Mobile Networks and Applications (2022) 27:1967–1976
https://doi.org/10.1007/s11036-022-01966-y

Feature Importance Analysis by Nowcasting Perspective to Predict COVID-19

André Vinícius Gonçalves[1,2] · Gustavo Medeiros de Araujo[2] · Leandro Pereira Garcia[3] · Fernanda Vargas Amaral[4] · Ione Jayce Ceola Schneider[5]

Accepted: 15 November 2021 / Published online: 23 April 2022

Abstract
The present work raises an investigation about prediction and the feature importance to estimate the COVID-19 infection, using Machine Learning approach. Our work analyzed the inclusion of climatic features, mobility, government actions and the number of cases per health sub-territory from an existing model. The Random Forest with Permutation Importance method was used to assess the importance and list the thirty most relevant that represent the probability of infection of the disease. Among all features, the most important were: i) the variables per region health stand out, ii) period comprised between the date of notification and symptom onset, iii) symptoms features as fever, cough and sore throat, iv) variables of the traffic flow and mobility, and also v) wheathers features. The model was validated and reached an accuracy average of 81.82%, whereas the sensitivity and specificity achieved 87.52% and the 78.67% respectively in the infection estimate. Therefore, the proposed investigation represents an alternative to guide authorities in understanding aspects related to the disease.

Keywords Feature importance · Feature engineering · Machine learning · Prediction model · COVID-19

1 Introduction

In December 2019, a new coronavirus, called SARS-CoV-2, was recognized in the city of Wuhan, China, and spread quickly to other countries in the world. In January 2020, the World Health Organization declared a Public Health Emergency of International Importance, and in March, the COVID-19 pandemic. At the beginning of September 2021 there were already more than 220.5 millions confirmed cases and more than 4.5 millions deaths from the disease [34]

When infecting the human body, there is a period of latency, followed by an infectious period. During this period, the infected person can transmit to others through coughing and sneezing. The virus mainly affects the respiratory tract and the first symptoms appear after the incubation period. In the most common cases,symptoms include fever, cough and fatigue, which will do so within on average for 11 to 14 days of infectiion [25]. Other symptoms, such as mucus production, headache, hemoptysis, diarrhea, dyspnoea, lymphopenia can also appear. The main clinical diagnosis is pneumonia [3, 32, 41, 46]. Furthermore, the risk of symptomatic infection increases with age. Thus, older individuals are more likely to have symptomatic infection and worse outcomes [3].

Laboratory diagnosis is an important tool for diagnosis, as well as for follow-up, evaluation and evolution of the case. The recommended diagnostic test is the real-time polymerase chain reaction (RT-PCR) of nasal and oropharyngeal swab samples. Other serological tests can be used to detect immune responses, such as class M (IgM) and class G (IgG). However, it is important to use resources rationally in conducting diagnostic tests [49].

✉ André Vinícius Gonçalves
avgandre@gmail.com

Gustavo Medeiros de Araujo
gustavo.araujo@ufsc.br

Leandro Pereira Garcia
lpgarcia18@gmail.com

Fernanda Vargas Amaral
fevmaral@hotmail.com

Ione Jayce Ceola Schneider
ione.schneider@ufsc.br

1 Federal Institute of Northern Minas Gerais, Montes Claros, Minas Gerais, Brazil

2 PGCIN, Federal University of Santa Catarina, Florianópolis, Brazil

3 Florianópolis Municipal Health Department, Florianópolis, Brazil

4 University of Malaga, Malaga, Spain

5 Federal University of Santa Catarina, Araranguá, Brazil

Regarding the rational use of resources for detection of infection spread, artificial intelligence techniques have been used to predict the diagnosis of COVID-19. The algorithms are managing to predict the stage of disease by means of several features such as age, comorbidities, symptoms, diagnosis and outcome [33, 51].

A very useful approach in this context is Nowcasting, mainly due to the transmission dynamics of an epidemic or pandemic. It's a technique used for prediction of the present, that is, an estimate of the current number given an event [5]. Although it generally uses time-series, recent advances in machine learning techniques have diversified the possibilities [29, 48].

This paper is an expanded version of the other article [19] and complements the initially developed proposal. Here, the data from the public health system in the capital of Santa Catarina – Brazil was enlarged with variables of weather, mobility and non-pharmacological government measures.

Thus to create predictive model we performed an investigation to assess the main features that can determine Covid-19 infection of an individual. In our work, we conduct several experiments with 221 features to label the 30th most important features that represent the high Covid-19 infection likelihood.

1.1 Contributions

Among the contributions of our work, we can highlight:

1. The verification of the high importance of the features of confirmed, discarded, and removed per territories and sub-territories of health, as well as the features of symptoms (fever, cough, and sore throat), all along the time of the notification date.
2. An intensive feature importance investigation results in findings that also highlighted the importance of traffic load and mobility, which reflects the people's isolation level.
3. The accuracy of the model achieved an average of 81.82% of correctness in determining whether the individual is infected with Covid-19.

The remainder of this paper is structured as follows: In Section 2, we describe the more relevant related works on the effort to determine the Covid-19 infection; Section 3 introduces the methodology applied to feature engineering; Section 4 detail the experimental assessments; Section 5 outline the discussion about results and finally, in Section 6, we present our final remarks and future work.

2 Related work

COVID-19 had a significant impact on the life and economy of several countries [21]. In addition to collapsing economies, the moral values of nations have been strongly affected by the pandemic [43]. All the impact, economic and social, motivated the Pan American Health Organization to seek to better understand the signs and symptoms of Sars-cov-2, in order to disseminate this knowledge. The challenge of the pandemic is to find the best model that elucidates the initial growth trajectory and the epidemiological characteristics of the new coronavirus [40]. In this sense, the predictive models has been useful to deal with the dynamic behavior of this virus [44].

Sars-cov-2 is a respiratory virus transmitted through droplets of saliva, sneezing or by close contact. In their study, [47] described 69 cases of COVID-19 in China, where it was identified that 15% of individuals had fever, cough and dyspnoea. However, a survey conducted in the United States, showed that 50% of patients affected by this virus did not have a fever, however cough and dyspnea were reported by 88% of people with the virus [6]. Still, in other studies, reports of symptoms were difficult to measure objectively, such as anosmia (loss of smell), hyposmia (decreased smell) and ageusia (loss of taste) [24].

In addition, infected individuals may never develop symptoms, others may have mild symptoms or develop moderate to severe Sars-cov-2 disease [35]. In order to understand the symptoms that best represent the pandemic, researchers around the world try to understand the behavior of the virus [24]. A group of researchers from Spain found five patterns of skin infection that may be associated with COVID-19. These patterns were repeated in patients with varying demographic characteristics, in different periods and with different severities of the disease. Among these patterns are maculopapular rashes (47% of cases), vesicles or pustules (19% of cases), hives (19% of cases) and other vesicular rashes (9% of cases) and livedo or necrosis (in 6% of cases) [15].

A preliminary analysis by the World Health Organization (WHO) shows that in relation to gender, there is a relatively uniform distribution of infections between women and men (47% versus 51 respectively), however, it seems that men have a higher rate mortality rate (58%) in relation to women [35]. Nevertheless, due to the need to know the outbreak of COVID-19, some studies are being carried out considering exogenous factors such as the social environment, climatic variables, pollution and population density [44]. Other studies point to the role of room temperature in the survival and transmission of viruses. According to the WHO, several environmental factors can influence the spread of communicable diseases that can cause epidemics. The underlying theory is that the number of cases and the spread of previous infectious viruses demonstrate seasonal patterns, affected by the climate, and therefore Covid-19 is likely to be similar in this respect [30].

Therefore, the prediction of a pandemic can be made based on several parameters, such as the impact of environmental factors, incubation period, impact of quarantine, age,

sex and many more. The difficulty in predicting the number of cases of a pandemic is the fact that the number of cases to be studied does not match the total infected population. [37].

Considering the importance of knowing this difficult epidemiological scenario, the Forecasting Models are an alternative to unravel the impacts of the pandemic. This technique assess past situations, which allows for better predictions about the situation that will occur in the future [43, 44]. Another is the Nowcasting Models that provide a prediction for cases that have not yet been reported [31]. It's very useful in situations where there is a delay in the response of the control instruments.

Both techniques help managers to create strategic planning and carry out decision making in the most assertive manner possible [30]. Then, several models have been developed which allow governments not to focus only on underlying methods such as personal judgment. Many use mathematical and statistical methods, as well as Artificial Intelligence techniques, to predict epidemic trajectories [11, 13, 26, 50], evaluation of non-pharmaco-logical interventions [4, 14, 22, 38], among others.

Another relevant theme in this context is the analysis of the importance of features that contributes to predictive modeling through the recognition of related variables. In one article, for example, the authors managed to develop a new algorithm, called Variance Based Feature Weighting, which not only ranks the COVID-19 symptoms but also assigns a quantitative importance measure [2]. Results indicated fever - 75%, cough - 39.8%, fatigue - 16.5%, sore throat - 10.8% and shortness of breath - 6.6% as quantitative measures of relevance for disease detection.

In another, researchers proposed a classification to predict the clinical severity of patients with COVID-19 [12]. The authors used 37 features, including basic patient information, a physical index, initial examination findings, clinical findings, comorbid diseases, and general blood test results at an early stage. The feature importance analysis was performed with AdaBoost, Random Forest and XGBoost, which selected the 20 most important to be processed in the Deep Learning classifier. The results showed that age, lymphocyte level, platelet count and shortness of breath or dyspnea were the most relevant factors in predicting severity.

And finally, a research in which the importance of various features in 2,787 US counties was investigated during the COVID-19 transmission trajectory [27]. The period involved the stages of outbreak, social distancing and reopening of activities. Through data-driven machine learning models, 23 features distributed into six categories were evaluated: social demographics of counties, population activities, mobility within the counties, movement across counties, disease attributes and social network structure. The results reveal that in municipalities with high population densities, mobility resources have a greater impact. As for municipalities with a low population density, the importance of the social network structure is smaller and the index of social distancing is greater. This allow policymakers to adjust control measures and strategies according to different levels and different time points.

3 Methodology

The main goal of our work is to analyze which features most contribute to the diagnosis of suspected cases of COVID-19 using the classification technique with Machine Learning. The methodology steps that can be seen in Fig. 1 are: 1) data selection and extraction, 2) data pre-processing and feature engineering, 3) hyperparameterization and feature selection and 4) model validation. Each step is detailed further next.

3.1 Data selection and extraction

The model considered four datasets for the classification task with representation of climate, mobility, government policy measures and, above all, patient data.

The first dataset, on climate, was obtained from the National Institute of Meteorology (INMET) [23]. It contains meteorological data from several cities in Brazil.

The second was Google Mobility Report. It has daily data on the movement trend of people in more than 200 countries and their respective cities, since February 2020 [28]. It's organized into six categories of places, such as retail and recreation, groceries and pharmacies, parks, public transit stations, workplaces and residential areas. In this article only the variables of Percent Change from Baseline were used.

The third database concerns non-pharmacological policy actions to control COVID-19 transmission. For this, mitigation measures and flexibility of activities defined by the decrees of the city of Florianópolis (SC), between March and May 2020, were analyzed [10].

The functioning of the society's activities were divided into several types as sports, gyms, beach, schools, mandatory internships, open fairs, commerce, shopping, public transport, food services, cultural establishments, churches, hotels, public agencies, civil construction, essential services, among others. According to the delimitation of the decrees, each activity have been classified daily into levels: open, open with restrictions and closed [17].

The last dataset used has been set corresponding to 1,927 reported cases of COVID-19 in the period between 02/18/2020 and 05/25/2020, obtained according to data availability. It was extracted from the Health Department of Florianópolis, capital of the State of Santa Catarina in southern Brazil and is available to be analysed [18].

According to the [16], the database comes from three sources: 1) anonymized data on suspected and confirmed

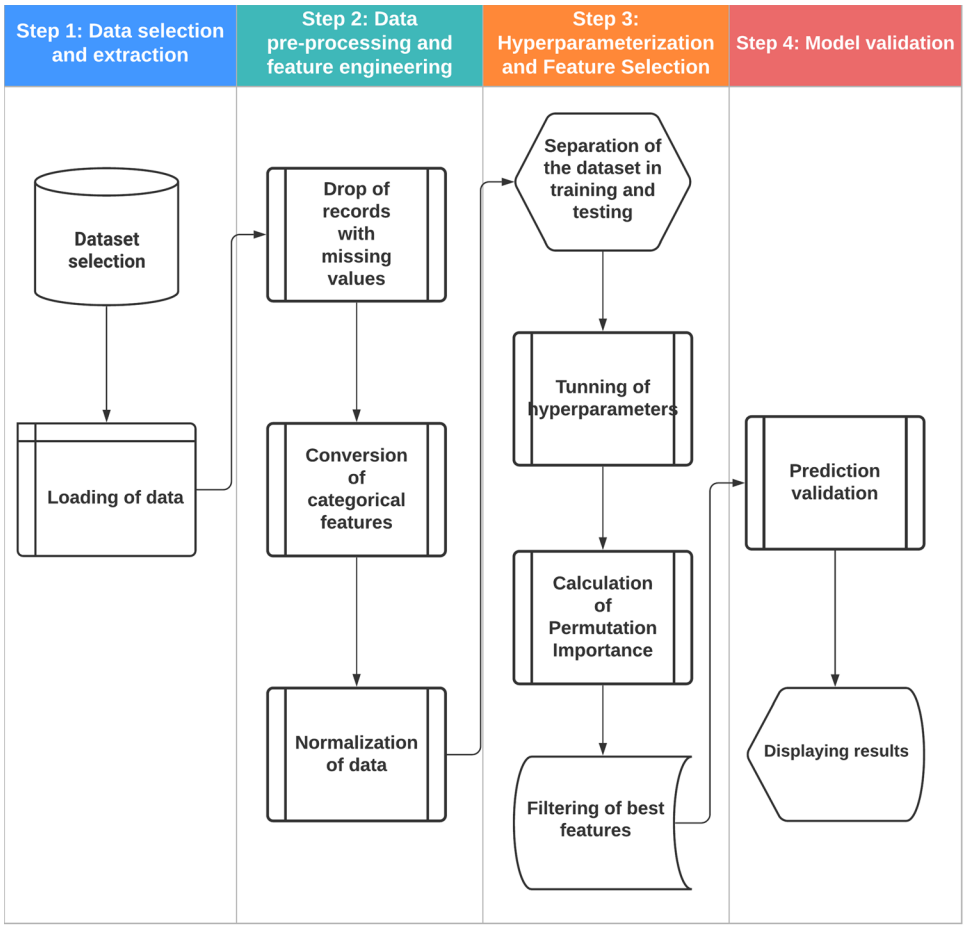

Fig. 1 Methodology Flow [19]

cases resident in Florianópolis; 2) demographic data of the 49 health regions that make up the municipality; and 3) data on the mobility represented by the traffic flow in the municipality.

The database contains individual data on the diagnosis (confirmed or discarded), sex, age (in years), and age groups (under 10 years old, 10 years old under 20 years old, 20 years old under 40 years old) years old, 40 years old to under 60 years old, 60 years old to under 80 years old, 80 years old or more), skin color (white and not), date of onset of symptoms, in addition to the following clinical data of symptoms of the disease: pain throat, dyspnoea, fever and cough.

There is also data on health regions in the city of Florianópolis. In total, there are 49 territories and 104 subterritories that correspond to regional divisions of the city.

Furthermore, the database contains the following: demographic data for health territories the total number of inhabitants and by sex; the number of persons aged 1 year, 2 years and so on up to 100 years or more; the number of people by skin color (white, black, yellow, Brazilian, indigenous and ignored); the number of people by years of schooling (from 1 to 17 years completed or more, in addition to literate, non-literate, literate through youth and adult literacy programs and with uninformed schooling); the total income per household, the average income of the households, the total income of the heads of households, the average income of the heads of households, the total income per person and the average income per person, the proportion of males, persons with 60 years of age or more, of people with non-white skin and of people with 10 years or less of education, as possible indicators of vulnerability.

Regarding mobility features, the database provides data on the average daily traffic on four major avenues in the city. The time window for calculating the average considers it starts on the day of symptom detection until the thirteenth day before, that is, it is a window delayed in time.

3.2 Data pre-processing and feature engineering

Initially, the patient's dataset was processed. All records with the value 'Missing' in the attributes of symptoms (Sore Throat, Dyspnea, Fever and Cough) and Diagnosis were removed. Then, the categorical attributes were converted to numerical ones, using the One-hot-encoding technique for Race / Color, Age group, Screening Method and symptoms, and Feature Hashing [45] for Territory and Sub-territory. The Table 1 has the conversion process result detailed.

Another procedure performed was the creation of new attributes. As suggested by [16], the number of infected people (with a positive diagnosis and up to 14 days after the onset of symptoms) in each health territory was calculated. Moreover, according to the principle of the SIR model of epidemiology [7], it was proposed to include the number of people discarded (with a negative diagnosis) and the number of people removed (with a positive diagnosis and more than 14 days after the onset of symptoms).

Furthermore, it was included the rate of people infected by the number of inhabitants of their respective health territory, as well as the rate of discarded and removed rate.

Following the idea of grouping the number of cases in each compartment of the SIR model by health territory, the process was repeated for sub-territory, creating three new variables: number of infected people in last 14 days, number of people discarded and the number of people removed, both per sub-territory.

Besides that, the Google Mobility Report was joined to the initial base with including the variables pertinent to population moviment. The key among the datasets was the date considering a three-day delay between the dynamics of daily activities and the onset of the patient's first symptoms, in an eventual contamination.

The process was repeated for climate dataset. The temperature data (mean, minimum and maximum), hourly humidity (mean, minimum and maximum), wind speed (mean, minimum and maximum), radiation (mean, minimum and maximum) and the sum of precipitation were consolidated in daily values and added to the main base.

Data on non-pharmacological government measures, obtained from municipal decrees, were also added. Before, the categorical attribute of the level of openness of each activity was transformed into a numerical value with the following scale: -1 (open), 0 (open with restriction) and 1 (closed). An index equivalent to the daily arithmetic mean of all activities was also created [17].

Finally, the data were normalized, transforming them to values within the range [0, 1] and, thus, establishing a common scale.

Table 1 Conversion of categorical features

Categorical Feature	Factors	Method
Race/Color	White; Yellow; Black; Parda; Unknown	One-hot-encoding
Age range	< 10; 10 <= *years* < 20; 20 to 80 step 20; > 80	One-hot-encoding
Screening Method	3 methods	One-hot-encoding
Fever	Yes; No	One-hot-encoding
Cough	Yes; No	One-hot-encoding
Sore throat	Yes; No	One-hot-encoding
Dyspnea	Yes; No	One-hot-encoding
Health Territory	48 regions	FeatureHasher
Health Sub-territory	104 subregions	FeatureHasher

3.3 Hyperparameterization and feature selection

The database was divided into training and test basis, 70% for training and 30% for testing. As there is an imbalance in the amount of data between the discarded and confirmed cases, the first being a larger amount, the sample was balanced using the Undersampling technique.

In the training stage, cross-validation was adopted as a way to assess the model's generalization. According to [39], the technique consists of dividing the database into k folds, one of which is selected at a time to be the test set and the other k-1 are used as a training set. The test is repeated until each of the k folds is used as a test set. In the end, the accuracy is given by the mean of the accuracy obtained for each of the k folds.

Hyperparametrization was performed using a random combination of parameters with 10 iterations in each tuning process. Accuracy was chosen as the maximization score.

After defining the parameters of the algorithm, the feature selection was performed considering the values of permutation importance as a criterion for assessing the degree of importance [1]. The criterion used was to select only those features with a value greater than zero. In this way, the features with values above this threshold remained in the model and the rest were removed from the database.

3.4 Model validation

In the last step of the process, with the algorithm trained and configured with the best parameters that fit the model, the algorithm was validated with the test base to assess its prediction capacity.

Steps 3 and 4 were repeated 100 times and the results for each were stored. Then the data were used to calculate the mean and stantard devation of evaluation metrics and permutation feature importances.

The equipment used to carry out the experiments had: i) Intel (R) Xeon (R) Gold 6126 CPU @ 2.60GHz CPU with 12 CPUs, ii) 32.0 GB of RAM, iii) 250 GB of hard disk and iv) Linux Ubuntu 16.04. The entire implementation was developed in the Python programming language, version 3.8.

4 Experiment assessments

We carried out experiments to analyze the evaluation metrics that measure the accuracy, in addition to ascertaining the features that had the most contribution to the performance.

Table 2 Random Forest Hyperparameters

Parameter	Value
criterion	[entropy, gini]
n_estimators	[5...100]
max_depth	[None, 1...5]
min_samples_split	[2...5]
min_samples_leaf	[1...5]
min_weight_fraction_leaf	[0.0...0.5]
max_features	['auto', 0.1...0.5]
bootstrap	False, True

The specific parameters of the Random Forest are presented in Table 2 as well as the possible value ranges. Through them, the best configuration is adjusted by means of a random search of hyperparameters.

The parameters described in Table 3 relate to the general settings of the environment.

In the experiments, the metrics used in the analysis of the proposed model to assess performance were accuracy, sensitivity and specificity. The data samples were obtained by running the algorithm repeatedly and they are presented in Table 4 as mean and standard deviation.

The Random Forest algorithm had an accuracy of 0.82695 0.02344 (mean standard deviation) on the training set and 0.81819 0.02331 on the test set. These results achieved a AUC ROC mean of 0.90890 0.02515 in the test data. Similarly, the precision recall curve obtained an average result of 0.82344 0.04823.

To make a parallel with the results obtained by [19] and analyze whether the changes contributed to an improvement in the metrics, the Mann Whitney test was used. It is a non-parametric statistical test that compares two independent samples [20].

In the accuracy analysis of the data sample obtained from both experiments that involved 100 executions, the initial article had 0.79275 and 0.03843 of median and interquartile range against 0.82383 and 0.03152 achieved in this paper. By applying the aforementioned test, it is possible to state with 95% confidence that there is a difference between the two groups and, therefore, the alternative hypothesis was accepted ($U = 2{,}053.5$; $p < 0.001$).

Table 3 General SettingsGeneral Settings

Parameter	Value
Execution Amount	100
Folds	10
Training/Test	70/30
RandomizedSearchCV Interactions	10
Features Selection Threshold	above zero (> 0)

Table 4 Prediction Metrics

Metric	Training (M SD)	Test (M SD)
Accuracy	0.82695 0.02344	0.81819 0.02331
Sensibility	0.87193 0.04853	0.87524 0.06030
Specificity	0.78198 0.03418	0.78668 0.04048

The same could be seen for sensitivity ($U = 3{,}546.0$; $p < 0.001$), 0.85194 and 0.06189 against 0.87452 and 0.06189 of median and interquartile range, respectively. And also specificity ($U = 3{,}170.0$; $p < 0.001$), 0.75603 and 0.5565 against 0.79089 and 0.05094 of same measurements. Therefore, the proposals presented here served to improve prediction in the three metrics.

Table 5 Features Permutation Importance of Accuracy Score

Feature	P. Importance (M SD)
1a Confirmed Sub-territory 14 Days	0.08326 0.03291
2a Notification Date	0.04021 0.02775
3a Fever0	0.02140 0.01564
4a Fever1	0.01812 0.01371
5a Cough1	0.01187 0.00970
6a Cough0	0.01093 0.00995
7a Confirmed Rate Territory 14 Days	0.00865 0.00618
8a Confirmed Territory 14 Days	0.00390 0.00365
9a Discarded Territory	0.00128 0.00262
10a Sore Throat0	0.00089 0.00195
11a Sore Throat1	0.00073 0.00173
12a Symptoms Start Date	0.00065 0.00294
13a Removed Rate Territory	0.00058 0.00197
14a Discarded Rate Territory	0.00056 0.00213
15a Mean Temperature	0.00049 0.00231
16a Workplaces Percent	0.00049 0.00171
17a Essential Services	0.00041 0.00166
18a Age	0.00035 0.00125
19a Common Areas Condominiums	0.00035 0.00159
20a Maximum Temperature	0.00033 0.00169
21a Mean Traffic-lag1	0.00028 0.00131
22a Grocery and Pharmacy Percent	0.00024 0.00104
23a Mean Traffic-lag 8	0.00024 0.00114
24a Age Above 91	0.00021 0.00233
25a Sub-territory 1	0.00019 0.00083
26a Removed Territory	0.00018 0.00122
27a Minimum Temperature	0.00017 0.00130
28a Residential Percent	0.00016 0.00151
29a Mean Traffic All Lags	0.00016 0.00129
30a Non-Essential Public Agencies	0.00016 0.00130

The main features selected with their respective Permutation Importance percentages are shown in Table 5. The results presented below are in the form of mean and standard deviation for the set of executions.

For a better visual understanding of the importance of the resources, the 15th most important variables obtained by the model are shown in Fig. 2.

Lastly, the response time of the algorithm had an average result of 11.74 of 1.94 seconds, considering the training step

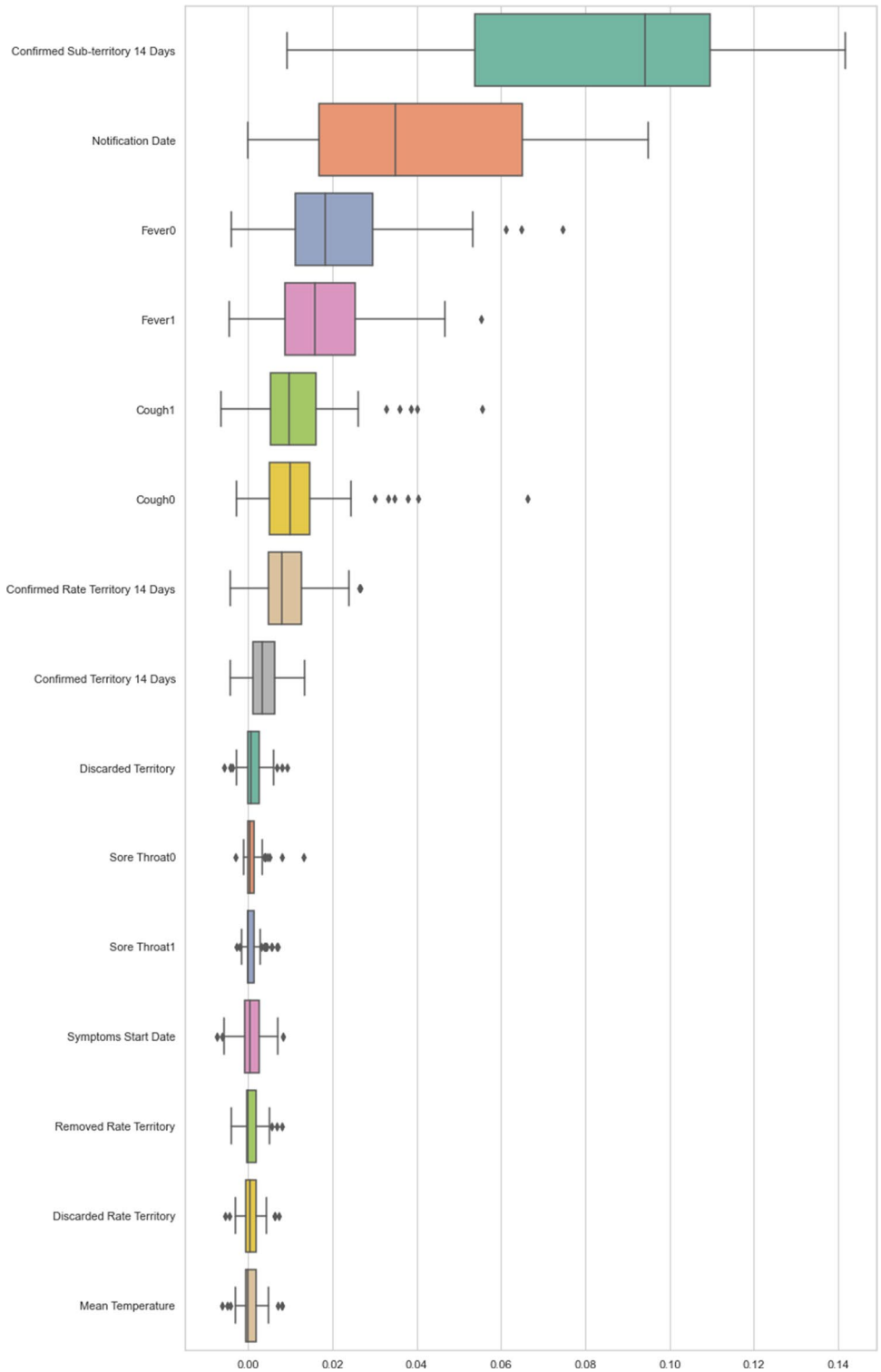

Fig. 2 Features Permutation Importance of Accuracy Score

that involved the hyperparameter tuning process and feature selection, in addition to the test step that consisted of the model validation.

5 Discussion

After nearly two years of its discovery, COVID-19 is a disease that arouses much interest because of its great impact on humankind. Wherefore, this work proposed to investigate the relevance of a set of variables in the diagnostic prediction of the disease.

The investigation started with the acquisition of a preliminary database with 221 features, which after going through pre-processing, increased to 277 due to the techniques of coding categorical variables. Then, the model was processed and analyzed by the Permutation Feature Importance method to assess the impact of each feature on the accuracy metric.

The most important feature was the number of Confirmed per Sub-territory in the last 14 days. All symptom features appeared among the thirty most significant, with the exception of dyspnea. This fact corroborates with the researches that investigate the symptoms and indicate fever, cough and sore throat among some of the most common ones [8, 9, 36].

The weather features had a good representation on the prediction importance scale. Among the top thirty classified are Mean Temperature, Maximum Temperature and Minimum Temperature. These results reinforce the high correlation of climatic variables in the prediction of COVID-19, as other studies have point out [42, 44].

Mobility features were also meaningful. There were fourteen features to represent traffic on the city's four major avenues and three of them were among the most important features. In the same direction, among the six variables extracted from Google Mobility Report, three arised listed in Table 5: Workplaces Percent, Grocery and Pharmacy Percent and Residential Percent.

The features related to non-pharmacological government measures had a more timid presence. Only the Essential Services, Common Areas of Condominium and Non-Essential Public Agencies appeared, given a total of twenty-seven.

In this paper, the feature engineering process carried out in the pre-processing step resulted in the creation of new health sub-territory variables contributed significantly to the model performance. In particular "Confirmed_subterritory_14days", the attribute most significant with a Permutation Importance value greater than the double of the second.

Thus, it is clear that the "health region" factor is very relevant in the correct classification of the disease diagnosis, as highlighted in [19].

Finally, the results of the model were somewhat satisfactory. This matter is associated with the addition of new variables such as data of climate, mobility, government actions and, above all, confirmed and discarded cases related to the health sub-territory, that were not present in the previous article.

6 Final remarks and future work

The present work shows a investigation about the feature importance in a prediction diagnostic model for cases of COVID-19, using the classification technique with Machine Learning.

These classification approaches are fundamental for monitoring the number of virus reproductions and for making decisions in the face of the pandemic. The advantage of them is to produce quick responses and relatively low cost compared to laboratory diagnosis.

The methodology section emphasized the hyperparametrization and feature selection techniques, as the research aimed to investigate two aspects: the features that best contributed to the performance of the model and the results of the hit rates in the validation of the test step.

In the first investigation step, the Random Forest with Permutation Importance method was used to assess the impact of the features on the results. Among the 277 variables that make up the database, the most relevant are: number confirmed per sub-territory of the last 14 days, date of notification and onset of symptoms, attributes of fever and cough, rate and number of confirmed per territory of the last 14 days, number discarded from territory and symptom variable sore throat. Other also were importants as temperature, age, flow traffic and mobility.

Regarding the second stage of the investigation, the metrics showed consistent results. Accuracy had a mean of 81.82%, whereas sensitivity reached 87.52% and specificity 78.67% of cases. All of them showed significant results compared to [19].

Therefore, the research conducted has shown that there is a feasible alternative in the process of underdiagnosis COVID-19 disease, considering the most relevant characteristics for the determination of infection. The limitation of the created model is directly related to the dataset used, as the importance of the features can change according to the applied environment (algorithms, hyperparameters, database, etc.).

In the near future, we intend to evaluate the model in the context of other cities. Another possibility is to add vaccination data and, later, to analyze the behavior of the classifier model.

Data Availability The dataset analysed and processed during this study are available in the Havard Dataverse repository: https://doi.org/10.7910/DVN/YGG0TQ.

References

1. Altmann A, Toloşi L, Sander O, Lengauer T (2010) Permutation importance: a corrected feature importance measure. Bioinformatics 26(10):1340–1347
2. Alzubaidi MA, Otoom M, Otoum N, Etoom Y, Banihani R (2021) A novel computational method for assigning weights of importance to symptoms of covid-19 patients. Artif Intell Med 112:102018. https://doi.org/10.1016/j.artmed.2021.102018
3. Ashour HM, Elkhatib WF, Rahman M, Elshabrawy HA, et al. (2020) Insights into the recent 2019 novel coronavirus (sars-cov-2) in light of past human coronavirus outbreaks. Pathogens 9(3):186
4. Awaidy SA, Mahomed O (2020) Impact of non-pharmaceutical interventions on the covid-19 epidemic: a modelling study. SAGE Open Medicine 8:2050312120979462
5. Bańbura M, Giannone D, Reichlin L (2010) Nowcasting. Tech. rep., ECB Working Paper
6. Bhatraju PK, Ghassemieh BJ, Nichols M, Kim R, Jerome KR, Nalla AK, Greninger AL, Pipavath S, Wurfel MM, Evans L, et al. (2020) Covid-19 in critically ill patients in the seattle region—case series. N Engl J Med 382(21):2012–2022
7. Brauer F (2005) The kermack–mckendrick epidemic model revisited. Math Biosci 198(2):119–131
8. Burke RM, Killerby ME, Newton S, Ashworth CE, Berns AL, Brennan S, Bressler JM, Bye E, Crawford R, Morano LH, et al. (2020) Symptom profiles of a convenience sample of patients with covid-19—united states, january–april 2020. Morb Mortal Wkly Rep 69(28):904
9. Carfì A, Bernabei R, Landi F et al (2020) Persistent symptoms in patients after acute covid-19, vol 324
10. CHF (2021) City hall of florianópolis. http://www.pmf.sc.gov.br/transparencia/index.php?pagina=legislacaoCOVID&menu=11&cms=legislacao+referente+a+covid19&IdEntidade=17. Accessed Sept 2021
11. Chowell G, Tariq A, Hyman JM (2019) A novel sub-epidemic modeling framework for short-term forecasting epidemic waves. BMC Med 17(1):164. https://doi.org/10.1186/s12916-019-1406-6
12. Chung H, Ko H, Kang WS, Kim KW, Lee H, Park C, Song HO, Choi TY, Seo JH, Lee J (2021) Prediction and feature importance analysis for severity of covid-19 in South Korea using artificial intelligence: Model development and validation. J Med Internet Res 23(4):e27060. https://doi.org/10.2196/27060
13. Fanelli D, Piazza F (2020) Analysis and forecast of covid-19 spreading in china, Italy and france. Chaos, Solitons & Fractals 134:109761. https://doi.org/10.1016/j.chaos.2020.109761
14. Flaxman S, Mishra S, Gandy A, Unwin HJT, Mellan TA, Coupland H, Whittaker C, Zhu H, Berah T, Eaton JW, et al. (2020) Estimating the effects of non-pharmaceutical interventions on covid-19 in europe. Nature 584(7820):257–261
15. Galván Casas C, Catala A, Carretero Hernández G, Rodríguez-jiménez P, Ferná ndez-Nieto D, Rodríguez-Villa Lario A, Navarro Fernández I, Ruiz-Villaverde R, Falkenhain-López D, Llamas Velasco M, et al. (2020) Classification of the cutaneous manifestations of covid-19: a rapid prospective nationwide consensus study in spain with 375 cases. British J Dermatol 183(1):71–77
16. Garcia LP, Goncalves AV, de Andrade MP, Pedebos LA, Vidor AC, Zaina R, de Luca Canto G, de Araujo GM, Amaral FV (2020) Estimating underdiagnosis of covid-19 with nowcasting and machine learning: experience from Brazil. medRxiv
17. Garcia LP, Traebert J, Boing AC, Santos GFZ, Pedebôs LA, d'Orsi E, Prado PI, Veras MADSM, Boava G, Boing AF (2020) O potencial de propagação da COVID-19 e a tomada de decisão governamental: uma análise retrospectiva em Florianópolis, Brasil. Revista Brasileira de Epidemiologia 23. https://doi.org/10.1590/1980-549720200091. http://www.scielo.br/j/rbepid/a/WJLGyZfwFkfPGVLMVW5y8ch/?lang=pt. Publisher: Associação Brasileira de Saúde Coletiva. Accessed Nov 2021
18. Gonçalves AV (2021) Florianópolis COVID-19. https://doi.org/10.7910/DVN/YGG0TQ
19. Gonćalves AV, Schneider IJC, Amaral FV, Garcia LP, Medeiros de Araújo G (2021) Feature importance investigation for estimating covid-19 infection by random forest algorithm. In: Bisset Álvarez E (ed) Data and information in online environments. Springer International Publishing, Cham, pp 272–285
20. Hart A (2001) Mann-Whitney test is not just a test of medians: differences in spread can be important. BMJ : British Med J 323(7309):391–393 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1120984/. Accessed Sept 2021
21. He S, Tang S, Rong L (2020) A discrete stochastic model of the covid-19 outbreak: Forecast and control. Math Biosci Eng 17:2792–2804
22. Hsiang S, Allen D, Annan-Phan S, Bell K, Bolliger I, Chong T, Druckenmiller H, Huang LY, Hultgren A, Krasovich E, et al. (2020) The effect of large-scale anti-contagion policies on the covid-19 pandemic. Nature 584(7820):262–267
23. INMEP (2021) Brazilian institute of meteorology. https://portal.inmet.gov.br/. Accessed Sept 2021
24. Iser BPM, Sliva I, Raymundo VT, Poleto MB, Schuelter-Trevisol F, Bobinski F (2020) Definição de caso suspeito da covid-19: uma revisão narrativa dos sinais e sintomas mais frequentes entre os casos confirmados. Epidemiologia e Serviços de Saúde 29:e2020233
25. Lauer SA, Grantz KH, Bi Q, Jones FK, Zheng Q, Meredith H, Azman ASA, Reich NG, Lessler J (2020) The incubation period of coronavirus disease 2019 (covid-19) from publicly reported confirmed cases: Estimation and application, vol 172. https://doi.org/10.7326/M20-0504. PMID: 32150748
26. Li L, Yang Z, Dang Z, Meng C, Huang J, Meng H, Wang D, Chen G, Zhang J, Peng H, Shao Y (2020) Propagation analysis and prediction of the covid-19. Infectious Disease Modelling 5:282–292. https://doi.org/10.1016/j.idm.2020.03.002. https://www.sciencedirect.com/science/article/pii/S2468042720300087. Accessed Nov 2021
27. Li Q, Yang Y, Wang W, Lee S, Xiao X, Gao X, Oztekin B, Fan C (2021) Mostafavi, a.: unraveling the dynamic importance of county-level features in trajectory of covid-19. Scient Reports 11(1):1–11
28. LLC G (2021) Google mobility report. https://www.google.com/covid19/mobility/. Accessed Sept 2021
29. Lu FS, Hattab MW, Clemente CL, Biggerstaff M, Santillana M (2019) Improved state-level influenza nowcasting in the United States leveraging Internet-based data and network approaches. Nat Commun 10(1):147. https://doi.org/10.1038/s41467-018-08082-0
30. Malki Z, Atlam ES, Hassanien AE, Dagnew G, Elhosseini MA, Gad I (2020) Association between weather data and covid-19 pandemic predicting mortality rate: Machine learning approaches. Chaos, Solitons & Fractals 138:110137
31. McGough SF, Johansson MA, Lipsitch M, Menzies NA (2020) Nowcasting by Bayesian Smoothing: A flexible, generalizable model for real-time epidemic tracking. PLOS Comput Biol 16(4):1–20. https://doi.org/10.1371/journal.pcbi.1007735. Publisher: Public Library of Science

32. Meo S, Alhowikan A, Al-Khlaiwi T, Meo I, Halepoto D, Iqbal M, Usmani A, Hajjar W, Ahmed N (2020) Novel coronavirus 2019-ncov: prevalence, biological and clinical characteristics comparison with sars-cov and mers-cov. Eur Rev Med Pharmacol Sci 24(4):2012–2019
33. Muhammad LJ, Algehyne EA, Usman SS, Ahmad A, Chakraborty C, Mohammed IA (2020) Supervised machine learning models for prediction of COVID-19 infection using epidemiology dataset, vol 2. https://doi.org/10.1007/s42979-020-00394-7
34. Organization WH (2020) Who coronavirus disease (covid-19) dashboard. https://covid19.who.int/. Accessed Oct 2021
35. Organization WH, et al. (2020) Diagnostic testing for sars-cov-2: interim guidance 11 september 2020. Tech. rep,. World Health Organization
36. Pan L, Mu M, Yang P, Sun Y, Wang R, Yan J, Li P, Hu B, Wang J, Hu C, et al. (2020) Clinical characteristics of covid-19 patients with digestive symptoms in hubei, china: a descriptive, cross-sectional, multicenter study. American Journal of Gastroenterology, 115
37. Petropoulos F, Makridakis S (2020) Forecasting the novel coronavirus covid-19. Plos One 15 (3):e0231236
38. Ren J, Yan Y, Zhao H, Ma P, Zabalza J, Hussain Z, Luo S, Dai Q, Zhao S, Sheikh A, Hussain A, Li H (2020) A novel intelligent computational approach to model epidemiological trends and assess the impact of non-pharmacological interventions for covid-19. IEEE J Biomed Health Inform 24 (12):3551–3563. https://doi.org/10.1109/JBHI.2020.3027987
39. Rodriguez JD, Perez A, Lozano JA (2009) Sensitivity analysis of k-fold cross validation in prediction error estimation. IEEE Trans Pattern Anal Mach Intell 32(3):569–575
40. Roosa K, Lee Y, Luo R, Kirpich A, Rothenberg R, Hyman J, Yan P, Chowell G (2020) Real-time forecasts of the covid-19 epidemic in China from february 5th to february 24th, 2020. Infect Dis Model 5:256–263
41. Russell CD, Millar JE, Baillie JK (2020) Clinical evidence does not support corticosteroid treatment for 2019-ncov lung injury. The Lancet 395(10223):473–475
42. Shi P, Dong Y, Yan H, Zhao C, Li X, Liu W, He M, Tang S, Xi S (2020) Impact of temperature on the dynamics of the covid-19 outbreak in china, vol 728. https://doi.org/10.1016/j.scitotenv.2020.138890. https://www.sciencedirect.com/science/article/pii/S0048969720324074. Accessed Sept 2021
43. Shinde GR, Kalamkar AB, Mahalle PN, Dey N, Chaki J, Hassanien AE (2020) Forecasting models for coronavirus disease (covid-19): a survey of the state-of-the-art. SN Computer Sci 1(4):1–15
44. da Silva RG, Ribeiro MHDM, Mariani VC, dos Santos Coelho L (2020) Forecasting brazilian and american covid-19 cases based on artificial intelligence coupled with climatic exogenous variables. Chaos, Solitons & Fractals 139:110027
45. Soheily-Khah S, Wu Y (2019) A novel feature engineering framework in digital advertising platform. 10,21. https://doi.org/10.5121/ijaia.2019.10403
46. Vannabouathong C, Devji T, Ekhtiari S, Chang Y, Phillips SA, Zhu M, Chagla Z, Main C, Bhandari M (2020) Novel coronavirus covid-19: current evidence and evolving strategies. J Bone Joint Surg American 102(9):734
47. Wang Z, Yang B, Li Q, Wen L, Zhang R (2020) Clinical features of 69 cases with coronavirus disease 2019 in wuhan, China Clinical infectious diseases
48. Wu JT, Leung K, Leung GM (2020) Nowcasting and forecasting the potential domestic and international spread of the 2019-ncov outbreak originating in wuhan, china: a modelling study. The Lancet 395 (10225):689–697. https://doi.org/10.1016/S0140-6736(20)30260-9
49. Xavier AR, Silva JS, Almeida JPC, Conceição JFF, Lacerda GS, Kanaan S (2020) Covid-19: clinical and laboratory manifestations in novel coronavirus infection. Jornal Brasileiro de Patologia e Medicina Laboratorial, 56
50. Yan L, Zhang HT, Goncalves J, Xiao Y, Wang M, Guo Y, Sun C, Tang X, Jing L, Zhang M, Huang X, Xiao Y, Cao H, Chen Y, Ren T, Wang F, Xiao Y, Huang S, Tan X, Huang N, Jiao B, Cheng C, Zhang Y, Luo A, Mombaerts L, Jin J, Cao Z, Li S, Xu H, Yuan Y (2020) An interpretable mortality prediction model for COVID-19 patients, vol 2. https://doi.org/10.1038/s42256-020-0180-7
51. Zoabi Y, Deri-Rozov S, Shomron N (2021) Machine learning-based prediction of COVID-19 diagnosis based on symptoms, vol 4. https://doi.org/10.1038/s41746-020-00372-6. https://www.nature.com/articles/s41746-020-00372-6. Accessed Oct 2021