Data-Driven Inference of COVID-19 qRT-PCR Results

Abstract

Background: The ultimate clue for the diagnosis of COVID-19 is a molecular test based on qRT-PCR, which is expensive, time-consuming and requires a specialized infrastructure for its application. In the face of medical suspicion of COVID-19, an alternative is to perform the analysis of existing clinical information using data-driven methods.

Aim of the study: In this research, we assess at which level of performance one can apply a machine learning method to infer the qRT-PCR positive/negative molecular results out of the clinical data of patients.

Methods: For the selection of characteristics, we use the algorithm of Boruta, a feature selector wrapper method based on Random Forest. Boruta generates an importance score that allows quantifying the relative importance of the descriptors. Using these features, we train a Random Forest classifier. We evaluate the performance of the classifier using *Precision-Recall* and ROC curves, and establish the ranges at which risk assessment permits effective decision-making.

Results: Using publicly available data set forth by the Mexican government, the Boruta algorithm obtained 20 features deemed important out of 23 initial ones. Cross-validation showed that the area under the ROC and Precision-Recall curves was $0.74 \pm 3.9 \times 10^{-3}$ and $0.43 \pm 7.2 \times 10^{-3}$, respectively, at one standard deviation.

Conclusion: As testing becomes critical in the back-to-normality phase of the COVID-19 pandemic, data-driven methods may result in fast, reliable, and inexpensive alternatives to support effective decision-making. This study offers bounding limits, which are above pure chance.

Keywords: Feature Selection, COVID-19, Confirmed/Discarded Discrimination.

1 Introduction

COVID-19 is an infectious disease caused by the SARS-CoV-2 virus. The most frequent symptoms at the beginning are fever, cough, and fatigue. Other signs that appear afterward include sputum production, myalgia, headache, hemoptysis, diarrhea, dyspnea, and lymphopenia [1]. Clinical diagnosis allows advancing in the detection of suspicious cases. However, the definitive diagnosis of COVID-19 is made by quantitative reverse-transcription polymerase chain reaction (qRT-PCR) assays, which are molecular techniques that consist of obtaining a large number of copies of a nucleic acid fragment. In principle, amplification makes it possible to identify or rule out SARS-CoV-2 in the sample with a high probability, requiring, on average, 5.2 copies of the envelope protein marker of the E genome and 3.8 copies of the RdRp genome polymerase for a probability detection rate of 95% [2]. Although the World Health Organization (WHO) has established the protocols to examine the samples, few laboratories have the necessary infrastructure to carry out the tests, the qRT-PCR study takes a couple of days under optimal conditions, and the cost of the equipment and reagents can be high. These conditions offer a formidable obstacle to a) the massive application of tests to determine the levels of immunology that persist; b) the assessment of whether it is feasible to carry out daily activities in areas that may have been exposed to infection; and c) to prioritize among those who receive vaccines, once these are available. Therefore, alternative, economical, fast, accessible, and robust forms are required to obtain information to support decision-making.

As the amount of information available to us increases, machine learning emerges as an alternative that provides supplemental information to diagnose the disease. Its response speed is its strength in the presence of exponential growth in the number of infections. Thus, the machine

learning community is responding through methodologies that allow predicting the change in the rate of infection over time [3], drug discovery [4], genome classification [5], and the prediction of patient survival [6]. Still, machine learning methods require adequate assessment of its individual components. Consider the problem of feature selection, which consists on finding a subset of those available predictors that serve to improve modeling. Having a reduced number of characteristics allows simplifying the interpretation of the model, reducing training time, avoiding overfitting, and reducing the amount of data necessary to classify the observations. The exhaustive solution to the problem is computationally challenging. Just consider that given n characteristics, one would need to verify the performance for groups of $i = 1, 2, \ldots, n$ of them, which is prohibitive even in relatively small cases. For example, if we have $n = 20$ characteristics and want to select $m = 10$, we would have a number of subsets in the order of 10^5. In other words, exhaustively testing all combinations is impractical even with powerful computers.

2 Dataset Preparation

In our study, we took the open dataset made public by the Mexican Health Minister corresponding to the government records related to COVID-19 [7]. A database record contains 35 variables, some of them associated with administrative issues, such as the registration number, or post-diagnosis treatment of COVID-19, such as whether a hospital admits the patient to the Intensive Care Unit (ICU). As a pre-process, we remove these features to leave only those that will make up a pre-diagnostic clinical verification list. We interpret the dates as days from the beginning of the symptoms. In our analysis, we focused attention on patients whose results for COVID-19 as been declared positive or negative. For our purposes, a positive case occurs when the qRT-PCR test is positive and negative otherwise. We remove from the set of predictors those for which several records, a number more significant than a certain threshold τ, have not been completed. Finally, we attribute values to those predictors that still have uncaptured entries. To do this, we use the library `missForest` in `R`, which trains a Random Forest [8] on the observed values of the data matrix to predict the missing values. The library can complete values for continuous or categorical variables, even in nonlinear relationships.

Before begining to process the data, we proceed to gain some intuition about how the nature of the classification problem using visualization of the data. Figure 1 shows the variables described in Appendix A. Note the large overlap along the marginals corresponding to each predictor. Also, note that this representation may hinder the observation of effects related to joint variables or conditional distributions questions about the data that the classifier may potentially exploit.

3 A Data-Driven Classifier

Given the database, in our study, we carry out an analysis to select the variables that best distinguish between confirmed and discarded COVID-19 patients. To do this, we use the implementation of the `Boruta` algorithm [9] contained in `R`, a wrapper-type method based on the Random Forest classifier [8]. Many other selection methods consider all the characteristics at the same time, such as Mutual Information and Super Casual Correlation [10]. However, we selected Boruta's algorithm because it considers multiple associations at the same time, without an exhaustive search. In wrapper-type feature selection approaches, the methods consider a feature important when its removal degrades the performance of the classifier. However, when a feature does not degrade the classifier performance, it cannot be considered unimportant [11]. Boruta's algorithm begins by defining shadow variables for each predictor. Then, the method shuffles the values of the shadow variable over all the observations of the same predictor. Boruta's algorithm compares the relative importance of the original variables and the shadow variables. If the importance value of an original

variable is statistically higher/lower than the maximum importance of the best-valued shadow variable, the method labels the predictor as important/not important. Then, Boruta discards the tagged variables and repeats the procedure until it tags all the variables. The Random Forest algorithm naturally handles a measure of importance. For a tree, the importance of a variable is related to the increase in precision achieved with the partitions made in the nodes, typically measured in terms of the Gini index or the entropy/information gain. For forest, the method averages the importance of a predictor over all trees. However, due to the haphazard nature of the Random Forest algorithm, each run may result in a different selection. To attack this problem, we run the Boruta algorithm k times, recording the characteristics and relative importance assigned to them. With each of the resulting subsets of attributes, one could construct a classifier, remaining to identify which offers the best generalization capacity.

To do this, using the characteristics considered important, we built a new classifier based on Random Forests. Random Forest are special extension of Decision Trees where one trains a particular tree with a subset of the datatset available, the algorithm selects a random subset of the predicts at each split, and one creates a large number of trees, naturally constructing an ensemble of classifiers. Like other tree-based methods, Random Forest handle equally well regression and classification problems and are equally well adapted to deal with numerical and categorical variables. Furthermore, the independence of the ensemble reduces the variance during voting for classification or averaging for regression.

During the learning process, we split the data into training and validation sets. An issue that we must address is the class imbalance in our data, although Random Forest classifiers are adept to this issue [12]. In general, the methods to balance the classes either undersample or oversample [13]. In the former case, we obtain a portion of the larger class. In the latter, we sample the smaller class with replacement. In both cases, we end up with the same number of samples for the positive and negative class. We repeat this operation l times for each classifier, in a process known as cross-validation, each time performing a partition with the same percentage, but a random subset of the data. Next, for each classifier, we evaluated performance using ROC curves and *Precision-Recall*. As a result of repeated performance evaluation, we have variations that we can summarize statistically. Finally, a plausible selection criterion could be the selection of the classifier that provides the maximum among the minimum performance values during cross-validation.

4 Experimental Results

In our experiments, we determine the ability of the feature selector to produce relevant descriptors and reliable classifiers. We took 49,570 records of the Mexican Health Minister open dataset [7], corresponding to the government archives from January 1, 2020, to April 19, 2020, containing 8,261 positive and 31,170 negative cases to COVID-19 after a qRT-PCR test. The computer we use to process the information runs on the Windows 8.1 operating system and consists of a 64-bit Intel i7-3770 processor at a clock frequency of 3.4Hz, and operating with 16MB of RAM. We developed our computer programs in `R` version 3.6.3.

After removing the administrative related features, and those with more than $\tau = 1$ *missing values*, the number of predictors was 23 (see Appendix A). As an illustration, the execution of Boruta's algorithm on our dataset results in the selection of 20 critical characteristics (see Figure 2). The remaining ones turned out to be not important to distinguish between confirmed and discarded cases, given the suspicion of COVID-19 and with the considered predictors. We repeated the selection of features 10 times, and noticed that the same features consistently appear in the selector while predictors 13 (`asthma`), 22 (`migrant`) and 23 (`country.origin`) were always rejected as unimportant.

With each of these features, one could construct a classifier, remaining to identify its performance measured as its generalization capacity. To do this, we first fine-tune the hyperparameters

sampling	AUC *Precision-Recall*	AUC ROC
undersampling	$0.42 \pm 8 \times 10^{-3}$	$0.75 \pm 3.5 \times 10^{-3}$
original dataset	$0.43 \pm 7.2 \times 10^{-3}$	$0.74 \pm 3.2 \times 10^{-3}$
oversampling	$0.39 \pm 6.5 \times 10^{-3}$	$0.74 \pm 2.9 \times 10^{-3}$

Table 1: Performance result for the Random Forest Classifier. Here, we compare three different strategies for class balancing.

corresponding to the number of variables randomly sampled at each split and the number of trees in the random forest. We use ten-fold cross-validation, for the former, and grid search, for the latter, resulting in 4 (four) for the number of variables and 1,500 for the number of trees. Using these hyperparameters and the characteristics considered significant, we built a classifier based on Random Forests. For learning, we split the data into 50% for training and 50% for validation. To balance the classes, we observed the performance for undersampling, oversampling, and learning with the original training datset split. We repeat this process 30 times, each time performing a partition with the same percentage, but using uniform random sampling. Next, for the classifier, we evaluated its performance using ROC and *Precision-Recall* curves. As a result of repeated performance evaluation, we have variations that we summarize in Table 1. With $0.75 \pm 3.5 \times 10^{-3}$, the ROC Area under the Curve (AUC) is largest for undersampling, while at $0.43 \pm 7.2 \times 10^{-3}$ the *Precision-Recall* AUC is largest using r the original training partition. In these results, we express uncertainty at one standard deviation.

Figure 3 shows a performance level in the AUC of $0.74 \pm 3.9 \times 10^{-3}$ and $0.92 \pm 1.8 \times 10^{-3}$ for the ROC and *Precision-Recall* curves, respectively, at one standard deviation. An intriguing question corresponds to the limits of performance, *i.e.*, the decision levels at which the classifier makes no mistakes. These values define the levels at which we are not letting people infected with COVID-19 to leave the hospital unchecked or not retaining people at the hospital without need as the qRT-PCR test will come negative. For each iteration, we computed the minimum value of the precision at which the recall was still one, *i.e.*, the number of false negatives is zero. The result was 0.19 $\pm 1.5 \times 10^{-3}$, with the uncertainties expressed at one standard deviation[1].

Conclusion

Through a classifier, one may establish the mapping between input characteristics and an output label employed to assign the observation to a class. Given a preliminary diagnosis by a physician expressing suspicion about the occurrence of COVID-19, this document shows that it is feasible to construct a Random Tree classifier to distinguish between patients who will be confirmed/discarded by a qRT-PCR-based molecular test. Also, although characteristics associated with COVID-19 are now widely known [6], our research quantifies the importance of each of them. Likewise, the family of classifiers presented here offers the ability to assign a measure of certainty regarding the diagnosis. A decision-maker may set a threshold for assignment to one or the other class balancing risks and costs. Decision-makers may play out these tradeoffs using the performance curves as they permit to know the decision thresholds that allow assigning a patient with COVD-19 when she/he has not been subject to a molecular test, but perhaps more importantly, permit to avoid rejecting an infected patient without proper care. Our tool can thus have a critical complementary value for clinical diagnosis.

Furthermore, nowadays, there is an increasing interest in defining a back-to-normality path forward. Most of the options available rely on extensive testing, which is complicated by the cost

[1]To facilitate the confirmation of our findings and improve on our framework, we are making our code available at *github. com/ joaquinsalas/ COVID19-DataDriven-Classifier*

implied, delayed results, and reliability. In this research, we show that a data-driven approach may offer an alternative that is fast, inexpensive and error bounded.

A Data Dictionary

The features used in our analysis to identify the result of the analysis of the sample reported by the laboratory of the National Network of Epidemiological Surveillance Laboratories include [7]:

ID	Name	Concept
1	`origin`	Medical surveillance is carried out through the Respiratory Disease Monitoring Health Unit System (USMER) or not.
2	`sector`	Type of institution of the National Health System that provided care.
3	`state.med.unit`	State code location for the medical unit.
4	`sex`	Genre of the patient.
5	`state.res`	Patient's residence state name.
6	`patient.type`	Type of care the patient received in the unit, *i.e.*, outpatient or inpatient.
7	`num.days`	Number of days between first symptoms and the patient's admission to the care unit.
8	`pneumonia`	Whether the patient was diagnosed with pneumonia.
9	`age`	Age of the patient in years.
10	`pregnant`	Whether the patient is pregnant.
11	`diabetes`	Identifies whether the patient has a diagnosis of diabetes.
12	`COPD`	Whether the patient has a diagnosis of Chronic Obstructive Pulmonary Disease.
13	`asthma`	Whether the patient has a diagnosis of asthma.
14	`immunosuppression`	Whether the patient has immunosuppression.
15	`hypertension`	Whether the patient has a diagnosis of hypertension.
16	`other.diseases`	Whether the patient is diagnosed with other diseases.
17	`cardiovascular`	Whether the patient has a diagnosis of cardiovascular disease.
18	`obesity`	Identifies whether the patient is diagnosed with obesity.
19	`chronic.kidney`	Whether the patient is diagnosed with chronic kidney failure.
20	`smoking`	Whether the patient has a smoking habit.
21	`contact`	Whether the patient had contact with another case diagnosed with SARS CoV-2.
22	`migrant`	Whether the patient is a migrant.
23	`country.origin`	Identifies the country from which the patient left for Mexico.

References

[1] Hussin Rothan and Siddappa Byrareddy. The Epidemiology and Pathogenesis of Coronavirus Disease (COVID-19) Outbreak. *Journal of Autoimmunity*, page 102433, 2020.

[2] Victor Corman, Olfert Landt, Marco Kaiser, Richard Molenkamp, Adam Meijer, Daniel Chu, Tobias Bleicker, Sebastian Brünink, Julia Schneider, and Marie Schmidt. Detection of 2019 Novel Coronavirus (2019-nCoV) by Real-Time RT-PCR. *Eurosurveillance*, 25(3), 2020.

[3] Samir Bandyopadhyay and Shawni Dutta. Machine Learning Approach for Confirmation of COVID-19 Cases: Positive, Negative, Death and Release. *medRxiv*, 2020.

[4] Yiyue Ge, Tingzhong Tian, Sulin Huang, Fangping Wan, Jingxin Li, Shuya Li, Hui Yang, Lixiang Hong, Nian Wu, Enming Yuan, et al. A Data-Driven Drug Repositioning Framework Discovered a Potential Therapeutic Agent Targeting COVID-19. *bioRxiv*, 2020.

[5] Gurjit Randhawa, Maximillian Soltysiak, Hadi El-Roz, Camila de Souza, Kathleen Hill, and Lila Kari. Machine Learning using Intrinsic Genomic Signatures for Rapid Classification of Novel Pathogens: COVID-19 Case Study. *bioRxiv*, 2020.

[6] Li Yan, Hai Zhang, Yang Xiao, Maolin Wang, Chuan Sun, Jing Liang, Shusheng Li, Mingyang Zhang, Yuqi Guo, and Ying Xiao. Prediction of Survival for Severe COVID-19 Patients with Three Clinical Features: Development of a Machine Learning-based Prognostic Model with Clinical Data in Wuhan. *medRxiv*, 2020.

[7] Datos Abiertos: Información Reference a Casos COVID-19 en México. `https://tinyurl.com/mexico-covid`. Accessed: 2020-04-20.

[8] Leo Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001.

[9] Miron Kursa and Witold Rudnicki. Feature Selection with the Boruta Package. *Journal of Statistical Software*, 36(11), 2010.

[10] E. Tuv, A. Borisov, G. Runger, and K. Torkkola. Feature Selection with Ensembles, Artificial Variables, and Redundancy Elimination. *Journal of Machine Learning Research*, 10:1341–1366, 2009.

[11] R. Nilsson, J. Peña, J. Björkegren, and J. Tegnér. Consistent Feature Selection for Pattern Recognition in Polynomial Time. *Journal of Machine Learning Research*, 8(Mar):589–612, 2007.

[12] Byron Wallace, Kevin Small, Carla Brodley, and Thomas Trikalinos. Class Imbalance, Redux. In *International Conference on Data Mining*, pages 754–763. IEEE, 2011.

[13] Harsurinder Kaur, Husanbir Singh Pannu, and Avleen Kaur Malhi. A Systematic Review on Imbalanced Data Challenges in Machine Learning: Applications and Solutions. *ACM Computing Surveys*, 52(4):1–36, 2019.

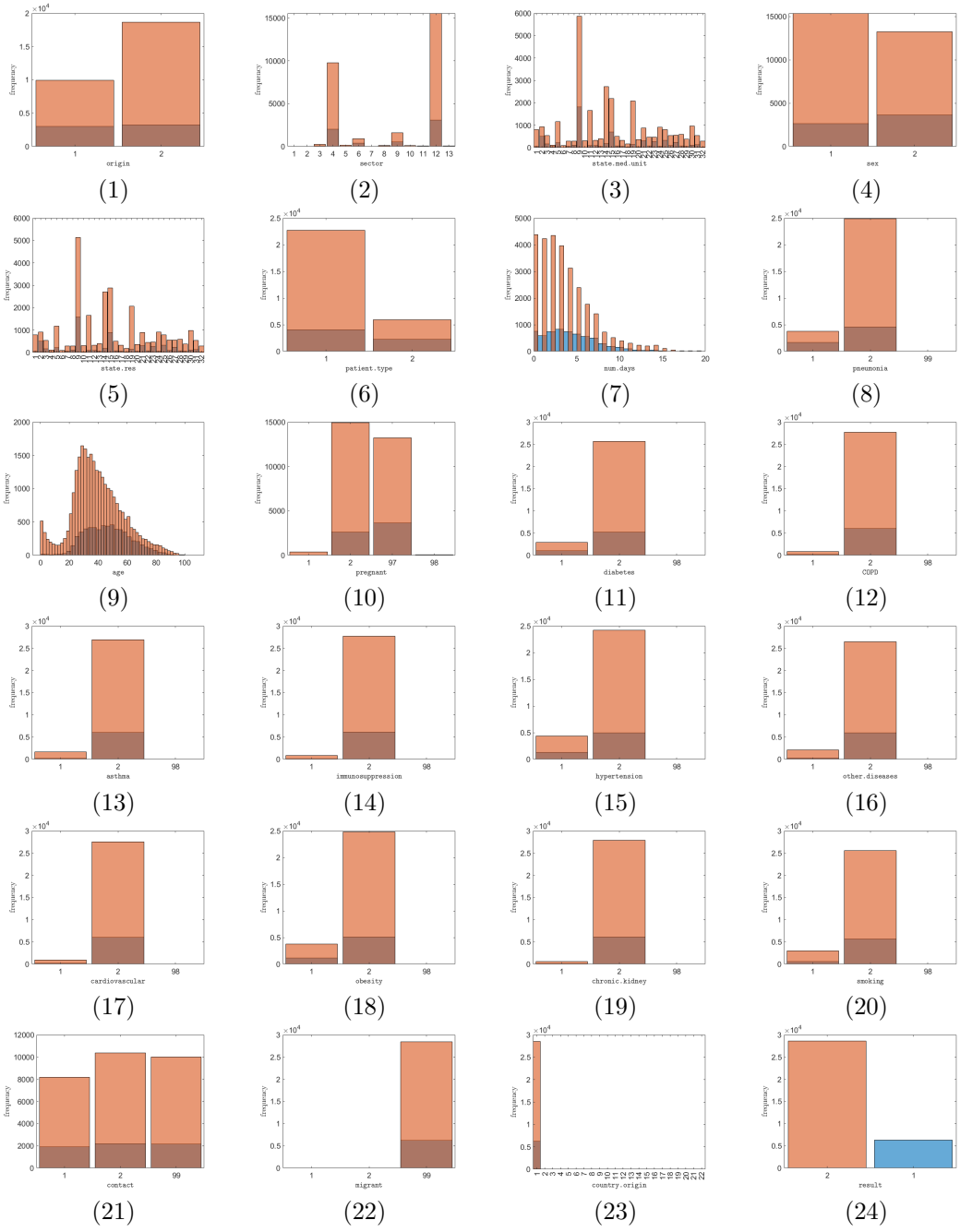


Figure 1: Distribution of predictors. Class positive is blue, negative is brown and overlapp is darker brown. The label number corresponds to the feature number described in Appendix A. Subfigure (24) corresponds to the result of the qRT-PCR molecular test.

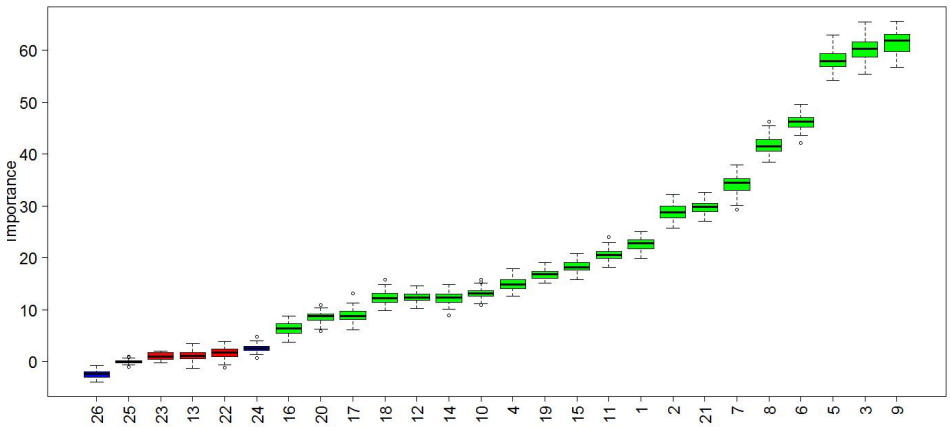
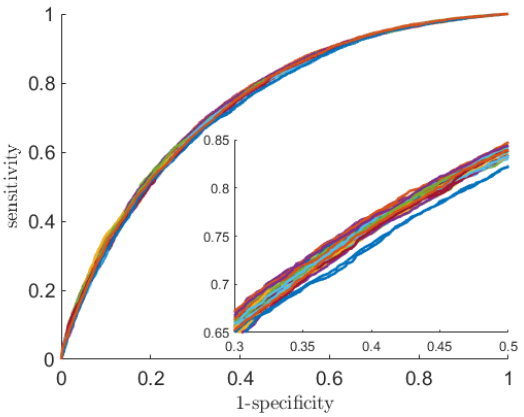
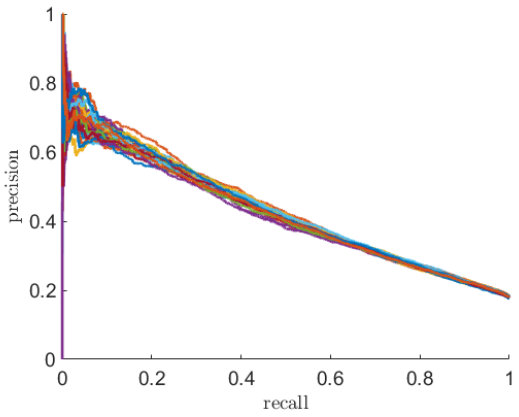


Figure 2: Feature Selection. Important characteristics in the process of discriminating between confirmed and discarded cases for COVID-19 using the Boruta algorithm (best seen in color, see Appendix A for the data dictionary).



(a) ROC, AUC $= 0.74 \pm 3.9 \times 10^{-3}$



(b) *Precision-Recall*, AUC $=0.4253 \pm 7.2 \times 10^{-3}$

Figure 3: Evaluation of performance. The ROC and *Precision-Recall* curves summarize combinations of elements in the confusion table at different threshold values.