an important measure in infectious COVID epidemiology by reflecting several factors such as control measures, contact rates and climatic conditions (9).

Meanwhile, non-pharmaceutical interventions (NPIs) measures have been implemented by countries worldwide with the aim of lowering new cases and new fatalities (2). NPIs refer to a wide range of top-down measures which are promoted by governments and bottom-up measures that refer to each citizen self-initiated behaviours who aim to interrupt the chains of infection by changing key aspects of human behaviour (10). In our study, we look at NPIs from the top-down point of view which corresponds to the implementation of NPIs established by the governments for each country throughout the COVID-19 pandemic months. By identifying the most effective NPIs, it is possible for stakeholders to implement a selection of key measures in a timely manner to fight COVID-19 or other future pandemics (11,12). In addition to the impact of the NPIs, considering socioeconomic and cultural factors can be an asset to identify vulnerable cultural minority groups that become a hotbed of infection by COVID-19 in order to address their needs and then also to mitigate the spread of the virus (13,14). Tang and colleagues (2022) (15) provide an exploratory analysis of the political, social, economic and cultural factors affecting how different global regions initially reacted to the COVID-19 pandemic. These authors argue that both government as well as population-related factors can contribute to success or failure in managing a pandemic. Particularly, the level of government trust by their country populations and of pandemic tiered preparedness have influence in a pandemic management outcome.

Several factors can affect the risk of transmission of COVID-19 including demographic, economic, and environmental characteristics (1,2,16-22). In addition to these factors the underlying culture of each society can also influence the impact of NPIs since, it can influence the collective responses of a group of individuals in their environment (23). To date very few studies (24, 25) applied machine learning models to compare the use of NPIs, socioeconomic and demographic characteristics, and cultural dimensions in predicting the reproduction rate. Zehtabian et al. (25) apply and compare four learning-based models, using 12 NPIs for predicting the average of 7-day daily new cases in a particular area. These authors also use Hofstede's cultural dimensions and argue that the severity and implementation of the NPIs measures were determined not just by the progress of the pandemic, but also by regional political developments, prevailing attitudes, and influential individuals shaping public discourse. They also defend that including information about the culture of the modelled region, along with traditional mathematical models and modern deep learning architectures, can enhance the accuracy of short-term predictions. Yeung et al. (2021) (24) used machine learning techniques with the goal of predicting COVID-19 confirmed infection growth (CIG) over 14 days considering NPIs and Hofstede's cultural dimensions. This work by Yeung and colleagues in 2021 showed that both Random Forest and AdaBoost regression techniques could potentially forecast the 14-day growth rate of confirmed COVID-19 infections. The novelty of our work is to predict the R_t, that reflects the anticipated quantity of subsequent infections stemming from an initial case that became infected at a specific point in time t, instead of the CIG, and combine R_t with cultural dimensions, socioeconomic and demographic variables and, also the restriction degree of NPIs. In this sense, the research question we propose to answer with our research is the following: *Can cultural characteristics be useful information for predicting the R_t?* with the objective to analyse the influence of cultural dimensions to improve the R_t evolution predictions. In this study we use data from 101 countries from March 2020 to May 2021 for testing several machine learning models with and without the Hofstede's cultural dimensions (26) so that we can compare the results. This study contributes to existing literature by measuring how the cultural context inherent to each country affected the way citizens reacted to the havoc caused by the pandemic spread worldwide throughout 2020 and 2021 relate to R_t. By including each country's cultural dimensions together with the already studied features related to NPIs and social and demographic indicators, the models developed in our study

contribute to a better understanding if a country's citizens collective culture played a role in the pandemic dissemination.

The remainder of the paper is organised as follows: firstly, a literature review is presented, then the methodology we used is explained, followed by the results and discussion. Finally, we present our conclusions, which include the limitations and future work.

2. Related work: NPIs and culture dimensions impact in COVID-19

Guan et al., (2020) (27) show collective culture-oriented actions and norms in response to the pandemic of COVID-19 will serve as a top-down influence on the individual behaviours of its members, and governments for each country must carefully balance collective and individual interests as well as effectively manage the trade-offs of social control and individual freedom to achieve the best outcomes for the public. Furthermore, according to Silva et al., (2021) (28) the social and clinical experience with COVID-19 will leave lasting marks on society and in the healthcare systems, from Latino cultural habits (e.g. proximity, greeting by kissing) to changes in the configuration of the healthcare systems, leaving hospitals for more complex clinical situations and providing healthcare institutions with diagnostic and therapeutic means that avoid systematic recourse to hospital emergencies

Hoepner et al. (2023) (29) statistically analyse the impact of NPIs on the effective reproductive rate of COVID-19 and found out that stay-at-home orders and other extreme NPIs appear to have only marginal effectiveness, while financially motivated measures such as debt relief and income support show a greater degree of effectiveness. They also point out that the effectiveness of NPIs, which typically aim to limit and control population movement, may also be influenced by cultural, societal, and economic factors in a given region.

Culture refers to a system of values that are shared and the collective practices that distinguish one society from another (26). Hofstede (2011) (30) presents six dimensions that represent a country's culture:

1. **Power Distance**, referring to the different solutions to the basic problem of human

inequality. Citizens in countries showing a large power distance seem to accept more easily a hierarchical order in which everybody has its place, and which requires no further justification, while in countries exhibiting low power distance, societal forces are under strain to try to balance the distribution of power and plead justification for inequalities of power;

2. **Uncertainty Avoidance**, related to the level of stress in a society in the face of an unknown future;
3. **Individualism versus Collectivism**, with the former denoting societies where each individual is supposed to lookup for himself/herself and direct family only, versus collectivist societies where people integrate into larger groups that are supposed to take care of the whole group, to which their members are loyal;
4. **Masculinity versus Femininity**, where the former indicates a society that is driven by success, rather than by values related to attaining a sustainable quality of life;
5. **Long Term versus Short Term Orientation**, relating to the choice of focus for people's efforts: the future or the present and past;
6. **Indulgence versus Restraint**, related to the gratification versus control of basic human desires related to enjoying life.

Lu et al., (2021) (31) state that culture fundamentally shapes how people respond to crises such as the COVID-19 pandemic and plays an important role in mask usage. According to Gelfand et al., (2021) (32), people in loose cultures, which have weaker norms and are much more permissive, had less fear of COVID-19 than people in tight cultures, which have stricter rules and punishments for deviance.

Erman & Medeiros (2021) (33) emphasise that during the initial period of the pandemic policy makers should more explicitly consider the cultural attributes of a society along with other important parameters such as demographic characteristics and healthcare system constraints in order to develop more tailored and effective policy responses.

Some studies already included Hofstede's cultural dimensions along with NPIs to study their impact on the pandemic. In particular, they have focused on data from the first wave of the pandemic. Wang (2021) (34) combined government policies and national culture to explain their influence on social distancing which showed that, the stringency index (a composite measure based on nine response indicators such as school closures, workplace closures, and travel bans, rescaled to a value from 0 to 100, where 100 represents the strictest response), and Hofstede's cultural dimensions long-term orientation and indulgence significantly and negatively affect social distancing (35). On the other hand, in Gokmen et al., (2021) (36) national cultural dimensions of 31 European countries were related to the rise in the total number of COVID-19 cases per million and obtained that power distance has a significant and negative effect while, individualism and indulgence have a positive and significant effect. The later result diverges from the previous study in which indulgence decreased social distance.

Guss & Tuason (2021) (37) show that the higher the GDP per capita, the higher the number of COVID-19 deaths, and also that countries with high individualistic values and high intellectual autonomy were associated with a higher number of COVID-19 deaths, while countries with higher collectivistic values were associated with fewer COVID-19 deaths.

Luu & Huynh (2020) (38) studied the impact of the cultural factor on complying with social distancing measures, obtaining that countries with high uncertainty avoidance prognosticate the lowest proportion of people gathering in public (e.g., retail and recreation, grocery and pharmacy, parks, transit stations, workplaces).

Although previous studies (e.g., Wang (2021) (34)) indicate that culture impacts NPIs, governmental strictness has a higher impact on social distancing than national culture since people may drop their guard due to exhaustion with social distancing rules after living for a long time under the COVID-19 pandemic restrictions.

Zehtabian et al., (2022) (25) tested four machine learning models to predict the number of people infected on a given day of the pandemic in a given area. For this, in one of the models tested Hofstede's cultural dimensions were used as input, however, the authors state as a limitation that there was no method to rigorously incorporate the cultural dimensions.

The work of Xie et al., (2022) (14) considers internal characteristics (economic and cultural characteristics) and external variables (government interventions) to train a machine learning back propagation (BP) neural network model with the objective of creating a model to predict risk perception of the Chinese population towards an event and for this, the COVID-19 pandemic is used as one of the events.

Table 1 shows the studies found during our literature analysis in which the authors consider cultural dimensions to understand their impact on the evolution of the pandemic and the implementation of the NPIs. The period of the data to which the study refers is clearly identified, when it is provided by the corresponding publication (second column of Table 1). The studies available focus on the early period of the pandemic and include different COVID-19 dependent variables such as total COVID-19 cases per million, the growth of confirmed COVID-19 infections, and the number of COVID-19 deaths. They mostly applied regressions' analysis with the aim of understanding the impact of culture on the evolution of COVID-19. Zehtabian et al., (2022) (25) combine the Oxford Covid-19 Government Response Tracker (OxCGRT) data on NPIs with the Hofstede cultural dimensions to predict the average of 7-day daily new cases, using LSTM (long short-term memory) and transformer encoder multihead attention (TEMA). However, as the authors state, their results were limited due to not being able to use the cultural dimensions rigorously and the sophisticated machine learning TEMA models are not making a decisive difference in prediction accuracy.

Table 1: Previous studies that consider cultural dimensions when analysing COVID-19 data

Reference	Period	Countries	Objective	Datasets	Features types used	Method	Results
Gelfand et al., (2021) (32)	Beginning of the pandemic to Oct 16th, 2020	57 countries	Examine the association between the strength of social norms and the number of cases and deaths in COVID-19	• Our World in Data • OxCGRT • Hofstede Insights	• COVID-19 measures • Socioeconomic measures • Hofstede dimensions	Ordinary least squares regression	Tight groups cooperate much faster under threat and have higher survival rates than loose groups
Erman & Medeiros (2021) (33)	First wave of the pandemic	73 countries	Explore the global variability of public health outcomes during the first wave of the pandemic	• WHO Global Health Observatory Data Repository • Hofstede Insights	• Socio-demographics • Health system capacity • COVID-19 measures • Cultural dimensions • Political dimensions	Random-effects meta-analysis models	Cultural attributes do account for some of the global disparities in COVID-19 attributed health outcomes
Gokmen et al., (2021) (36)	N/A	31 European countries	Investigate the impact of national cultural dimensions on the increasing rate of the total COVID-19 cases per million (IRTCCPM)	• Our World in Data; • VHD European countries; • Hofstede Insights	• COVID-19 measures: cases; • Hofstede dimensions: power distance and collectivism	Explanatory modelling	Power distance has a significant and negative effect on IRTCCPM whereas, individualism and indulgence have significant and positive effects on IRTCCPM
Wang (2021) (34)	First wave of the pandemic	80 countries	Assess the impact of national culture and government policies on social distancing to fight COVID-19	• OxCGRT • Hofstede Insights • Google mobility reports	• COVID-19 measures • Hofstede culture dimensions • Google Community Mobility Reports • Socio-economic measures	Linear regression	Social distancing decreases with long-term orientation and the opposite is true for indulgence
Yeung et al., (2021) (24)	April 1st to September 30th, 2020	114 countries	Predict the national growth of confirmed infections growth in COVID-19 (CIG) across 14 days	• OxCGRT • Hofstede Insights • Johns Hopkins University dataset	• NPIs measures • Current infection numbers • Hofstede cultural dimensions	Ridge, decision tree, random forest, AdaBoost and SVR	Non–time series machine learning models can predict future CIG to an appreciable degree of accuracy
Luu & Huynh (2020) (38)	February 16th to March 29th 2020	58 countries	Examine the role of the cultural dimensions in practising social distancing across the world	• Google • Hofstede Insights	• Social distancing taking place in various locations • Hofstede cultural dimensions	Ordinary Least Squares regression	Countries with higher uncertainty avoidance index predict the lower proportion of people gathering in public
Güss & Tuason (2021) (37)	July 14th to December 29th, 2020	75-76 countries	Study culture values impact in deaths by COVID-19	• Johns Hopkins University dataset • Hofstede Insights • World Health Organization dataset	• COVID-19 measures • Socioeconomic and demographical measures • Hofstede cultural dimensions	Regression analyses	Countries with high GDP per capita, high individualistic values and high intellectual autonomy were significantly and consistently associated with high numbers of COVID-19 deaths

Zehtabian et al., (2022) (25)	January 2020 to end of December 2020	N/A	Test four alternative learning-based models (using LSTM and TEMA) to predict the number of infected people in a given day of the pandemic in a particular area	• OxCGRT • Hofstede Insights • XPrize Pandemic Response Challenge dataset	• NPIs measures • Hofstede cultural dimensions • Geographical region	Long short-term memory networks (LSTMN) Susceptible-infectious-removed model	The short-term forecasts have been improved with the addition of the cultural component in the models tested by the authors nevertheless, the accuracy of the medium-term forecasts is still very low

3. Methodology

The methodology we used is the CRoss-Industry Standard Process for Data Mining (CRISP-DM). The CRISP-DM methodology defines a non-rigid sequence of six phases, which allow the construction and implementation of a Data Mining model to be used in a real environment (39). The Figure 1 summarizes the CRISP-DM methodology we applied in our study.

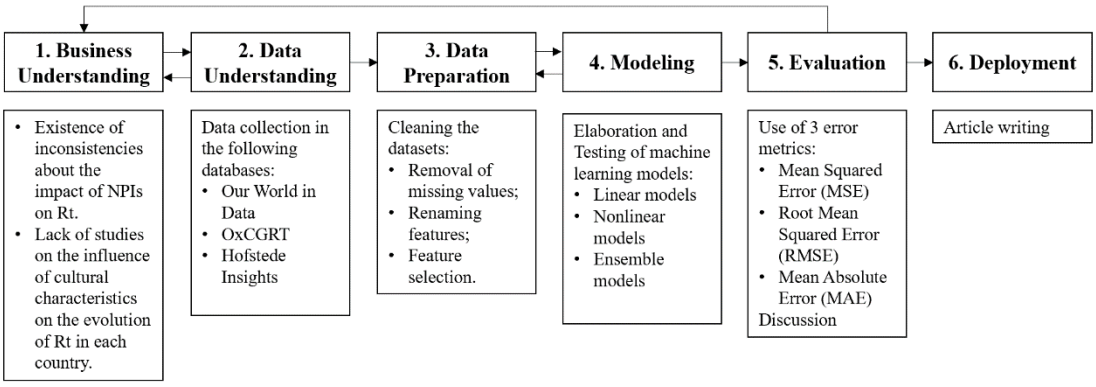


Figure 1: CRISP-DM methodology

Initially, we started by identifying a problem recognised during the literature review in the scope of the impact of the NPIs. In countries considering their cultural differences more specifically, throughout our literature review to date we found no studies that apply machine learning models to compare the use of NPIs, socioeconomic and demographic characteristics, and cultural dimensions in predicting the reproduction rate (R_t). Existing studies, and in particular Yeung et al., (2021) (24), consider the impact of NPIs and cultural dimensions with COVID-19 cases, deaths and CIG_{14} (24).

In the business understanding phase, we started by collecting the data at its sources where we gathered daily data of COVID-19 and NPIs implemented considering their degree of restriction throughout the pandemic from March 2020 to May 2021. Data on cultural dimensions were also considered to understand the cultural differences of the countries. The data related to COVID-19 was obtained from the Our World in Data COVID-19 dataset[1].

[1] https://ourworldindata.org/coronavirus-source-data

NPIs data for COVID-19 was obtained using the Oxford COVID-19 Government Response Tracker (OxCGRT)[2]. From this dataset NPI related information of COVID-19 was collected. The NPIs' policy values are ordinal measures that associated values are integers between 0 and 5, which represent the restriction degree of the given measure.

The cultural characteristics' data were obtained from the Hofstede Insights[3] website which includes a score ranging from 0 to 100 for each of the six cultural dimensions: power distance, individualism, masculinity, uncertainty avoidance, long term orientation and indulgency.

After extracting and loading the data, we proceed to its preparation by cleaning its null values and removing variables which are not relevant to this study. As so, the dataset was refined by eliminating redundant and uninformative features. This included removing country codes, total and cumulative values, and variables with significant missing data. Additionally, features like gender-specific smoking rates, extreme poverty indicators, diabetes prevalence and cardiovascular mortality rates were excluded due to very limited data availability for these features. Our approach to cleaning the Rt dataset establishes a baseline while preserving the unique characteristics of each country.

In the Our World in Data dataset we removed a total of 42 features (Table 2) such as the ISO code because the name of the country is the variable that identifies the COVID-19 data origin and is the feature used to merge the datasets. We decided to remove features that represented total, rounded and weekly values in order to allow for cross-country comparisons. We also decided to rename the feature location to country to standardize this feature for later merging the datasets. Additionally, there were R_t negative values which should be positive so we replaced with their absolute value. Regarding the missing values, we decided to remove the rows that had no value associated such as population, population density, median age, people aged 65 or older, GDP per capita, hospital beds per thousand, life expectancy, human development index and stringency index. In relation to the reproduction rate variable, the initial values are null because we assumed there was no data for each country yet, so we replaced them with 0 until there was registered data, whereas after the first appearance of a value the next null values that appear are replaced by the average R_t value for that country.

Additionally, we have removed the global scale values from the dataset since we want to do an analysis per country.

[2] https://github.com/OxCGRT/covid-policy-tracker/blob/master/documentation/interpretation_guide.md
[3] https://www.hofstede-insights.com/country-comparison/

The tools chosen to perform these activities were Python[4] as the programming language and the Jupyter Notebook[5] platform.

Table 2: Cleaning Our World in Data Dataset

Cleaning Our World in Data Dataset	Features
Removed features	A total of 42 features – e.g., iso_code, tests_units
Renamed features	location → country
Negative values	Replaced negative values with abs()
Missing values	Replaced initial missing values with 0 and after the first non-null value replace them with the mean of those values: • new_cases_per_million • new_deaths_per_million • reproduction_rate • icu_patients_per_million • hosp_patients_per_million Removed countries where socioeconomic and demographical features are missing: • population • population_density • median_age • aged_65_older • gdp_per_capita • life_expectancy • human_development_index

From the Oxford dataset (Table 3), we removed features that were of no interest for the analysis such as the country code, flag and initial note. The "CountryName" feature was renamed to country for the purpose of later merging the datasets. In this dataset there were missing values in the "end date" and in the "policy value". In this sense, after analysing the data, the lines that had the NPI end date ("EndDate") set to null refer to NPIs that were implemented and still in force and, therefore, these missing values were replaced by the date of the dataset extraction. Regarding the "policy values" with no value we decided to remove them as they would not be useful data for our analysis.

Table 3: Cleaning Oxford Dataset

[4] https://www.python.org/
[5] https://jupyter.org/

Cleaning Oxford Dataset	Features
Removed features	CountryCode, Flag, InitialNote
Renamed features	CountryName → country
Missing values	Removing PolicyValue missing values EndDate filled with the current date

In the Hofstede Insights dataset (Table 4) we removed some features with metadata information such as the "adjective" and the "slug" as well as features with useless information to our work such as the "id" and "ivr". We also decided to rename the "title" to "country" and the cultural dimensions by their long name instead of their short name.

Table 4: Cleaning Culture Dataset

Cleaning the Culture Dataset	Features
Removed features	id, slug, adjective, ivr
Renamed features	title : country, pdi : power_distance, idv : individualism, mas : masculinity, uai : uncertainty_avoidance, lto : long_term_orientation, ind : indulgence
Removed rows with missing values	Qatar and Tunisia

After cleaning and combining the data we obtained a dataset of 101 countries with data between March 2020 and May 2021.

This dataset is composed of variables values related to COVID-19 and the NPIs in which each row represents the variable's value per day. It also includes the variables related to the socioeconomic and demographic characteristics, and as well the Hofstede's cultural dimensions in which each row represents the cultural values for that specific country. We can observe these variables in Table 5.

Table 5: The characterization of the datasets with their final features

Features Type	Features
COVID-19	Reproduction rate (R_t)

Socioeconomic and demographic	Population Population density Median age Aged 65 older GDP per capita Life expectancy Human development index
NPIs	Stringency index Testing policy School closing International travel controls Public information campaigns Protection of elderly people Contact tracing Workplace closing Stay at home requirements Cancel public events Restrictions on gatherings Restrictions on internal movement Debt/contract relief Income support Facial Coverings Vaccination policy Close public transport
Hofstede's cultural dimensions	Power distance Individualism Masculinity Uncertainty avoidance Long term orientation Indulgence

We used the data from all the 101 included countries of our dataset to train several machine learning models so that we could verify the impact of cultural dimensions in the prediction of the reproduction rate.

Through these models we aim to assess whether there is an improvement of the models with the addition of Hofstede's cultural characteristics in predicting the R_t. For data normalization we used the StandardScaler procedure which is one of the most used techniques in machine learning in which the features are normally distributed with a mean value close to 0 and a standard deviation of 1 (40). After normalizing the data, we selected the features for the machine learning models from the dataset we obtained after the data preparation phase. Thus, we tested two data selection methods: correlation and mutual information (MI). Through correlation analysis and MI results, we validated the relevance of the features, provided insights into variable interactions, and reduced dataset noise.

Figure 2 shows the correlations of all independent variables with the R_t. We can observe that one of the most relevant correlations with the reproduction rate we obtained was a weak negative correlation with the facial coverings NPI (-0.16). We also obtained other correlations close to 0.10 and -0.10 such as, human development index (0.09), vaccination policy (-0.09), new deaths per million (-0.08), median age (0.08), life expectancy (0.08) and uncertainty avoidance (0.07).

Although individually each of these variables did not obtain a relevant linear correlation with the target R_t, this does not mean that all these variables are not important to train the predictive models. For this reason, it was decided to explore nonlinear correlations using MI.

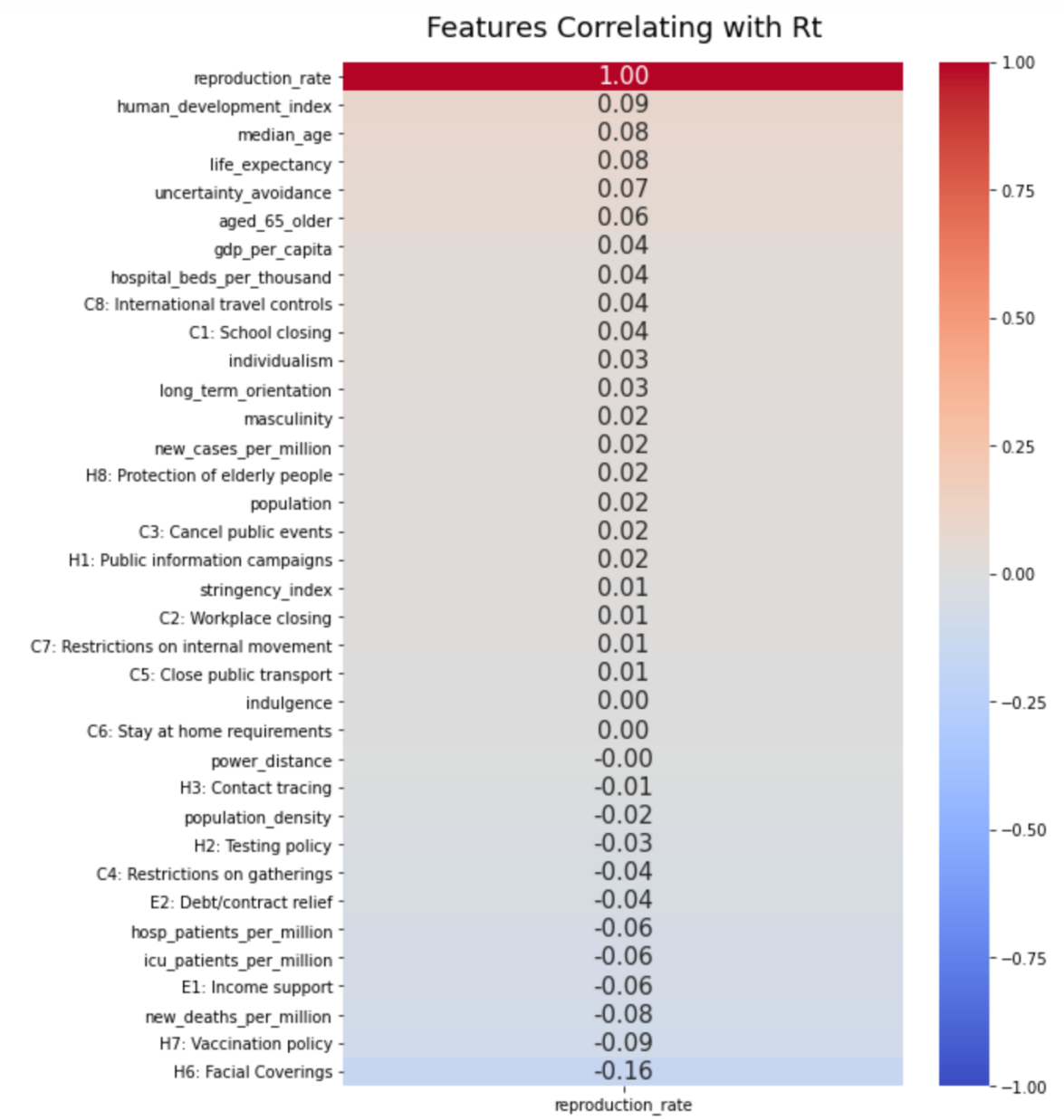


Figure 2: Features correlation with R_t

MI is used as a criterion to select the best input variables (from a set of possible variables) for the long-term forecasting objective, in other words, between two variables, say X and Y, it refers to the amount of information obtained from X in the presence of Y, and vice versa (41).

This dependency measure can capture linear as well as nonlinear dependencies but it does not capture multivariate dependencies or indicate collinearity between features (24). Accordingly, we use the Python package ennemi[6] based on MI (42). In Table 6 we can observe that all features obtained a MI score above 0.10. The features that obtained the highest correlation were the socioeconomic and demographic ones such as population size, GDP per capita, portion of the elderly population (older than or equal to 65 years), population density, average life expectancy, human development index and average age with values ranging from 0.25 to 0.29. Then the features that obtained the highest correlation were the cultural dimensions of long-term orientation, indulgence, power distance, individualism, masculinity, and uncertainty avoidance with values between 0.18 and 0.21. Finally, NPIs such as close public transport, facial coverings, income support and the restriction index obtained correlations between 0.10 and 0.27.

Through the analysis of these results, we can argue that in addition to the socioeconomic and demographic dimensions, the Hofstede's cultural dimensions are relevant to help predict the R_t value. Regarding the NPIs, these values can be explained due to the volatile nature of their values, as there are countries that rarely change them (e.g., Brazil) as opposed to countries that do it regularly (e.g., Portugal).

Yeung et al., (2021) (24) state that all features are important to be considered in the machine learning models because features are selected if they achieve MI greater than 0.10. Therefore, we consider all NPIs, socioeconomic and demographical features and Hofstede's cultural dimensions.

Table 6: Mutual Information score

Feature	Score
Stringency index	0.27
Population	0.29
Population density	0.28
Median age	0.26
Aged 65 older	0.29
GDP per capita	0.29
Hospital beds per thousand	0.25
Life expectancy	0.28
Human development index	0.27
H2: Testing policy	0.12
C1: School closing	0.11
C8: International travel controls	0.12
H1: Public information campaigns	0.17

[6] https://github.com/polsys/ennemi

H8: Protection of elderly people	0.10
H3: Contact tracing	0.13
C2: Workplace closing	0.13
C6: Stay at home requirements	0.11
C3: Cancel public events	0.14
C4: Restrictions on gatherings	0.11
C7: Restrictions on internal movement	0.10
E2: Debt/contract relief	0.11
E1: Income support	0.12
H6: Facial Coverings	0.15
H7: Vaccination policy	0.15
C5: Close public transport	0.11
Power distance	0.18
Individualism	0.19
Masculinity	0.18
Uncertainty avoidance	0.18
Long term orientation	0.21
Indulgence	0.21

In order to evaluate the applied models, we apply three error metrics (43):

- The **Mean Absolute Error or MAE** measures the average of the sum of absolute differences between observation values and predicted values.
- The **Mean Squared Error or MSE** measures the average squared difference between the estimated and true values.
- The **Root Mean Squared Error or RMSE** is a monotonic transformation of the MSE that corresponds to its root.

4. Results

We tested several machine learning models to compare the results of data from 101 countries where each row contains the value of the R_t per day in a particular country. Table 6 shows all the features used for the models. We used the following tree types of models for predicting the R_t: linear models, nonlinear models, and ensemble models.

Due to their simplicity and low complexity, linear regression models are among the most fundamental and widely used techniques to model and predict the behaviour of massive data flows (44). The linear regression models we applied are the Linear Regression, Least Absolute Shrinkage and Selection Operator (LASSO), Ridge Regression, Elastic Net, Huber Regression, Least-Angle Regression (LARS), Passive Aggressive Regressor, and the Stochastic Gradient Descent (SGD) Regressor.

In general, nonlinear regression models are similar to linear regression models, where an important difference from nonlinear regression models is that the number of regression parameters is not necessarily directly related to the number of independent variables in the model (45).

In this study, we apply four nonlinear models: K Nearest Neighbours Regression (KNN), Decision Tree Regression, Extra-Tree Regression and Support Vector Regression (SVR).

Ensemble models aim to combine several models to improve the prediction accuracy in learning problems with a numerical target variable (46). For example, a Random Forest Regressor is an ensemble of Decision Trees. The ensemble models we used are the following: AdaBoost, Random Forest, Extra-Trees Regressor, Gradient Boosting and Extreme Gradient Boosting (XGBoost).

Table 7 shows that the results of the models improved with the addition of Hofstede's cultural dimensions. Comparing the results of all the models we can state that the linear models are the ones with the highest errors. This is explained by the fact that the data does not follow a linear trend, which can be seen through the correlations obtained in the heatmap (Figure 2).

Although the linear models were those with the worst results, the ones that applied the Huber regression obtained the lowest MAE error of 0.22292 with the use of the Hofstede's six cultural dimensions and an error of 0.22379 without them. Linear and Ridge regression both showed the lowest MSE of 0.11174 with the cultural dimensions and an error of 0.11329 without them.

Also, previous models both showed the lowest RMSE of 0.33427 with the cultural dimensions and 0.33659 without the cultural dimensions. In the nonlinear models the error values obtained are significantly better, which confirms the nonlinearity of the data, justifying the use of MI for the feature selection stage.

Comparing the nonlinear models, the Decision Tree regressor was the method that obtained the best results for the three metrics: a MAE of 0.08628 with the cultural dimensions and an error of 0.08603 without them; a MSE of 0.02122 with the cultural dimensions and an error of 0.02127 without them; and a RMSE of 0.14569 with the cultural dimensions and an error of 0.14584 without them.

Finally, we tested the ensemble models which performed even better compared to the nonlinear models, as expected. The model with the lowest error was the Extra-Trees regressor with a MAE of 0.08596, a MSE of 0.02088 and a RMSE of 0.14451 without considering the use of the cultural dimensions. These results further improve with the use of the cultural dimensions, being the Random Forest regressor the best model with a MAE of 0.08575, a MSE of 0.02077 and a RMSE of 0.14412.

Table 7: Machine learning models error results, with and without the cultural dimensions

Error / Predictive Model	Without the cultural dimensions			With the cultural dimensions		
	MAE	MSE	RMSE	MAE	MSE	RMSE
Linear Models						
Linear Regression	0.22795	0.11329	0.33659	0.22758	0.11174	0.33427
LASSO	0.22903	0.12197	0.34924	0.22903	0.12197	0.34924
Ridge	0.22795	0.11329	0.33659	0.22758	0.11174	0.33427
Elastic Net	0.22903	0.12197	0.34924	0.22903	0.12197	0.34924
Huber Regression *	0.22379	0.11693	0.34195	0.22292	0.11554	0.33992
LASSO LARS	0.22903	0.12197	0.34924	0.22903	0.12197	0.34924
Passive Aggressive Regressor	0.34407	0.22909	0.47863	0.29025	0.16326	0.40405
SGD Regressor	0.22842	0.11414	0.33785	0.22773	0.11261	0.33557
Non-linear Models						
KNN Regressor	0.09761	0.02753	0.16593	0.09643	0.02700	0.16433
Decision Tree Regressor	0.08603	0.02127	0.14584	0.08628	0.02122	0.14569
Extra-Tree Regressor	0.08673	0.02175	0.14748	0.08642	0.02261	0.14931
SVR	0.12457	0.03815	0.19532	0.12099	0.03511	0.18738
Ensemble Models						
AdaBoost Regressor	0.29101	0.13456	0.36682	0.27001	0.11759	0.34281
Random Forest Regressor	0.08659	0.02101	0.14495	0.08575	0.02077	0.14412
Extra-Trees Regressor	0.08596	0.02088	0.14451	0.08587	0.02084	0.14437
Gradient Boosting Regressor	0.18593	0.07183	0.26802	0.18566	0.07004	0.26466
XGBoost Regressor	0.10535	0.02552	0.15976	0.10443	0.02504	0.15824

To optimize the three best ensemble models, their hyperparameters were adjusted using the GridSearch tool. Table 8 shows the best hyperparameters used for the Extra Trees and Random Forest.

Table 8: Hyperparameter tuning for the three best models

Models	Hyperparameters
Extra Trees Regressor	Estimators: 50 Maximum depth: 100 Bootstrap: False Minimum samples split: 2 Maximum features: auto Minimum samples leaf: 1
Random Forest Regressor	Estimators: 200 Maximum depth: 100 Bootstrap: False Minimum samples split: 2 Maximum features: sqrt Minimum samples leaf: 1

In general, the regression models obtained lower R_t prediction errors with the addition of the features of the six cultural dimensions along with the features of the socioeconomic, demographic, and various degrees of restriction levels of the NPIs. Nevertheless, the improvements are only slightly incremental. One of the facts that may have contributed to hinder the expected impact of the cultural dimensions is the heterogeneity of the countries among themselves combined with the false homogeneity of characterizing very large countries with the same measures for the six cultural dimensions, as larger and more populous countries tend to have larger cultural differences inside their borders.

By testing the best parameters of each of the three best ensemble models we concluded that the Random Forest regressor obtained the lowest error. Thus, this was selected to predict R_t for five countries that are very different from each other culturally, socioeconomically and demographically: India, Brazil, the UK, Israel and Portugal. Table 9 shows the results obtained for predicting the R_t for each of these countries. It can be observed that India was the country that obtained the best result with a MAE of 0.02637, MSE of 0.00163 and a RMSE of 0.04046, whereas the UK was the worst performing country with a MAE of 0.09994, a MSE of 0.02941and a RMSE of 0.17151. While it is not possible to concrete identify the reason for this performance discrepancy, it is interesting to note that the countries with higher "power distance" (dimension), India and Brazil, present a better performance for predicting the R_t. One can hypothesize that countries characterized by a lower "power distance" may be underrepresented.

Table 9: Results of the errors for the random forest regressor regarding the data of the five considered countries

Errors / Countries	MAE	MSE	RSME
India	0.02637	0.00163	0.04046
Brazil	0.03083	0.00225	0.04749
UK	0.09994	0.02941	0.17151
Israel	0.07953	0.01677	0.12951
Portugal	0.07347	0.01142	0.10687

5. Discussion

According to Yeung et al., (2021) (24) machine learning models can predict the Confirmed Infection Growth in 14 days (CIG_{14}) with a significant level of accuracy. In their study, the authors aimed to predict CIG_{14} using NPIs and Hofstede's cultural dimensions, for which they tested five regression machine learning models: Ridge, Decision Tree, Random Forest, AdaBoost and Support Vector.

Similarly, in our study we use machine learning models to predict the R_t, however, we consider different NPIs than the ones chosen on the above-mentioned study.

Yeung et al., (2021) (24) consider 13 ordinal policies and 4 computerized indices namely, Stringency Index, Government Response Index, Containment Health Index and Economic Support Index. Whereas, in our study we consider 16 ordinal policies as we also considered the NPIs testing policy, international travel controls, public information campaigns, protection of elderly people, contact tracing, workplace closing, debt/contract relief, facial coverings and vaccination policy. Additionally, we also deem socioeconomic and demographic features, namely population, population density, median age, population aged 65 older, GDP per capita, life expectancy, and human development index. From our results, we can see that the use of the Hofstede's cultural dimensions improved the models when compared to those without the cultural features.

The work of Zehtabian et al., (2022) (25) corroborates these findings as they state a lower cumulative 7-day MAE for their TEMA approach including the Hofstede's cultural dimensions, when compared with the other models the authors applied, namely LSTM.

Regarding the evaluation metrics in Yeung et al., (2021) (24), MAE and median percent error were used while, we adopt the MAE, MSE and RMSE. Also, in the study of Yeung et al., (2021) (24) three validation methods were used (in-distribution, out-of-distribution and cross-validation), whereas, we applied the hold-out technique.

In the two abovementioned studies, the ensemble models were the ones that obtained the best results because these are the combination of several predictions coming from a set of models to make the final prediction through averaging and voting so that the ensemble model provides better results than any of the individual models (47). The Random Forest model is the one that obtain better results in both, being that in Yeung et al., (2021) (24) the authors obtained a MAE of 0.031 for predicting the CIG_{14} and in our study we obtained a MAE of 0.085 for predicting the R_t. The study of Zehtabian et al., (2022) (25) applies the hold-out technique on the OxCGRT data considering the following 12 NPIs: school closing, workplace closing, cancel public events, restrictions on gatherings, close public transport, stay at home requirements, restrictions on internal movement, international travel controls, public information campaigns, testing policy, contact tracing, and facial coverings. The Zehtabian et al., (2022) (25) study tests the following models: LSTM as developed by the Pandemic Response Challenge ("LSTM-UT-Cogn" (48)), a simple LSTM network ("LSTM-Baseline"), LSTM with Hofstede cultural dimensions ("LSTM-CultD-SIR") and TEMA with Hofstede cultural dimensions ("TRANSENC-CultD-SIR"). This last one, the TEMA "TRANSENC-CultD-SIR", provides their lowest cumulative MAE regarding the prediction of average daily number of new cases per 100k people over a span of 7 days.

To better understand the impact of each feature and, particularly, the contribution of the proposed cultural dimensions, we computed the feature importance as extracted from the best performing model, the Random Forest (Figure 3). Even though MI emphasized the value of the additional information provided by the cultural dimensions (Table 6), we can observe that none is included in the top ten most relevant features. This strengthens the results show on Table 7, showing just a slightly incremental increase in the models' performance when the cultural dimensions are included, especially in comparison to NPI. However, one can hypothesize that the Hofstede's cultural dimensions are too broad and in fact may hinder their real impact, especially in medium-large countries with cultural contrasts in their boarders. Further studies are recommended, maybe more focused on regional populations and their cultural characteristics.

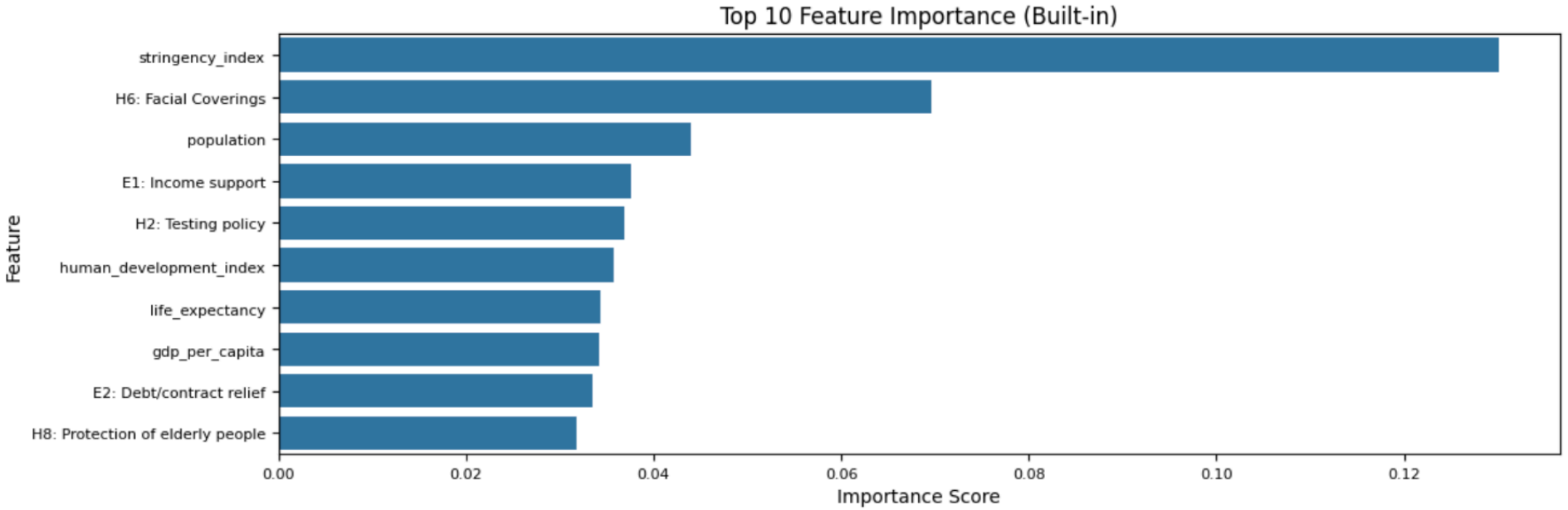


Figure 3: Feature importance

Table 10 shows a comparison between the study of Yeung et al., (2021) (24), of Zehtabian et al., (2022) (25) and our study to spotlight their main differences and similarities.

Table 10: Summary comparison of our study with the work of Yeung et al., (2021) (24) **and of Zehtabian et al., (2022)** (25)

	Yeung et al., (2021) (24)	**Zehtabian et al., (2022)**	**Our study**
Objective	Predict the national growth of confirmed COVID-19 infections across 14 days	Predict the average of 7-day daily new cases	Understand if cultural dimensions are useful data to improve the machine learning models regarding the reproduction rate (R_t) prediction
Countries	114	N/A	101
Target	Confirmed Infection Growth in 14 days (CIG_{14})	7-day daily new cases	R_t
Features	• 17 NPIs • 6 Hofstede cultural dimensions	• 12 NPIs • 6 Hofstede cultural dimensions	• 17 NPIs • 7 Socioeconomic and demographic characteristics • 6 Hofstede cultural dimensions
Evaluation metrics	• MAE • Median percent error	MAE	• MAE • MSE • RMSE
Validation schema	• In-distribution • Out-of-distribution • Cross-validation	Hold out	Hold out
Best models	• Random Forest • AdaBoost	TEMA, with 6 Hofstede cultural dimensions (TRANSENC-CultD-SIR)	• Random Forest • Extra Trees

From a practical perspective, we highlight that all the 8 linear models, 3 out of the 4 non-linear models, and all the 5 ensemble models trained obtained higher accuracy when including the cultural dimensions (as opposed to not include them). Although the differences are small, the results are consistent (16 out of 17 models achieved better performance). Additionally, the difference is more significant for the more complex ensemble models. Therefore, we can positively answer to our research question, i.e., cultural characteristics are useful information for predicting the Rt. We believe this is an important result, useful for helping manage future pandemics evolution, since it is not enough to implement NPIs to stop the dissemination, although their undisputed effectiveness, proved during COVID-19 spread; instead, for mathematical and computational prediction models to be more accurate, they need to encompass a wider array of features that include the characterization of the population, in a sense that each population's response is different, depending on the cultural and demographic context.

Our study differs from the other studies also shown in comparative Table 10 because it uses several contextual variables such as, cultural dimensions, socioeconomic and demographic variables and, also the restriction degree of NPIs to predict the Rt. Noteworthy, we did not find any other study that used these variables to predict Rt. Since, until now, Rt is a variable calculated by mathematical formulas, this study takes an innovative approach by applying machine learning models to test new variables to predict it.

From these results we recommend that policymakers should give greater consideration to society's cultural nuances, in addition to crucial factors like population demographics and healthcare infrastructure limitations, when crafting more customized and impactful policy solutions based on NPIs.

While national cultural dimensions may reflect the majority's values, the dataset's composition could skew representation. Potential imbalances in geographic, socioeconomic, and cultural diversity within samples might lead to over- or under-representation of certain regions. This sampling bias could explain the varying accuracy of predictions across countries, with more precise forecasts for Brazil and India compared to the UK, Israel, and Portugal. The model's effectiveness may depend on how well the sampled data captures each nation's true cultural diversity.

6. Limitations and future studies

Our study has a few relevant limitations that need to be highlighted. First, it is dependent upon the data, since this is a secondary data-based study. Specifically, any eventual lack of homogeneity by each country's authority in collecting the data is directly reflected in data quality, affecting overall models' performance. Additionally, the hold-out validation technique was used to split the data, and as it is the simplest splitting data technique, we propose for future studies the use of other validation techniques. Also, the future studies can test other machine learning models.

Another important limitation is the fact that the Hofstede's cultural dimensions are on a national (country-wide) basis, whereas there are many countries with cultural contrasts within their borders (e.g. the United States, Spain, Switzerland), thus country-level cultural dimensions are limited due generalizing for the whole country. To address this limitation, we propose as future research to focus on a reduced set of specific countries and conduct local surveys that may complement the results here presented by addressing specific regional and local contexts.

We recommend that further studies should be done, involving a finer grain of socioeconomic, cultural and demographic characterization of the population inside countries, so that one can then use these to apply and compare various set of predictive machine learning models with greater detail. A more refined comprehension of what characteristics can help prediction and what cannot would prove beneficial to future policymakers.

Finally, we need to highlight the limitations of current state-of-the-art machine learning models for pandemic prediction. For instance, the dynamic nature of pandemics and the emergence of new variants of the virus could lead to changing trends that might not be captured by models trained on historical data. Additionally, the complexity of human behavior and policy responses, which are influenced by many factors beyond NPIs and cultural dimensions, could introduce unpredictability in the models' performance over time. As we aim to predict a pandemic evolution with a higher granularity in time and population scopes, we face the challenges of computational limitations that hinder implementations of complex systems that address a larger number of dynamic variables in an uncertainty pandemic context (48).

7. Conclusions

Since the beginning of the pandemic, countries around the world have implemented various NPIs aimed at stopping the spread of COVID-19. We tested several machine learning models, with and without the Hofstede's six cultural dimensions, to compare and understand whether these improve the prediction of the reproduction rate R_t.

This study set out to answer the following research question: "Can cultural characteristics be useful information for predicting the R_t?" and to answer this question, we used data from 101 countries to train several machine learning models to compare the results between the models with and without the Hofstede's cultural dimensions. Our results show the use of cultural dimensions help improve the models, and that the ones that obtained a better prediction of the R_t were the ensemble models, especially the Random Forest which without the cultural dimensions obtained a MAE of 0.08659 and with the cultural dimensions obtained an improved error of 0.08575. Additionally, we used the best parameters for Random Forest Regression to predict the value of the R_t for five different countries where we found a better result for India (MAE=0.02637) and Brazil (MAE=0.03083).

Author contributions

All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Margarida Duarte. The first draft of the manuscript was written by Margarida Duarte and all authors contributed, reviewed and commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding

This work was supported by the Fundação para a Ciência e Tecnologia (FCT) within the following Projects: UIDB/04466/2020 and UIDP/04466/2020, by the European Commission Horizon Europe programme under the grant agreement number 101071330 (project InCITIES), and also by the European Commission - European Education and Culture Executive Agency (Eacea) - Erasmus-Edu-2024-Eur-Univ-1 programme under the grant agreement number 101177236 (project PIONEER).

Financial interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

8. References

1. Bo Y, Guo C, Lin C, Zeng Y, Li HB, Zhang Y, et al. Effectiveness of non-pharmaceutical interventions on COVID-19 transmission in 190 countries from 23 January to 13 April 2020. Int J Infect Dis. 2021;102:247–53.
2. Castex G, Dechter E, Lorca M. COVID-19: The impact of social distancing policies, cross-country analysis. Econ Disasters Clim Chang. 2020;135–59.
3. González-Bustamante B. Evolution and early government responses to COVID-19 in South America. World Dev. 2021;137.
4. Yang F, Zhang S, Wang Q, Zhang Q, Han J, Wang L, et al. Analysis of the global situation of COVID-19 research based on bibliometrics. Heal Inf Sci Syst [Internet]. 2020;8(1):30. Available from: https://doi.org/10.1007/s13755-020-00120-w
5. Sohrabi C, Alsafi Z, O'Neill N, Khan M, Kerwan A, Al-Jabir A, et al. World Health Organization declares global emergency: A review of the 2019 novel coronavirus (COVID-19). Vol. 76, International Journal of Surgery. Elsevier Ltd; 2020. p. 71–6.
6. Fontanet A, Autran B, Lina B, Kieny MP, Karim SSA, Sridhar D. SARS-CoV-2 variants and ending the COVID-19 pandemic. Lancet. 2021;397(10278):952–4.
7. Li Y, Campbell H, Kulkarni D, Harpur A, Nundy M, Wang X, et al. The temporal association of introducing and lifting non-pharmaceutical interventions with the time-varying reproduction number (R) of SARS-CoV-2 : a modelling study across 131 countries. Lancet Infect Dis. 2020;21(2):193–202.
8. Liu Y, Morgenstern C, Kelly J, Lowe R, Jit M. The impact of non-pharmaceutical interventions on SARS-CoV-2 transmission across 130 countries and territories. BMC Med. 2020;1–12.
9. Arroyo-Marioli F, Bullano F, Kučinskas S, Rondón-Moreno C. Tracking $\mathcal{R}$ of COVID-19: A New Real-Time Estimation Using the Kalman Filter. medRxiv. 2020 Jan;2020.04.19.20071886.
10. Perra N. Non-pharmaceutical interventions during the COVID-19 pandemic: A review. Phys Rep. 2021 May;913:1–52.
11. Haug N, Geyrhofer L, Londei A, Dervic E, Desvars-Larrive A, Loreto V, et al. Ranking the effectiveness of worldwide COVID-19 government interventions. Nat Hum Behav [Internet]. 2020;4(12):1303–12. Available from: https://doi.org/10.1038/s41562-020-01009-0
12. Cevasco KE, Roess AA, North HM, Zeitoun SA, Wofford RN, Matulis GA, et al. Survival analysis of factors affecting the timing of COVID-19 non-pharmaceutical interventions by U.S. universities. BMC Public Health [Internet]. 2021;21(1):1985. Available from: https://doi.org/10.1186/s12889-021-12035-6
13. Waitzberg R, Davidovitch N, Leibner G, Penn N, Brammli-Greenberg S. Israel's

response to the COVID-19 pandemic: tailoring measures for vulnerable cultural minority populations. Int J Equity Health [Internet]. 2020;19(1):71. Available from: https://doi.org/10.1186/s12939-020-01191-7

14. Xie Q, Xue Y. The Prediction of Public Risk Perception by Internal Characteristics and External Environment: Machine Learning on Big Data. Int J Environ Res Public Health. 2022;19(15).
15. Tang JW, Caniza MA, Dinn M, Dwyer DE, Heraud J-M, Jennings LC, et al. An exploration of the political, social, economic and cultural factors affecting how different global regions initially reacted to the COVID-19 pandemic. Interface Focus [Internet]. 2022;12(2). Available from: https://doi.org/10.1098/rsfs.2021.0079
16. Auger KA, Shah SS, Richardson T, Hartley D, Hall M, Warniment A, et al. Association between Statewide School Closure and COVID-19 Incidence and Mortality in the US. JAMA - J Am Med Assoc. 2020;324(9):859–70.
17. Chernozhukov V, Kasahara H, Schrimpf P. Causal impact of masks, policies, behavior on early covid-19 pandemic in the U.S. J Econom. 2021;220(1):23–62.
18. Duhon J, Bragazzi N, Kong JD. The impact of non-pharmaceutical interventions, demographic, social, and climatic factors on the initial growth rate of COVID-19: A cross-country study. Sci Total Environ. 2021;760:144325.
19. Islam N, Sharp SJ, Chowell G, Shabnam S, Kawachi I, Lacey B, et al. Physical distancing interventions and incidence of coronavirus disease 2019: Natural experiment in 149 countries. BMJ. 2020;370:1–10.
20. Koh WC, Naing L, Wong J. Estimating the impact of physical distancing measures in containing COVID-19: An empirical analysis. Int J Infect Dis. 2020;100:42–9.
21. White ER, Hébert-Dufresne L. State-level variation of initial COVID-19 dynamics in the United States. PLoS One. 2020;15(10 OCTOBER):1–13.
22. Fuller JA, Hakim A, Victory KR, Date K, Lynch M, Dahl B, et al. Mitigation Policies and COVID-19-Associated Mortality - 37 European Countries. MMWR Morb Mortal Wkly Rep. 2021 Jan;70(2):58–62.
23. Hofstede Insigths. Compare Countries. 2021.
24. Yeung AYS, Roewer-Despres F, Rosella L, Rudzicz F. Machine Learning–Based Prediction of Growth in Confirmed COVID-19 Infection Cases in 114 Countries Using Metrics of Nonpharmaceutical Interventions and Cultural Dimensions: Model Development and Validation. J Med Internet Res. 2021;23(4):e26628.
25. Zehtabian S, Khodadadeh S, Turgut D, Bölöni L. Predicting Infections in the Covid-19 Pandemic—Lessons Learned. In: Shaban-Nejad A, Michalowski M, Bianco S, editors. Multimodal AI in Healthcare Studies in Computational Intelligence [Internet]. Cham: Springer International Publishing; 2022. p. 279–91. Available from:

https://doi.org/10.1007/978-3-031-14771-5_20
26. Hofstede G. Culture's consequences: International differences in work-related values. Vol. 5. sage; 1984.
27. Guan Y, Deng H, Zhou X. Understanding the impact of the COVID-19 pandemic on career development : insights from cultural psychology. J Vocat Behav. 2020;119:103438.
28. Silva CJ, Cruz C, Torres DFM, Muñuzuri AP, Carballosa A, Area I, et al. Optimal control of the COVID-19 pandemic: controlled sanitary deconfinement in Portugal. Sci Rep. 2021;11(1):1–15.
29. Hoepner AGF, Hoepner R, Leippold M, Lin M-T, Lin Y. Efficacy of Non-Pharmacological Interventions Before COVID Mass Vaccination: An Open Data Study Across 185 Countries. Swiss Financ Inst Res Pap Ser [Internet]. 2023;23(17). Available from: https://ssrn.com/abstract=4375783
30. Hofstede G. Dimensionalizing cultures: The Hofstede model in context. Online readings Psychol Cult. 2011;2(1):919–2307.
31. Lu JG, Jin P, English AS. Collectivism predicts mask use during COVID-19. Proc Natl Acad Sci U S A. 2021;118(23).
32. Gelfand MJ, Jackson JC, Pan X, Nau D, Pieper D, Denison E, et al. The relationship between cultural tightness–looseness and COVID-19 cases and deaths: a global analysis. Lancet Planet Heal. 2021;5(3):e135–44.
33. Erman A, Medeiros M. Exploring the Effect of Collective Cultural Attributes on Covid-19-Related Public Health Outcomes. Front Psychol. 2021;12(March):1–21.
34. Wang Y. Government policies, national culture and social distancing during the first wave of the COVID-19 pandemic: International evidence. Saf Sci. 2021;135(October 2020):105138.
35. Duarte M, Moro S, Ferreira da Silva C. Does cultural background influence the dissemination and severity of the COVID-19 pandemic? Heliyon. 2022;8(2).
36. Gokmen Y, Baskici C, Ercil Y. The impact of national culture on the increase of COVID-19: A cross-country analysis of European countries. Int J Intercult Relations. 2021;81:1–8.
37. Güss CD, Tuason MT. Individualism and Egalitarianism Can Kill: How Cultural Values Predict Coronavirus Deaths Across the Globe. Front Psychol. 2021;12(June):1–13.
38. Luu T, Huynh D. Does culture matter social distancing under the COVID-19 pandemic? Saf Sci. 2020;130:104872.
39. Moro S, Laureano R, Cortez P. Using data mining for bank direct marketing: An application of the CRISP-DM methodology. ESM 2011 - 2011 Eur Simul Model Conf Model Simul 2011. 2011;(Figure 1):117–21.

40. Raju VNG, Lakshmi KP, Jain VM, Kalidindi A, Padma V. Study the Influence of Normalization/Transformation process on the Accuracy of Supervised Classification. Proc 3rd Int Conf Smart Syst Inven Technol ICSSIT 2020. 2020;(Icssit):729–35.
41. Sorjamaa A, Hao J, Lendasse A. Mutual information and k-Nearest Neighbors approximator for time series prediction. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics). 2005;3697 LNCS:553–8.
42. Laarne P, Zaidan MA, Nieminen T. ennemi: Non-linear correlation detection with mutual information. SoftwareX. 2021;14:100686.
43. Steurer M, Hill RJ, Pfeifer N. Metrics for evaluating the performance of machine learning based automated valuation models. J Prop Res. 2021;38(2):99–129.
44. Akgün B, Öğüdücü ŞG. Streaming linear regression on spark MLlib and MOA. Proc 2015 IEEE/ACM Int Conf Adv Soc Networks Anal Mining, ASONAM 2015. 2015;1244–7.
45. Crawley MJ. Applied Linear Regression Models:Non-linear Regression. Book. 2012;661–84.
46. Mendes-Moreira J, Soares C, Jorge AM, De Sousa JF. Ensemble approaches for regression: A survey. ACM Comput Surv. 2012;45(1).
47. Ganaie M, Hu M, Tanveer M, Suganthan P. Ensemble deep learning: A review [Internet]. 2021. Available from: https://arxiv.org/pdf/2104.02395.pdf
48. Miikkulainen R, Francon O, Meyerson E, Qiu X, Sargent D, Canzani E, et al. From Prediction to Prescription: Evolutionary Optimization of Nonpharmaceutical Interventions in the COVID-19 Pandemic. IEEE Trans Evol Comput. 2021;25(2):386–401.
49. Sturmberg, J. P., & Martin, C. M. (2020). COVID-19–how a pandemic reveals that everything is connected to everything else. *Journal of Evaluation in Clinical Practice*, *26*(5), 1361.