Machine Learning for Prognosis of Patients with COVID-19: An Early Days Analysis

José Solenir L. Figuerêdo[1], **Renata F. Araújo-Calumby**[2], **Rodrigo T. Calumby**[1]

[1]Department of Exact Sciences, University of Feira de Santana
Feira de Santana – BA – Brazil

[2]College of Pharmacy, Feira de Santana Higher Education Unit
Feira de Santana – BA – Brazil

jslfigueredo@ecomp.uefs.br, farm.renata@hotmail.com, rtcalumby@uefs.br

Abstract. *This work proposes a machine learning approach to predict the prognosis of patients with COVID-19. To assist in this task, a descriptive analysis and relative risk estimation were performed. In addition, the importance of variables in the perspective of machine learning algorithms was computed and discussed. The experiments were performed with large-scale nation-wide dataset from Brazil. The results reveal that the model developed was able to predict the patient's prognosis with an AUC = 0.8382. The results also point out that the chance of death is greater among patients over 60 years old, with comorbidities, and symptoms such as dyspnea and Oxygen saturation ($< 95\%$), confirming results observed in other regions of the world.*

1. Introduction

Outbreaks of the COVID-19 epidemic have caused health problems worldwide, since its discovery in December 2019. Causing different symptoms such as fever, cough, fatigue, and mild respiratory diseases to serious complications, in many cases leads the patient the death [Yan et al. 2020]. Initially detected in the city of Wuhan, China, the virus, later called SARS-CoV-2 (Severe Acute Respiratory Syndrome Coronavirus), quickly spread around the world, leading the World Health Organization (WHO) to declare the state pandemic on March 11, 2019. Identified by researchers as a binuclear virus that has a broad clinical spectrum of infection [Wang et al. 2020a, Wang et al. 2020b], it has so far claimed millions of victims worldwide.

To date (July 2021), more than 196 million people have been infected and more than 4.2 million have died of COVID-19 worldwide. Considering the American continent, more than 76,788,166 cases and 2 million deaths were registered [WHO 2021b]. In this region, Brazil is the second most affected country, behind only the United States. Brazil has a total of 19,938,358 infected cases, with 556,834 deaths [Brazil 2021]. Besides, currently the country still has a high number of daily deaths. Despite the Brazilian situation, other regions of the globe have shown signs of improvement, evidenced mainly by the reduction in the number of daily cases. However, there are still challenges attached, like other possible waves [Leung et al. 2020] and emerging [Pollet 2020, Nania 2021], new virus variants [Mahase 2021, WHO 2021a], or even future pandemics. In this context, the doubts and uncertainties that still surround COVID-19 have imposed great challenges on health systems. One of these challenges is related to the care rationing process, which

is closely related to clinical decision-making. In this sense, discussing and understanding strategies that support the decision-making process about care rationing is essential, both considering the current scenario, but also as a way of preparing for other possible pandemics.

Care rationing reveals the need for a screening process for infected patients [White and Lo 2020]. This screening process can directly or indirectly influence lethality and mortality rates. In some countries, these rates have been quite expressive, as is the case in Brazil. Taking July 2021, Brazil had a lethality rate of 2.8%, with a mortality rate of 265 per 100,000 inhabitants [Brazil 2021]. The reduction of these rate could be achieved through different strategies, with a sophisticated screening process being one of the possible alternatives. This process could make use of effective prognostic biomarkers, for example. Although stratification is not the most ideal strategy, it is an alternative that is justified due to the scarcity of hospital resources, whether human or technical, and the possibility of better therapeutic definition, especially in time of pandemic. However, for this to be possible, the identification of reliable predictors of patient mortality is necessary. To contribute this, the current literature has already identified several clinical characteristics associated with the severity of COVID-19 infection. Therefore, these characteristics could be used to make a prognosis, especially based on solutions using Artificial Intelligence (AI). The result of this prediction would assist the professional in the decision-making process, regarding the best referral to be applied to a given patient, considering multiple criteria, among them, the availability of hospital resources.

In recent years, systems developed using AI have been applied in different areas, including health. In this one, AI systems find a wide spectrum of applications, for example, systems to assist in the diagnosis and prognostication of many conditions [Masood et al. 2018, Silva et al. 2020, Sousa et al. 2020]. In particular, these systems are often developed using a subfield of AI called machine learning. But the development of these systems is not restricted to this subfield, more advanced techniques such as deep learning can also be employed. With regard to COVID-19, recent studies have shown that machine learning can be applied, in different ways, to deal with the pandemic. The data collected, regardless of its nature, whether textual or visual, can be integrated and analyzed by different machine learning algorithms. Among the tasks that have explored its use, we have the taxonomic classification of COVID-19 genomes [Randhawa et al. 2020], COVID-19 detection based on computed tomography images (CT) [Ozkaya et al. 2020], monitoring of COVID-19 using IoT [Barbosa et al. 2020] and patient survival prediction [Yan et al. 2020]. Each task has its particularities and limitations that must be overcome. But, in a particular way, our work will deal exclusively with the latter case.

Thus, the aim of this work was the development of models to support the prognosis for patients with COVID-19, through machine learning algorithms. For this, data from patients in Brazil were used, considering the initial months of the pandemic in Brazil. We consider a set of predictive variables of basic individual information, such as gender and age group, the occurrence of symptoms, presence of comorbidities, among others. In addition to the development of predictive models, a descriptive analysis of the data was also carried out, as well as an odds ratio analysis. These analysis intended to identify possible markers of disease severity, as well as understanding the profile of people who died in the initial months of the pandemic. With a similar objective, an alternative procedure is

performed using a technique for variable importance estimation.

The remainder of this article is organized as follows: Section 2 presents the related works and Section 3 describes the experimental process. The results and discussions are presented in Section 4. Finally, Section 5 brings the conclusions and future work.

2. Related Works

So far, several studies have been developed in order to contribute to facing the pandemic. These are researches with different purposes, but all aiming develop tools to help fight the pandemic. Specifically, our work is focused on a particular type of study, which is the prediction of prognosis of patients. A peculiar system like this can find different uses in a hospital environment, especially in the decision-making process. For example, this system could be used to prioritize patients with a more serious condition, that is, patients who are more likely to die. For this purpose, some works have been developed [Yan et al. 2020, Xie et al. 2020, Souza et al. 2020].

In [Yan et al. 2020], the authors analyzed epidemiological, demographic, clinical and laboratory data from 485 patients with COVID-19. The study aimed to identify patients at high risk, through predictive biomarkers, in order to improve the prognosis, in order to reduce patient mortality. For this, the authors applied the XGBoost algorithm. The results showed that the model developed was able to select three biomarkers that predict the mortality of individual patients with an accuracy above 90%, namely: lactic dehydrogenase (LDH), highly sensitive C-reactive protein and lymphocytes.

Considering a survival analysis scenario, in [Xie et al. 2020], the authors developed a retrospective study seeking to identify predictive variables that would assist in predicting hospital mortality. The study was conducted with 140 patients who had pneumonia associated with COVID-19, and who used oxygen supplementation. Among the conclusions is the fact that hypoxemia is independently associated with in-hospital mortality. In addition, they found that oxygen saturation (S_pO_2) values greater than 90% with oxygen supplementation indicate a high probability of survival. As indicated by the researchers, these results may help guide the clinical management of patients with severe COVID-19, particularly in settings requiring strategic allocation of limited critical care resources [Xie et al. 2020].

The authors in [Souza et al. 2020] carried out a study similar to those previously mentioned. However, instead of using data from patients in the city of Wuhan, they used data from the state of Espirito Santo (Brazil). In addition to the difference in the location where the study was carried out, there were particular differences in the expressiveness of the data, such as the absence of data regarding laboratory tests. Souza *et al.* applied a set of machine learning algorithms to determine survival in patients with COVID-19. For that, the study dataset comprises information of 13,690 patients concerning closed cases due to cure or death. Considering the best model, the authors showed that the outcome by COVID-19 can be predicted with an AUC of 0.92. From the results achieved, the authors conclude that machine learning techniques can assist in the prognostic prediction and physician decision-making, allowing a faster response and contributing to the non-overload of healthcare systems [Souza et al. 2020]

The present study, similarly to those already mentioned, aims to determine the survival for patients with COVID-19. However, considering the absence of works with

large scale data and in-depth feature analysis, it presents significant contributions and innovations. For instance, [Yan et al. 2020] and [Xie et al. 2020] were conduct in a city-level (Wuhan, China). On the other hand, considering the work of [Souza et al. 2020], although the study was carried out in Brazil, a limited database was used, containing only data from the state of Espírito Santo. This indicates the need for a more in-depth study, seeing that regional variations may occur in a country with continental characteristics such as Brazil, as highlighted by the authors themselves [Souza et al. 2020]. Moreover, in a machine learning scenario, a deep understanding of variables importance is a demand to properly support data gathering, model development and, ultimately, disease treatment. Our large-scale prediction model and an in-depth analysis are described in the next sections.

3. Experimental Process

The experimental process carried out in this work is illustrated in Figure 1. There are four steps: data collection, data preprocessing, training (optimization and validation) and, finally, the evaluation and analysis of the model.

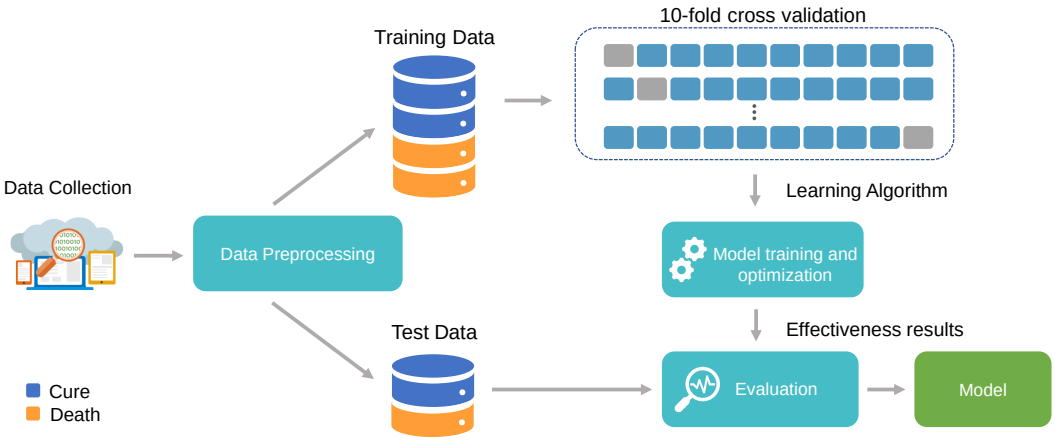

Figure 1. Experimental process used in this study.

3.1. Dataset

The experiments were conducted using the *Severe Acute Respiratory Syndrome Syndrome Database* (SRAG2020). Such data is publicly available on the OpenDATASUS portal [1]. This anonymized database is the result of an initiative by the Ministry of Health of Brazil, through the Secretariat of Health Surveillance (SHS), which develops surveillance for Severe Acute Respiratory Syndrome (SARS) in Brazil since 2009, due to the Influenza A(H1N1)pdm09 pandemic.

More recently, with the new Coronavirus pandemic, data from COVID-19 has been incorporated into the surveillance network. This data is updated frequently as new information becomes available. The version of the dataset used here refers to COVID-19 cases counted in the initial months of the pandemic in Brazil, specifically up to August 31, 2020.

[1]https://opendatasus.saude.gov.br/dataset/bd-srag-2020 As of August 02, 2021

3.2. Data Preprocessing

As the main objective is to predict the patient's prognosis, only completed cases are used (death or cure). In addition to information regarding the evolution of the infection, the dataset also includes basic individual information, such as gender and age group, symptoms, comorbidities, among others. After carrying out the aforementioned selection, the database included 274,493 patient records: 164,535 (59.94%) of cure and 109,958 (40.06%) of death.

The original database includes 156 attributes that characterize patients. However, a preliminary analysis found that some of these characteristics represent only marginal information that do not add relevant prediction-oriented content (e.g., city, ZIP code). Therefore, some variables considered irrelevant for the task were removed, resulting in a final set of 39 variables (including the outcome attribute): age, gender, race, education, geographic area, dyspnea, fever, cough, Oxygen saturation ($< 95\%$), sore throat, respiratory discomfort, diarrhea, vomiting, other symptoms, heart disease, diabetes, neuropathy, pneumopathy, kidney disease, asthma, immunodepression, hemopathy, liver failure, down syndrome, postpartum, obesity, other comorbidities, hospitalization (Yes, No, Unknown), ICU, antiviral treatment, antiviral type, severe syndrome outbreak, nosocomial[2], bird or swine contact, pregnancy, risk factor (Yes, No), X-ray (Normal, Interstitial infiltrate, Consolidation, Mixed, Other, Unrealized, Unknown), ventilatory support (Yes, invasive; Yes, non-invasive; No; Unknown), and evolution.

Still in the preprocessing stage, it was detected that the database had a high percentage of missing data, including more than 60% of data missing for some of the attributes. Thus, in order to avoid possible inconsistencies in the experiments, data standardization was performed considering missing data as non-occurrence of the event in particular (e.g., for the cough attribute, if the information was missing, it was indicated as the patient not having this symptom). All variables are categorical, with the exception of "Age". Thus, this attribute was discretized as: child (0-10), teenager (11-17), young adult (18-29), average adult (30-40), adult (41-59) and elderly (60 or more).

3.3. Experimental Setup

The database was partitioned into training and test sets. Notice that before partitioning a sub-sampling was applied to balance the dataset (cures and deaths). In this process, random records belonging to the majority class (cure) were discarded. This procedure was used mainly to avoid possible bias in the training stage of the predictive models. Hence, a simple train-test partitioning was performed by means of a stratified random procedure. A sample of 30,000 patients was fixed for the test set and the rest (202,164 patients) were assigned to the training set.

The training set was used to build the models which were enhanced with hyperparameter optimization through grid search based on 10-fold cross validation and using stratified random sampling. In turn, the test set was used to verify the effectiveness of the models developed. This study used two algorithms, the Decision Tree and Naive Bayes. The hyperparameters tested for the Decision Tree can be seen in the Table 1. These algorithms were selected due to the ease in explaining the decisions made by the models

[2]Refers to infection acquired in the hospital.

Table 1. Hyperparameters used in the decision tree grid search.

Hyperparameters	Tested values	Best value
Quality Measure	Gini index, Gain Ratio	Gain Ratio
Pruning method	Without Pruning, MDL	Without Pruning
Minimum number of records per node	(10, 20, 30, 50,100)	50

learned, which are considered simple and explainable. This is a very desirable characteristic in critical contexts such as for health decision support systems. Additionally, a set of experiments was carried out in order to find the most important attributes from the perspective of machine learning models. For this, the removal of one attribute at a time and the verification of the model's effectiveness without that attribute was conducted.

3.4. Evaluation

The developed models were evaluated with the Receiver Operating Characteristic (ROC) curve its Area Under the Curve (AUC). The ROC curve corresponds to a graphic technique widely used to evaluate the effectiveness of binary classifiers from the application of different confidence thresholds [Han et al. 2011]. On the other hand, the AUC provides a single representative value of the overall effectiveness. The AUC varies between 0 and 1, with 0 for a model that performed all predictions erroneously, while 1 for models that performed 100% of the predictions correctly.

4. Results and Discussions

Table 2 presents a descriptive analysis for selected sample of attributes. The data corroborate with previous studies carried out in different regions of the world [Zhou et al. 2020, Grasselli et al. 2020, Richardson et al. 2020, Onder et al. 2020] which report that older age, male gender and the presence comorbidities were associated with hospitalization by COVID-19 and, therefore, can be used as potential risk factors. Additionally, the descriptive analysis revealed that symptoms related to breathing such as dyspnea and respiratory distress cause a higher percentage of deaths among patients who develop them. The results also show that, although Brazil is located in a different geographic region with climate and sociodemographic differences from where these first studies were carried out, the risk factors are similar.

4.1. Odds ratio analysis

In addition to the previous analysis, an estimate of the odds ratio of death was performed. For that, odds ratios were computed with a 95% confidence interval (Table 3). The attributes that indicate greater chances of death, due to the presence of the risk factor, are highlighted in bold. According to the results, the chance of death is greater among patients admitted to the ICU, aged 60 years or older or who used invasive ventilation support. Signs and symptoms such as Oxygen saturation, dyspnea and respiratory distress were also indicated as factors that increase the chances of death. In addition, comorbidities also indicate an increased chance of death. On the other hand, symptoms such as fever, cough, sore throat and diarrhea were identified as having the least contribution to the patient's death. Peculiarly, obesity and asthma, commonly indicated as risk factors,

Table 2. Descriptive analysis. Selected variables.

	Category	All n (%)	Cure n (%)	Death n (%)
All		274493 (100)	164535 (59.9)	109958 (40.1)
Age	0-35	31186 (11.7)	27607 (88.5)	3579 (11.5)
	36-59	99811 (36.7)	75543 (75.7)	24268 (24.3)
	60 or more	**143496 (52.3)**	**61385 (42.8)**	**82111 (57.2)**
Gender	Female	120023 (43.7)	73998 (61.7)	46025 (38.4)
	Male	**154410 (56.3)**	**90499 (58.6)**	**63911 (41.4)**
	Unknown	60 (0.02)	38 (63.3)	22 (36.7)
Dyspnea	**Yes**	**189309 (69.0)**	**106204 (56.1)**	**83105 (43.9)**
	No	51998 (19.0)	38924 (74.9)	13074 (25.1)
	Unknown	3405 (1.2)	1681 (49.4)	1724 (50.6)
	Missing	29781 (10.9)	17726 (59.5)	12055 (40.5)
Fever	**Yes**	**182306 (66.4)**	**115290 (63.2)**	**67016 (36.8)**
	No	58699 (21.4)	33958 (57.9)	24741 (42.2)
	Unknown	4227 (1.5)	1654 (39.1)	2573 (60.9)
	Missing	29261 (10.7)	13633 (46.6)	15628 (53.4)
Cough	**Yes**	**198174 (72.2)**	**125367 (63.3)**	**72807 (36.7)**
	No	45881 (16.7)	25784 (56.2)	20097 (43.8)
	Unknown	3718 (1.4)	1347 (36.2)	2371 (63.8)
	Missing	26720 (9.7)	12037 (45.1)	14683 (55.0)
Oxygen saturation	**Yes**	**150698 (54.9)**	**78931 (52.4)**	**71767 (47.6)**
	No	73057 (26.6)	55368 (75.8)	17689 (24.2)
	Unknown	5817 (2.1)	2905 (49.9)	2912 (50.1)
	Missing	44921 (16.4)	27331 (60.8)	17590 (39.1)
Respiratory Discomfort	**Yes**	**151419 (55.2)**	82588 (54.5)	**68831 (45.5)**
	No	71790 (26.2)	51870 (72.3)	19920 (27.8)
	Unknown	4871 (1.8)	2549 (52.3)	2322 (47.7)
	Missing	46413 (16.9)	27528 (59.3)	18885 (40.7)
Heart disease	**Yes**	**88770 (32.3)**	**44716 (50.4)**	**44054 (49.6)**
	No	48011 (17.5)	27768 (57.8)	20243 (42.2)
	Unknown	1871 (0.7)	841 (45.0)	1030 (55.1)
	Missing	135841 (49.5)	91210 (67.1)	44631 (32.9)
Neuropathy	**Yes**	**10641 (3.9)**	**4002 (37.6)**	**6639 (62.4)**
	No	93384 (34.0)	51964 (55.7)	41420 (44.4)
	Unknown	3448 (1.3)	1509 (43.8)	1939 (56.2)
	Missing	167020 (60.9)	107060 (64.1)	59960 (35.9)
Pneumopathy	**Yes**	**9991 (3.6)**	**3974** (39.8)	**6017 (60.2)**
	No	93526 (34.1)	51819 (55.4)	41707 (44.6)
	Unknown	3519 (1.3)	1552 (44.1)	1967 (55.9)
	Missing	167457 (61.0)	107190 (64.0)	60267 (36.0)
Homeopathy	**Yes**	**2210 (0.8)**	**1072 (48.5)**	**1138 (51.5)**
	No	98698 (36.0)	53785 (54.5)	44913 (45.5)
	Unknown	3672 (1.3)	1572 (42.8)	2100 (57.2)
	Missing	169913 (61.9)	108106 (63.6)	61807 (36.4)
Obesity	**Yes**	**11385 (4.2)**	**6385 (56.1)**	**5000 (43.9)**
	No	89595 (32.6)	48652 (54.3)	40943 (45.7)
	Unknown	5588 (2.0)	2553 (45.7)	3035 (54.3)
	Missing	167925 (61.18)	106945 (63.69)	60980(36.3)

did not indicate a greater chance of the patient dying. Here we highlight that the large amount of missing and ignored data may have introduced some bias on these results.

Specifically, with regard to "ICU" and "Ventilatory Support", it is worth mentioning that the fact that they are among the characteristics that increase the chances of patient death does not necessarily mean that it was the cause of death. They appear possibly because patients who require the use of these resources typically represent more severe cases to which other risk factors may be associated.

Still in relation to the "ICU" and the "Ventilatory Support" it is possible to speculate on the possible factors that may be associated with an increased chance of patient death. The results may have been motivated by two main characteristics: i) the intervention occurs lately, that is, when the case is already in a severe stage of the disease, which could reduce the effectiveness of the treatment; ii) Or, in the worst case, even if the intervention is performed at the appropriate time, as these are substantially more serious cases, which may or may not be associated with risk factors, the patient still cannot resist. In this perspective, for both cases, the other attributes presented in Table 3 could be used to support decision making on what is the best measure of referral of the patient, before the disease gets worse.

Table 3. Evaluation of odds ratio. The attributes that indicate greater chances of death are highlighted in bold.

Condition	OR	95% CI	*P*	z-score
ICU	**5.04**	(4.94 - 5.13)	< 0.0001	170.96
Age >=60	**4.76**	(4.68 - 4.84)	< 0.0001	184.75
Ventilatory Support	**4.18**	(4.09 – 4.27)	< 0.0001	126.32
Oxygen saturation	**2.85**	(2.79 - 2.90)	< 0.0001	103.97
Dyspnea	**2.33**	(2.28 - 2.38)	< 0.0001	76.06
Risk factor	**2.25**	(2.21 - 2.28)	< 0.0001	96.16
Kidney disease	**2.19**	(2.11- 2.28)	< 0.0001	38.24
Heart Disease	**2.18**	(2.13 - 2.22)	< 0.0001	71.94
Respiratory Discomfort	**2.17**	(2.13 - 2.21)	< 0.0001	79.03
Neuropathy	**2.08**	(2.00 - 2.17)	< 0.0001	34.79
Pneumopathy	**1.88**	(1.80 - 1.96)	< 0.0001	29.42
Liver Failure	**1.81**	(1.67 - 1.96)	< 0.0001	14.21
Immunodepression	**1.41**	(1.34 - 1.48)	< 0.0001	14.22
Diabetes	**1.35**	(1.32 - 1.38)	< 0.0001	26.91
Hemopathy	**1.27**	(1.17 - 1.38)	< 0.0001	5.58
Down Syndrome	**1.20**	(1.03 - 1.39)	$= 0.0185$	2.35
Used antiviral	**1.07**	(1.05 - 1.10)	< 0.0001	7.25
Obesity	0.93	(0.90 - 0.97)	< 0.0001	3.59
Fever	0.80	(0.78 - 0.81)	< 0.0001	23.36
Cough	0.75	(0.73 - 0.76)	< 0.0001	28.02
Sore throat	0.74	(0.72 - 0.76)	< 0.0001	27.63
Diarrhea	0.74	(0.72 - 0.76)	< 0.0001	24.17
Asthma	0.56	(0.52 - 0.59)	< 0.0001	22.11
Postpartum	0.30	(0.26 - 0.36)	< 0.0001	14.37
Vomiting	0.00	(0.003 - 0.004)	< 0.0001	266.54

4.2. Machine learning models

Figure 2-a shows the effectiveness of the models discovered in Experiment 1: Decision Tree (DT) and Naive Bayes (NB). DT was more effective than NB. The AUC values achieved show that the models developed were able to obtain promising effectiveness, especially given the high percentage missing data in the original database.

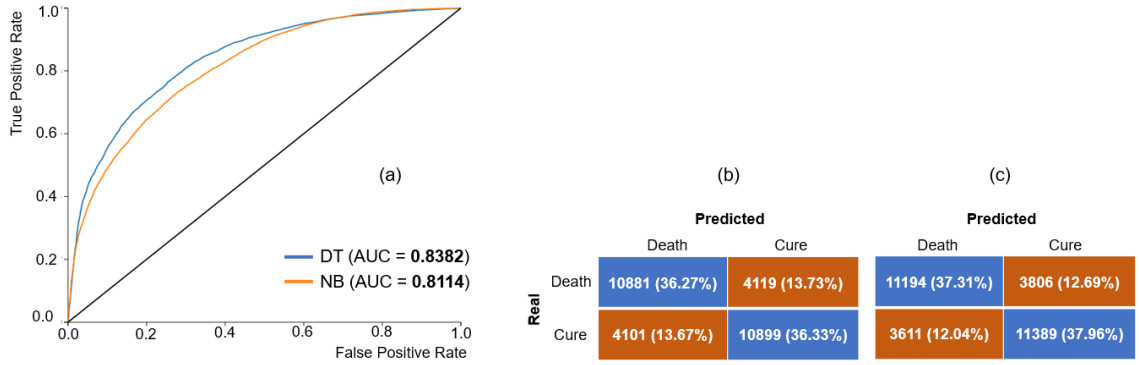

Figure 2. (a) ROC and AUC curve for the models developed; (b) Naive Bayes contingency matrix; (c) Decision Tree contingency matrix.

Considering the contingency matrices for the models developed (Figure 2-b and 2-c), there was a similar amount of false positive and false negative predictions. The algorithms were able to correctly predict between roughly 72% and 75% of the cases. Moreover, the small fraction of errors and the balance between the types of error suggest some cases are still hard to predict, but no explicit prediction bias towards an specific outcome was observed.

4.3. Variable importance

According to the ablation experiments, Figure 3 shows the ten variables considered most important for survival prediction based on DT (Figure 3-a) and NB (Figure 3-b). Some of the attributes were common to both models, such as "ventilatory support", "age", "ICU" and "race". Only "immunodepression", "dyspnea" and "fever" were model-specific. In general, this result can assist in the data collection and construction of models that prioritize them as important to the prognosis of patients.

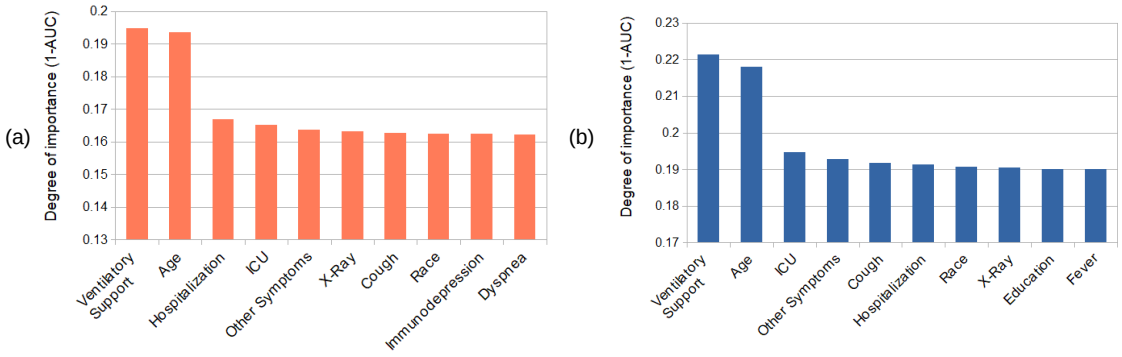

Figure 3. Variables considered most important: (a) Decision Tree. (b) Naive Bayes.

5. Conclusion

The descriptive analysis employed revealed that older age, male gender and the presence of comorbidities are factors that contribute to death, suggesting that these factors can be

used to support decision making. The calculation of the relative risk confirms that the chance of death is greater among patients aged 60 years or older, and among those who have comorbidities such as kidney, neurological, and cardiac diseases. Some symptoms were also pointed out as factors that increase the chance of death, such as dyspnea and Oxygen saturation ($< 95\%$). Some of these findings had already been pointed out as factors present in patients from other regions, such as the United States, Italy and China. Our results demonstrate that this behavior was also observed in the initial months of COVID-19 in Brazil. Analyzing the importance of variables in the perspective of machine learning algorithms, it was found that some of the attributes pointed out as the most significant in the relative risk analysis, were also selected by the algorithms, such as "age" and "dyspnea", but also factors such as "race" and "education".

The experiments carried out indicate that the model discovered is capable of predicting patient prognosis, and the model obtained by DT was slightly more effective than the model developed with the NB. The decision tree-based model obtained $AUC = 0.8382$, while the NB reached $AUC = 0.8114$, which is considered quite promising. This result indicates that machine learning techniques fed with demographic and clinical data along with patient comorbidities can help guide the clinical management of patients in more severe cases of COVID-19. It is worth noting that, although this study used data from COVID-19, given the current context, the experimental process could be extended to other scenarios.

As future work, we intend to evaluate other machine learning methods, as well as deep learning strategies. In addition, we intend to use more sophisticated techniques for explaining machine learning models, in order to better understand the relationship of attributes with prognosis prediction accuracy. Consequently, it may help the experts to comprehend the context and reliability for each prediction. We also intend to apply some feature selection technique, in order to find the most important subset of features to deal with this problem, which are relevant to the class of interest and not redundant with each other. Finally, we also intend to consider updated data on COVID-19 cases in Brazil, in order to analyze variations in the profile of deaths between the initial months and more current developments. The inclusion of these new data will also allow us to assess its impact on model's effectiveness. For this, a month-by-month iterative analysis will be conducted.

References

Barbosa, V., Ferreira, H., Gomes, L., Gomes, F., Barreto, I., Monteiro, O., and Oliveira, M. (2020). Smartres - uma plataforma iot para monitoramento inteligente em saúde e sua aplicação no contexto da covid-19. In *Proceedings of the 20th Brazilian Symposium on Computing Applied to Healthcare*, pages 297–307, Porto Alegre, RS, Brasil. SBC.

Brazil (2021). Ministério da saúde - painel coronavírus. `https://covid.saude.gov.br/`. Accessed: April 01, 2021.

Grasselli, G., Zangrillo, A., Zanella, A., Antonelli, M., Cabrini, L., Castelli, A., Cereda, D., Coluccello, A., Foti, G., Fumagalli, R., Iotti, G., Latronico, N., Lorini, L., Merler, S., Natalini, G., Piatti, A., Ranieri, M. V., Scandroglio, A. M., Storti, E., Cecconi, M., Pesenti, A., and for the COVID-19 Lombardy ICU Network (2020). Baseline

Characteristics and Outcomes of 1591 Patients Infected With SARS-CoV-2 Admitted to ICUs of the Lombardy Region, Italy. *JAMA*, 323(16):1574–1581.

Han, J., Pei, J., and Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

Leung, K., Wu, J. T., Liu, D., and Leung, G. M. (2020). First-wave COVID-19 transmissibility and severity in China outside Hubei after control measures, and second-wave scenario planning: a modelling impact assessment. *The Lancet*, 395(10233):1382–1393.

Mahase, E. (2021). Covid-19: What new variants are emerging and how are they being investigated? *BMJ*, 372.

Masood, A., Sheng, B., Li, P., Hou, X., Wei, X., Qin, J., and Feng, D. (2018). Computer-assisted decision support system in pulmonary cancer detection and stage classification on ct images. *Journal of Biomedical Informatics*, 79:117 – 128.

Nania, R. (2021). COVID-19's Fourth Wave: What You Need to Know Now. `https://www.aarp.org/health/conditions-treatments/info-2021/covid-4th-wave.html`. Accessed: August 02, 2021.

Onder, G., Rezza, G., and Brusaferro, S. (2020). Case-Fatality Rate and Characteristics of Patients Dying in Relation to COVID-19 in Italy. *JAMA*, 323(18):1775–1776.

Ozkaya, U., Ozturk, S., and Barstugan, M. (2020). Coronavirus (COVID-19) Classification using Deep Features Fusion and Ranking Technique.

Pollet, M. (2020). Coronavirus second wave: Which countries in europe are experiencing a fresh spike in covid-19 cases? `https://www.euronews.com/2020/10/05/is-europe-having-a-covid-19-second-wave-country-by-country-breakdown`. Accessed: August 02, 2021.

Randhawa, G. S., Soltysiak, M. P. M., El Roz, H., de Souza, C. P. E., Hill, K. A., and Kari, L. (2020). Machine learning using intrinsic genomic signatures for rapid classification of novel pathogens: Covid-19 case study. *PLOS ONE*, 15(4):1–24.

Richardson, S., Hirsch, J. S., Narasimhan, M., Crawford, J. M., McGinn, T., Davidson, K. W., , and the Northwell COVID-19 Research Consortium (2020). Presenting Characteristics, Comorbidities, and Outcomes Among 5700 Patients Hospitalized With COVID-19 in the New York City Area. *JAMA*, 323(20):2052–2059.

Silva, S., Simozo, F., Junior, L. M., and Tinós, R. (2020). Uso de redes neurais convolucionais para identificar displasia cortical focal em pacientes com epilepsia refratária. In *Anais do XVII Encontro Nacional de Inteligência Artificial e Computacional*, pages 211–221, Porto Alegre, RS, Brasil. SBC.

Sousa, I., Vellasco, M., and Silva, E. (2020). Classificações explicáveis para imagens de células infectadas por malária. In *Anais do XVII Encontro Nacional de Inteligência Artificial e Computacional*, pages 47–57, Porto Alegre, RS, Brasil. SBC.

Souza, F. S. H., Hojo-Souza, N. S., Santos, E. B., Silva, C. M., and Guidoni, D. L. (2020). Predicting the disease outcome in covid-19 positive patients through machine learning: a retrospective cohort study with brazilian data. *medRxiv*.

Wang, C., Horby, P. W., Hayden, F. G., and Gao, G. F. (2020a). A novel coronavirus outbreak of global health concern. *The Lancet*, 395(10223):470–473.

Wang, D., Hu, B., Hu, C., Zhu, F., Liu, X., Zhang, J., Wang, B., Xiang, H., Cheng, Z., Xiong, Y., Zhao, Y., Li, Y., Wang, X., and Peng, Z. (2020b). Clinical Characteristics of 138 Hospitalized Patients With 2019 Novel Coronavirus–Infected Pneumonia in Wuhan, China. *JAMA*, 323(11):1061–1069.

White, D. B. and Lo, B. (2020). A Framework for Rationing Ventilators and Critical Care Beds During the COVID-19 Pandemic. *JAMA*, 323(18):1773–1774.

WHO (2021a). Tracking sars-cov-2 variants. `https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/`. Accessed: August 02, 2021.

WHO (2021b). Who coronavirus (covid-19) dashboard. `https://covid19.who.int/`. Accessed: August 01, 2021.

Xie, J., Covassin, N., Fan, Z., Singh, P., Gao, W., Li, G., Kara, T., and Somers, V. K. (2020). Association Between Hypoxemia and Mortality in Patients With COVID-19. *Mayo Clinic Proceedings*, 95(6):1138–1147.

Yan, L., Zhang, H.-T., 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., and Yuan, Y. (2020). An interpretable mortality prediction model for COVID-19 patients. *Nature Machine Intelligence*, 2(5):283–288.

Zhou, F., Yu, T., Du, R., Fan, G., Liu, Y., Liu, Z., Xiang, J., Wang, Y., Song, B., Gu, X., Guan, L., Wei, Y., Li, H., Wu, X., Xu, J., Tu, S., Zhang, Y., Chen, H., and Cao, B. (2020). Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *The Lancet*, 395(10229):1054–1062.