





ARTICLE

Assessment of the evolution of patients hospitalized for COVID-19 in Paraná

 José Bruno Pereira Souza* and  Brian A. R. de Melo

Graduate Program in Biostatistics, State University of Maringá, Maringá - PR, Brazil

*Corresponding author. Email: josebruno.dte@gmail.com

(Received: December 18, 2023; Revised: May 15, 2024; Accepted: June 21, 2024; Published: February 11, 2025)

Abstract

The COVID-19 pandemic was marked by great fear, as it was a new disease of which we had no knowledge regarding its effects and prevention methods. However, during this period, we also made significant advances in research across various fields, from studying the causes and effects of the disease to the development of vaccines. In this study, we focus on assessing the evolution (recovery/death) of COVID-19 inpatients, who were hospitalized in the state of Paraná, Brazil in the year of 2022. To achieve this, we analyzed data from the System of Epidemiological Surveillance of Influenza (SIVEP) which provides information about Brazilian patients hospitalized with severe acute respiratory syndrome, using several machine learning techniques that allowed us to relate the patient evolution to possible associated factors. Results showed that age, gender, education, and neurological disorder, among other factors, have significant impacts in the evolution of inpatients. When predicting the patient outcome, we obtained an accuracy over 75%, which shows the efficiency of the models.

Keywords: COVID-19; prediction models; risk factors; SIVEP; supervised learning.

1. Introduction

According to the World Health Organization (WHO, 2023), as of October 9, 2023, there were 771,151,224 confirmed cases of COVID-19 worldwide, since the onset of the pandemic, with 37,827,912 cases occurring in Brazil. Due to the rapid growth of COVID-19 cases, many researchers have used machine learning techniques to assist in diagnosing patients (Alyasseri *et al.*, 2022). Jung *et al.*, 2022 applied machine learning techniques to data from a Paris hospital in order to assess severity considering the first three waves of the disease.

In Brazil, several studies were carried out to understand the regional effects of the pandemic. Oliveira *et al.*, 2024 evaluated the survival times of Brazilian inpatients and showed that the risk of death varies significantly depending on the region of these patients. Barreto *et al.*, 2023 used machine

and deep learning techniques to predict the evolution of inpatients from the state of Rio Grande do Norte and Oliveira & Nobre, 2023 used machine learning to predict hospitalization in Minas Gerais.

According to the Paraná Health Department, on March 12, 2020, the first six cases of Covid-19 were confirmed in the state. Since then, 2,946,293 cases have been confirmed until October 9, 2023. In this study, we consider the period from January 1, 2022, to September 12, 2022 and use data provided by the System of Epidemiological Surveillance of Influenza (SIVEP), which provides information about all Brazilian patients hospitalized with severe acute respiratory syndrome. This database shows that, in the time period considered in this study, the number of hospitalizations in the state of Paraná was 12,030 with 9,139 of them in of them being admitted to Intensive Care Units (ICUs) and with patients aged between 18 and 101 years. The SIVEP database also shows information about comorbidities, personal patient data, among other relevant information and the outcome is the patient evolution, which is a categorical variable with two levels: recovery and death (DATASUS, 2022).

Considering the rich information in the SIVEP database, our goal in this study is to assess and predict the evolution of patients, hospitalized in the state of Paraná, Brazil, based on the many variables available, such as gender, race, education, age, comorbidities, vaccine status, etc, while also evaluating the impact that these variables have on the outcome, using supervised machine learning techniques. (Hastie *et al.*, 2009; Monard & Baranauskas, 2003).

The research was previously exempt from approval by the Research Ethics Committee, as the database is public. In accordance with Resolution 466/12 of the Brazilian Research Council, the researchers reported a total ethical commitment in the handling, analysis and publication of data so that the research does not require approval.

2. Materials and Methods

2.1 Data

The data used in this study pertain to adult (age 18 to 59) and elderly (60 years or more) patients diagnosed with COVID-19 and hospitalized in the state of Paraná, Brazil, in the year 2022. These data were obtained from the "Severe Acute Respiratory Syndrome Database" constructed by the SIVEP-Gripe and are available at DATASUS (2022). The database comprises 12,030 observations, which are cases reported between January 1, 2022, and September 12, 2022. We only consider data from 2022, as 1 year of vaccination has been completed and we focus on the state of Paraná due to regional differences in health quality levels in Brazil (Júnior *et al.*, 2019; Oliveira *et al.*, 2024).

For the study, the inclusion criteria selected were the nationality of the patient, who had to be Brazilian residents in the state of Paraná and confirmed to have a COVID-19 infection. As predictor variables, we considered gender, race, education, nosocomial, puerperal, heart disease, hematology, Down syndrome, liver diseases, asthma, diabetes, chronic neurological disease, pulmonary disease, immunodeficiency, renal disease, obesity, COVID-19 vaccination, ICU admission, use of ventilatory support, and age group as shown in Table 1. Other possible predictors, such as the symptoms and type of vaccine were not included due to poor quality data.

2.2 Machine learning methods

In this section we briefly describe several machine learning techniques that will be used to predict the patient evolution using the described predictors.

2.2.1 Logistic regression

The first technique considered is the logistic regression (LR), which is a regression model that belongs to the Generalized Linear Models family (McCullagh & Nelder, 1989). The LR allows us

to estimate the probability (p) of an outcome as a function of several predictors x_j , $j = 1, \dots, k$, as shown in equation (1).

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k \quad (1)$$

In this model, the β_j are the regression coefficients which can be estimated via Maximum Likelihood. Once these coefficients are estimated it is straightforward to compute the estimated probabilities and, choosing a cutoff point, we are able to classify the observations. In this paper, we consider the best cutoff point as the one that maximizes the Youden's index (Martínez-Cambor & Pardo-Fernández, 2019; Youden, 1950).

2.2.2 Classification trees

Classification trees (CT), a non-parametric methodology used for decision-making, are well-known for their interpretability as they allow the visualization of decision logic in the form of a tree. The construction involves recursive partitioning of the predictor variables' space into nodes and leaves (Izbicki & dos Santos, 2020). The best partition is determined by minimizing the Gini index, given by equation (2),

$$\sum_R \sum_{c \in C} \hat{p}_{R,c}(1 - \hat{p}_{R,c}), \quad (2)$$

where $\hat{p}_{R,c}$ is the proportion of observations classified as category c in region R . This index is minimized when the leaves contain observations of a single class. (Breiman *et al.*, 2017).

The construction of the tree occurs in two stages: creating the complete tree and pruning to avoid overfitting. The prediction for the i -th observation in the k -th region is given by the mode of the responses observed in that region as shown in equation (3).

$$g(\mathbf{x}) = \text{mode}\{y : \mathbf{x} \in R_k\} \quad (3)$$

Classification trees are used in various applications, such as email spam classification (Sharma & Sahni, 2011), public health (Lemon *et al.*, 2003), demand forecasting (Bala, 2010), and fraud detection (Chiu *et al.*, 2011), due to their ability to handle both categorical and numerical data while providing interpretable predictions.

2.2.3 Random forests

Next, we consider the Random Forests (RF) methodology, which builds a classifier combining multiple trees of different sizes and built with a random subset of predictor variables, reducing overfitting (Izbicki & dos Santos, 2020). In this way, RF perform a random selection of predictor variables at each node of the trees, thus reducing the correlation between the trees and increasing the model's diversity.

The choice of the number of trees (B) and the variables to be selected randomly is related to the model's effectiveness and can be determined through cross-validation to balance the model's complexity and efficiency. To classify an observation with a random forest, predictions from all the trees are combined using the mode shown in equation (4),

$$g(\mathbf{x}) = \text{mode} \left\{ g^b(\mathbf{x}), b = 1, \dots, B \right\}, \quad (4)$$

where $g^b(\mathbf{x})$ is the prediction from the b -th tree.

2.2.4 Support vector machine

The Support Vector Machine (SVM) technique (Cortes & Vapnik, 1995) is a powerful machine learning tool used for classification and regression, especially in high-dimensional cases. Unlike other methods, SVM does not estimate probabilities $P(Y = c | \mathbf{x})$ but provides estimated classes for new observations.

SVM uses a linear function $f(\mathbf{x})$ to construct a classifier $g(\mathbf{x})$ and when the observations are linearly separable, SVM builds a hyperplane that perfectly separates the observations according to their classes. The margin between the hyperplane and the closest points is maximized and the points used to define the margins are called support vectors.

When the data is not linearly separable, SVM allows some points to be on the "wrong" side of the margins. This is controlled by a hyperparameter C , which determines the trade-off between fitting the training data correctly and seeking a larger margin. In more complicated scenarios, the Kernel Trick can be used to map the original space to a higher-dimensional space, improving the classification. Different types of kernels, such as polynomial, radial, and sigmoid kernels can be applied to perform this mapping. For more details, see Hastie *et al.*, 2009 and Izbicki & dos Santos, 2020.

2.2.5 Naive Bayes

The last methodology considered in this paper is the Naive Bayes (NB) which estimates the conditional probability of an event occurring given a set of predictors ($P(Y = c | \mathbf{x})$) using the Bayes Theorem.

$$P(Y = c | \mathbf{x}) = \frac{f(\mathbf{x} | Y = c) P(Y = c)}{\sum_s f(\mathbf{x} | Y = s) P(Y = s)} \quad (5)$$

To obtain the conditional probability shown in equation (5), we must estimate the marginal probabilities $P(Y = s)$ and the conditional densities $f(\mathbf{x} | Y = s)$, for all classes s . The marginal probabilities can be estimated as the sample proportions of each class, however, to estimate the conditional densities, we need to assume that the predictors are conditionally independent, given the class $Y = c$. With this assumption, the conditional density can be rewritten as:

$$f(\mathbf{x} | Y = s) = f((x_1, \dots, x_k) | Y = s) = \prod_{j=1}^k f(x_j | Y = s) \quad (6)$$

The assumption represented by equation (6) makes the algorithm computationally efficient and easy to implement and, even with these assumption, the NB classifier tends to be robust and works well in many scenarios. This makes the algorithm computationally efficient and easy to implement. For more details, see Morettin & Singer, 2022 and Sohil *et al.*, 2022.

3. Results and Discussion

Initially, a descriptive analysis was conducted and its results are summarized in Table 1, which presents all the predictors, the frequencies for each predictor, in general and according to the patient evolution, and the p-value of the Person chi-square test.

We note that the patients are, mostly, elderly (68.97%), white (81.11%), and have low levels of education. Nearly 29% of patients needed intensive care treatment (ICU) while 13% used invasive ventilatory support and 68.29% recovered. The most common comorbidity among the hospitalized patients is Cardiopathy (35,74%) followed by Diabetes (23.13%). At a 5% significance level, the predictors that were not associated with the patient evolution are Hematology, Down Syndrome, Asthma, and Obesity.

For the following analysis we randomly split the data in two: the *training* and *test datasets*, containing 70% and 30% of the data, respectively. The first one was used to train the algorithms and the second to evaluate the prediction quality from the trained algorithms. All data analysis was performed using the R software (R Core Team, 2021) and we used the *pROC* (Robin *et al.*, 2011) package to find the best cutoff points and compute the predictive measures show in in Table 5.

Table 1. Descriptive statistics of the associated factors.

Predictor	Levels	Recovery (%)	Death(%)	Total	p-value
Sex	M	3918 (47.69%)	2086 (54.68%)	6004(49.90%)	< 0.001
	F	4297 (52.31%)	1729 (45.32%)	6026(50.10%)	
Age	Adult	2966 (36.1%)	767 (20.1%)	3733(31.03%)	<0.001
	Elderly	5249 (63.9%)	3048 (79.9%)	8297(68.97%)	
Race	White	5964 (83.69%)	2667 (82.95%)	8631(81.11%)	0.142
	Yellow	63 (0.88%)	38 (1.18%)	101(0.95%)	
	Brown	913 (12.81%)	413 (12.85%)	1626(15.28%)	
	Black	176 (2.47%)	96 (2.99%)	272(2.56%)	
	Indian	10 (0.14%)	1 (0.03%)	11(10.35%)	
Education	No	194 (7.95%)	139 (12.07%)	333(9.27%)	<0.001
	Elementary 1	743 (30.45%)	422 (36.63%)	1165(32.43%)	
	Elementary 2	525 (21.52%)	277 (24.05%)	802(22.33%)	
	Highschool	684 (28.03%)	236 (20.49%)	920(25.61%)	
	College	294 (12.05%)	78 (6.77%)	372(10.35%)	
Nosocomial	Yes	187 (2.28%)	109 (2.86%)	293(2.44%)	0.064
	No	8028 (97.72%)	3706 (97.14%)	11734(97.56%)	
Postpartum	Yes	90 (1.1%)	1 (0.03%)	91(0.76%)	<0.001
	No	8125 (98.9%)	3814 (99.97%)	11939(99.24%)	
Cardiopathy	Yes	2696 (32.82%)	1571 (41.18%)	4267(35.47%)	<0.001
	No	5519 (67.18%)	2244 (58.82%)	7763(64.53%)	
Hematology	Yes	76 (0.93%)	49 (1.28%)	125(1.04%)	0.087
	No	8139 (99.07%)	3766 (98.72%)	11905(98.96%)	
Down Syndrome	Yes	29 (0.35%)	19 (0.5%)	38(0.32%)	0.308
	No	8186 (99.65%)	3796 (99.5%)	11992(99.68%)	
Liver Disease	Yes	68 (0.83%)	87 (2.28%)	155(1.29%)	<0.001
	No	8147 (99.17%)	3728 (97.72%)	11875(98.71%)	
Asthma	Yes	255 (3.1%)	114 (2.99%)	369(3.07%)	0.775
	No	7960 (96.9%)	3701 (97.01%)	11661(96.93%)	
Diabetes	Yes	1779 (21.66%)	1004 (26.32%)	2783(23.13%)	<0.001
	No	6436 (78.34%)	2811 (73.68%)	9247(76.87%)	
Neurological Disease	Yes	647 (7.88%)	495 (12.98%)	1142(9.49%)	<0.001
	No	7568 (92.12%)	3320 (87.02%)	10888(90.51%)	
Pneumopathy	Yes	535 (6.51%)	360 (9.44%)	895(7.44%)	<0.001
	No	7680 (93.49%)	3455 (90.56%)	11135(92.56%)	
Immunodeficiency	Yes	283 (3.44%)	234 (6.13%)	517(4.3%)	<0.001
	No	7932 (96.56%)	3581 (93.87%)	11513(95.7%)	
Kidney disease	Yes	391 (4.76%)	294 (7.71%)	685(5.69%)	<0.001
	No	7824 (95.24%)	3521 (92.29%)	11345(94.31%)	
Obesity	Yes	488 (5.94%)	250 (6.55%)	738(6.13%)	0.206
	No	7727 (94.06%)	3565 (93.45%)	11292(93.87%)	
Covid vaccine	Yes	6696 (83.52%)	3018 (81.04%)	9714(80.75%)	0.001
	No	1321 (16.48%)	706 (18.96%)	2316(19.25%)	
ICU	Yes	1525 (18.56%)	1954 (51.22%)	3479(28.92%)	<0.001
	No	6690 (81.44%)	1861 (48.78%)	8551(71.08%)	
Ventilatory support	Invasive	309 (3.76%)	1305 (34.21%)	1611(13.39%)	<0.001
	Non-invasive	3889 (47.34%)	1707 (44.74%)	5596(46.53%)	
	No	4017 (48.9%)	803 (21.05%)	4820(40.08%)	

3.1 Logistic Regression

Initially, we modeled the probability of a patient death via LR. Due to the large number of predictors, we applied the stepwise method for variable selection, using the Akaike Information Criteria (AIC) value to select the most important predictors. We used the *hnp* and *pROC* packages to evaluate the residuals and find the best cutoff point for the estimated Table 2 shows the results from the fitted LR model including estimated regression coefficients (β), its standard error (SE), the p-value, the Odds Ratio (OR) and a 95% confidence interval for the OR.

Table 2. Estimates of the logistic regression final model.

Predictor	Level	β	SE	p-value	OR	95% CI
Intercept	-	0.866	0.263	0.001	2.37	1.42 - 3.98
Sex	Male	0.385	0.107	<0.001	1.469	1.19 - 1.81
Education	Elementary 1	-0.401	0.174	0.021	0.669	0.48 - 0.94
	Elementary 2	-0.592	0.191	0.002	0.553	0.38 - 0.80
	High School	-0.620	0.196	0.001	0.538	0.37 - 0.79
	College	-0.909	0.244	<0.001	0.403	0.25 - 0.65
Postpartum	Yes	-12.895	249.248	0.959	<0.001	0.00 - 0.00
Neurological Disorder	Yes	0.621	0.171	<0.001	1.862	1.33 - 2.60
Immunodeficiency	Yes	0.750	0.280	0.007	2.116	1.22 - 3.66
Obesity	Yes	0.327	0.205	0.110	1.387	0.92 - 2.07
ICU	Yes	0.491	0.134	<0.001	1.633	1.25 - 2.12
Ventilatory Support	No	-3.029	0.213	<0.001	0.048	0.03 - 0.07
	Yes, non-invasive	-1.990	0.187	<0.001	0.136	0.09 - 0.19
Age	Elderly	0.641	0.138	<0.001	1.898	1.45 - 2.49

Observing the OR values obtained for statistically significant variables, we find that male patients have a 47% higher chance of death compared to female patients. Evaluating the Education predictor, we note that, as the patient education level increases, the probability of death decreases, so that patients with a elementary 1 level of education have a 33% lower chance of death than patients with no education. In the outermost case, the chance of death in patients with college education are 60% lower when compared to patients with no education.

For patients diagnosed with neurological diseases the chance of death is 86% higher while in patients with immunodeficiency this value is 111%. We also note that patients who were admitted to the ICU or had invasive ventilatory support have much higher chances of death. Finally, elderly patients have a 90% higher chance of death than adult patients. This significant increase in the occurrence of death is justified because some variables address risk factors presented in various articles, for example, Duprat & Melo, 2020 and Galvão & Roncalli, 2021.

Applying this model to the test dataset, we estimate the probability of death for each patient and, using this probabilities, we can predict the patient evolution. The best cutoff point for classification was found to be 0.399, that is, patients who have an estimated probability of death lower than 0.399 are predicted to recover. The predictive capacity measures obtained through this process are organized in Table 5 where we can see that 78% of the patients in the test dataset were classified correctly. Among the patients that recovered the proportion of correctly classified observation is 88% and among the deaths this number is 55%.

3.2 Classification Tree

The classification tree was built using all the variables in the dataset and was subsequently pruned to obtain a more effective and less complex model. Initially, the complexity parameter for tree

construction was set to 0.001. For pruning, we evaluated the cross-validation error and its minimum value was 0.737 corresponding to a total of 10 splits. This analysis was conducted using the *rpart* (Therneau & Atkinson, 2019) and *rpart.plot* (Milborrow, 2022) packages.

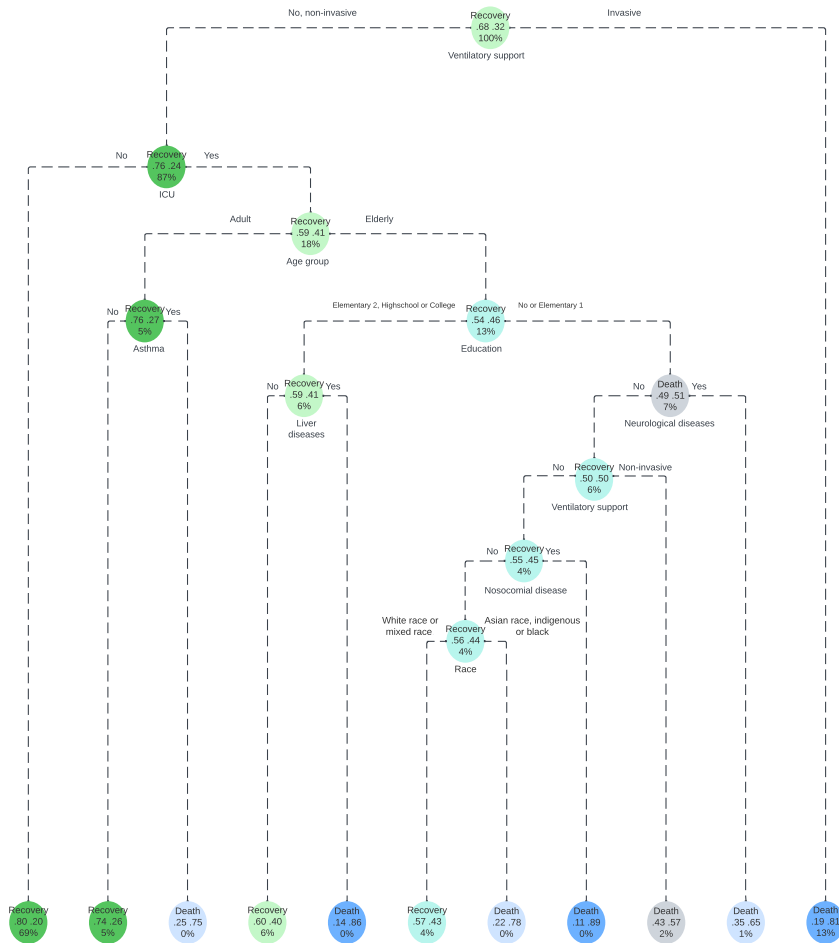


Figure 1. Classification tree pruned tree with 10 splits.

Figure 1 shows the pruned tree with 10 splits. In this tree, the selected variables were ventilatory support, ICU admission, the patient’s age group, asthma, education, neurological diseases, liver diseases, nosocomial, and race. The ventilatory support variable appeared twice in the model because it has three categories. Among patients requiring invasive ventilatory support, the proportion of deaths was 81%. For patients where treatment was done non-invasively or there was no support, the proportion of recovery was 76%.

The second split was generated by the ICU admission variable. Patients who were not admitted to the ICU had a recovery proportion of 80% while those who were admitted had a recovery proportion of 59%. The third split was given by the age group, with a 73% recovery proportion for adult patients. These patients were directed to a fourth split defined by the asthma variable, where

asthmatic patients had a death proportion of 75%, while non-asthmatic patients had a recovery proportion of 74%.

Elderly patients were directed to the fifth split given by the education variable. Patients with a high school level of education, or higher, had a recovery proportion of 59% and were directed to a sixth node given by the liver disease variable. Patients with liver diseases had an 86% death proportion; otherwise, they had a 60% recovery proportion. Patients with a grade school level 1 or no education had a 51% death proportion and were directed to a seventh split given by the neurological disease variable. The death proportion for patients with this condition was 65%; otherwise, an eighth split was made.

In the eighth split, the use of non-invasive ventilatory support showed a 57% death proportion; otherwise, the recovery proportion was 55%. The ninth split, given by the nosocomial variable, showed a death proportion of 89%; otherwise, the recovery proportion was 56%. The last split was generated by the race variable. Whites or people of mixed race had a recovery proportion of 57%, while other cases had a death proportion of 78%.

To assess the predictive power of the model, we applied it to the test sample and computed the predictive metrics in the Table 5. We note that the CT classifier obtained an accuracy value of 77%. Despite the model having a low sensitivity value, we had a high specificity, indicating a good proportion of correct negative classifications. The Negative Predictive Value (NPV) and Positive Predictive Value (PPV) metrics were balanced, 76% and 77%, respectively, showing that this classifier does not favor the cure or death of patients.

3.3 Random Forests

To implement the RF model, we considered the *randomForest* package (Liaw & Wiener, 2002) and explored several configurations to maximize its predictive power. Specifically, we varied the number of variables from 1 to 20 and the number of trees among 500, 1000, and 1500, selecting the combination that achieved the highest accuracy. The results, presented in Table 3, show that the highest accuracy was obtained with forests containing either 500 or 1000 trees and using 5 variables. Since both configurations yielded the same accuracy, we selected the model with 500 trees for its computational efficiency.

Table 3. Number of variables, trees and resulted accuracy of the generated Random Forests.

Variables	Trees	Accuracy	Variables	Trees	Accuracy	Variables	Trees	Accuracy
1	500	0.686	8	500	0.759	15	500	0.743
1	1000	0.686	8	1000	0.756	15	1000	0.742
1	1500	0.686	8	1500	0.757	15	1500	0.742
2	500	0.759	9	500	0.752	16	500	0.741
2	1000	0.759	9	1000	0.750	16	1000	0.741
2	1500	0.760	9	1500	0.754	16	1500	0.744
3	500	0.758	10	500	0.752	17	500	0.741
3	1000	0.759	10	1000	0.751	17	1000	0.741
3	1500	0.757	10	1500	0.749	17	1500	0.741
4	500	0.763	11	500	0.749	18	500	0.742
4	1000	0.762	11	1000	0.749	18	1000	0.742
4	1500	0.762	11	1500	0.749	18	1500	0.740
5	500	0.765	12	500	0.745	19	500	0.739
5	1000	0.765	12	1000	0.747	19	1000	0.739
5	1500	0.762	12	1500	0.749	19	1500	0.741
6	500	0.764	13	500	0.745	20	500	0.741
6	1000	0.764	13	1000	0.746	20	1000	0.741
6	1500	0.764	13	1500	0.742	20	1500	0.741
7	500	0.759	14	500	0.745			
7	1000	0.761	14	1000	0.743			
7	1500	0.758	14	1500	0.744			

Applying the chosen forest to the test dataset, the probability of death was estimated for each observation. The cutoff point was chosen in the same way as in the LR and its value is 0.295. Table 5 presents the predictive metrics of the forest using this cutoff point.

The RF classifier showed a satisfactory accuracy value, when compared to the previous two classifiers. The values of specificity and VPN were also high, demonstrating that the model had a good correct classification proportion for negatives, i.e., patients classified as recovery.

3.4 Support Vector Machines - SVM

To build the best the SVM classifier we considered different values for the hyperparameters that control the number of observations that can be on the "wrong" side of the margins, and several options for the kernel, including radial, linear, sigmoid, and polynomial. The cost parameter varied among the values 0.001, 0.01, 0.1, 1, 10, 100, 1000, and gamma between 0.5, 1, and 2. For each combination of these parameters, we evaluated the errors and the results are shown in Table 4. This analysis was conducted using the *e1071* (Meyer *et al.*, 2021).

Table 4. Error of the SVM classifier for each combination of parameters values and kernels.

Radial			Linear		Sigmoid			Polynomial		
Cost	Gamma	Error	Cost	Error	Cost	Gamma	Error	Cost	Gamma	Error
0.001	0.5	0.326	0.001	0.319	0.001	0.5	0.326	0.001	0.5	0.251
0.01	0.5	0.326	0.01	0.259	0.01	0.5	0.253	0.01	0.5	0.262
0.1	0.5	0.326	0.1	0.258	0.1	0.5	0.325	0.1	0.5	0.280
1	0.5	0.294	1	0.257	1	0.5	0.333	1	0.5	0.285
10	0.5	0.303	10	0.252	10	0.5	0.353	10	0.5	0.284
100	0.5	0.305	100	0.253	100	0.5	0.351	100	0.5	0.306
1000	0.5	0.308	1000	0.263	1000	0.5	0.353	1000	0.5	0.299
0.001	1	0.326			0.001	1	0.326	0.001	1	0.262
0.01	1	0.326			0.01	1	0.251	0.01	1	0.278
0.1	1	0.326			0.1	1	0.331	0.1	1	0.287
1	1	0.311			1	1	0.351	1	1	0.287
10	1	0.313			10	1	0.337	10	1	0.309
100	1	0.319			100	1	0.335	100	1	0.298
1000	1	0.316			1000	1	0.327	1000	1	0.405
0.001	2	0.326			0.001	2	0.326	0.001	2	0.277
0.01	2	0.326			0.01	2	0.252	0.01	2	0.285
0.1	2	0.326			0.1	2	0.341	0.1	2	0.287
1	2	0.310			1	2	0.352	1	2	0.287
10	2	0.314			10	2	0.355	10	2	0.300
100	2	0.316			100	2	0.352	100	2	0.413
1000	2	0.316			1000	2	0.343	1000	2	0.508

Based on the results in Table 4, we selected the sigmoid kernel with a cost value of 0.01 and gamma of 1, as it yielded the lowest error among all tested configurations. We applied this classifier to the test data and computed the predictive metrics presented in Table 5. The results indicate that the SVM classifier achieved good accuracy, with high sensitivity and positive predictive value (PPV), suggesting effective classification of positive cases.

3.5 Naïve Bayes

In the naïve Bayes prediction model, the parameter selection was done through ten-fold repeated cross-validation, varying the Laplace smoothing and bandwidth of the model. The metric for model selection was accuracy. Unlike the models presented earlier, we did not choose the variables to be used in the model, thus utilizing all variables in the dataset. The best results were obtained with a bandwidth value of 3 for the model as it provided the highest accuracy, up to the fifth decimal place and the application or absence of Laplace smoothing did not influence the model fit. Thus, we assumed the model without Laplace smoothing.

The NB outputs are the probabilities of death and a cutoff point was computed to make the classification as in the LR and RF scenarios. Applying it to the test data, we obtain the results in Table 5 where we note that the naïve Bayes technique showed the lowest accuracy among all techniques, (72%), but high specificity and NPV values suggesting effective classification of negative cases.

3.6 Discussion

In Table 5 we observe that the highest accuracy values were obtained in the LR and CT methods. Additionally, both methods facilitate the identification of predictors that are significantly associated

Table 5. Metrics of predictive quality for the fitted techniques.

Fit	Sen	Esp	PPV	NPV	Acc
LR	55	88	69	81	78
CT	40	94	76	77	77
RF	60	83	63	82	76
SVM	77	67	90	43	75
NB	59	79	57	80	72

with the outcome. Comparing the results from these techniques, we note several similarities: both identified the *Ventilatory support* as the main predictor while also recognizing *Education*, *Age* and *ICU* as associated factors. However, we see a certain advantage in the CT as it can combine the predictors levels to produce a better model. For example, for the *Ventilatory support*, the CT only considers the split in *No* and *Yes, no invasive* after the data was split by *Age*, *Education* and *Neurological diseases*. The CT also identifies predictors that did not show up in the LR, such as *Asthma*, *Liver diseases*, *Neurological diseases*, *Nosocomial diseases* and *Race*. This structure resulted in a more balanced prediction, with a PPV and NPV of 76% and 77%, indicating that it can be used to predict deaths and recoveries while maintaining a high correct classification while the LR is not so good in predicting deaths. This shows us that the associated factors identified by only the CT are also important and should be considered in future studies.

The predictions in the other three classifiers are very unbalanced, that is, the RF and NB perform well in identifying negatives, but show much lower PPV values while the SVM has the opposite behavior. However, it can be argued that, in this situation, a good prediction of positives (deaths) is more important, as it would identify patients at higher risk, allowing them to be treated differently, reducing the total number of deaths. Following this idea, SVM appears to be the best technique that could be applied to patient data upon admission.

Regarding the risk factors, similar results were found in Oliveira *et al.*, 2024 where higher education levels were associated with a lower risk of death. Other factors, such as sex, age, and ICU admission, were also identified in both studies, but the *Asthma* predictor was only identified in the CT. Oliveira & Nobre, 2023 identified in their study *Age* and *Race* as associated factors for hospitalization with *Comorbidity* being the most important predictor.

Due to the large number of COVID-19 cases and different levels of worsening of the disease, the research was limited by the cases included in the database, where only hospitalized patients were included. We are also limited by the lack of data for other important variables in this predictive process, such as the patient's clinical condition.

4. Conclusions

In this study, we applied various classification methods to predict the outcomes of COVID-19 inpatients in Paraná. The models used include Logistic Regression, Decision Trees, Random Forests, Support Vector Machines and Naïve Bayes.

When comparing the applied techniques using prediction metrics, we observed that the LR achieved the highest accuracy, demonstrating its usefulness in understanding how the the associated factors impact the outcome and as a classifier. However, its PPV and NPV values were unbalanced, indicating that the LR performs well in classifying recoveries cases but is less effective in predicting the deaths. The NB and RF results showed a similar behavior to the LR, in the sense that they performed better when classifying negatives than positives, but with a smaller overall accuracy.

Opposite to that, the SVM showed the best results when classifying the positives, with a PPV of 90% indicating that it can be used to evaluate whether a patient has a high risk of death. But, it

has a poor performance when classifying the recoveries, with a NPV of only 43%. Finally, the CT provided the most balanced prediction results, with a PPV and NPV of 76% and 77%, indicating that it can be used to predict deaths and recoveries while maintaining a high correct classification overall. The CT also gives an understanding of the relationships between the outcome and the predictors and among the predictors themselves.

Regarding the impact of the predictors on patient outcomes the LR and CT showed some similar results, identifying the use of ventilatory support, education, age, ICU, and neurological diseases as risk factors for COVID-19 inpatients.

Acknowledgments

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES).

Conflicts of Interest

The authors declare no conflict of interest.

Author Contributions

Conceptualization: SOUZA, J.B.P.; MELO, B.A.R. **Data curation:** SOUZA, J.B.P.; MELO, B.A.R. **Formal analysis:** SOUZA, J.B.P.; MELO, B.A.R. **Funding acquisition:** SOUZA, J.B.P.; MELO, B.A.R. **Investigation:** SOUZA, J.B.P.; MELO, B.A.R. **Methodology:** SOUZA, J.B.P.; MELO, B.A.R. **Project administration:** SOUZA, J.B.P.; MELO, B.A.R. **Software:** SOUZA, J.B.P.; MELO, B.A.R. **Resources:** SOUZA, J.B.P.; MELO, B.A.R. **Supervision:** SOUZA, J.B.P.; MELO, B.A.R. **Validation:** SOUZA, J.B.P.; MELO, B.A.R. **Visualization:** SOUZA, J.B.P.; MELO, B.A.R. **Writing - original draft:** SOUZA, J.B.P.; MELO, B.A.R. **Writing - review and editing:** SOUZA, J.B.P.; MELO, B.A.R.

References

1. Alyasseri, Z. A. A. *et al.* Review on COVID-19 diagnosis models based on machine learning and deep learning approaches. *Expert systems* **39**, e12759 (2022).
2. Bala, P. K. *Decision tree based demand forecasts for improving inventory performance in 2010 IEEE International Conference on Industrial Engineering and Engineering Management* (2010), 1926–1930.
3. Barreto, T. d. O. *et al.* Artificial intelligence applied to analyzes during the pandemic: COVID-19 beds occupancy in the state of Rio Grande do Norte, Brazil. *Frontiers in Artificial Intelligence* **6**, 1290022 (2023).
4. Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. *Classification and regression trees* (Routledge, 2017).
5. Chiu, C., Ku, Y., Lie, T. & Chen, Y. Internet auction fraud detection using social network analysis and classification tree approaches. *International Journal of Electronic Commerce* **15**, 123–147 (2011).
6. Cortes, C. & Vapnik, V. Support-vector networks. *Machine learning* **20**, 273–297 (1995).
7. DATASUS. SRAG 2021 e 2022 – Banco de Dados de Síndrome Respiratória Aguda Grave- Incluindo dados da COVID-19 <https://opendatasus.saude.gov.br/dataset/srag-2021-e-2022>. (accessed: 14.09.2022).
8. Duprat, I. P. & Melo, G. C. d. Análise de casos e óbitos pela COVID-19 em profissionais de enfermagem no Brasil. *Revista Brasileira de Saúde Ocupacional* **45** (2020).

9. Galvão, M. H. R. & Roncalli, A. G. Fatores associados a maior risco de ocorrência de óbito por COVID-19: análise de sobrevivência com base em casos confirmados. *Revista brasileira de epidemiologia* **23** (2021).
10. Hastie, T., Tibshirani, R., Friedman, J. H. & Friedman, J. H. *The elements of statistical learning: data mining, inference, and prediction* (Springer, 2009).
11. Izbicki, R. & dos Santos, T. M. *Aprendizado de máquina: uma abordagem estatística* (Rafael Izbicki, 2020).
12. Jung, C., Excoffier, J.-B., Raphaël-Rousseau, M., Salaün-Penquer, N., Ortala, M. & Chouaid, C. Evolution of hospitalized patient characteristics through the first three COVID-19 waves in Paris area using machine learning analysis. *Plos one* **17**, e0263266 (2022).
13. Júnior, P. *et al.* Hospitalizações e óbitos por influenza no Brasil: uma estimativa de incidência no período de 2010 a 2016 (2019).
14. Lemon, S. C., Roy, J., Clark, M. A., Friedmann, P. D. & Rakowski, W. Classification and regression tree analysis in public health: methodological review and comparison with logistic regression. *Annals of behavioral medicine* **26**, 172–181 (2003).
15. Liaw, A. & Wiener, M. Classification and Regression by randomForest. *R News* **2**, 18–22. <https://CRAN.R-project.org/doc/Rnews/> (2002).
16. Martínez-Cambor, P. & Pardo-Fernández, J. C. The Youden index in the generalized receiver operating characteristic curve context. *The international journal of biostatistics* **15**, 20180060 (2019).
17. McCullagh, P. & Nelder, J. A. *Generalized linear models* (Chapman & Hall, 1989).
18. Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A. & Leisch, F. *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien* R package version 1.7-9 (2021). <https://CRAN.R-project.org/package=e1071>.
19. Milborrow, S. *rpart.plot: Plot 'rpart' Models: An Enhanced Version of 'plot.rpart'* R package version 3.1.1 (2022). <https://CRAN.R-project.org/package=rpart.plot>.
20. Monard, M. C. & Baranauskas, J. A. Conceitos sobre aprendizado de máquina. *Sistemas inteligentes-Fundamentos e aplicações* **1**, 32 (2003).
21. Morettin, P. A. & Singer, J. d. M. *Estatística e ciência de dados* (2022).
22. Oliveira, G. G. R. & Nobre, C. N. *The Use of Machine Learning to Predict Hospitalization of Covid-19: A Case Study in the State of Minas Gerais-Brazil*. in *HEALTHINF* (2023), 392–399.
23. Oliveira, M. M., de Melo, B. A. R. & Salci, M. A. Evaluation of survival time in people hospitalized for COVID-19 in Brazil. *Acta Scientiarum. Health Sciences* **46** (2024).
24. R Core Team. *R: A Language and Environment for Statistical Computing* R Foundation for Statistical Computing (Vienna, Austria, 2021). <https://www.R-project.org/>.
25. Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C. & Müller, M. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **12**, 77 (2011).
26. Sharma, A. K. & Sahni, S. A comparative study of classification algorithms for spam email data analysis. *International Journal on Computer Science and Engineering* **3**, 1890–1895 (2011).
27. Sohil, F., Sohali, M. U. & Shabbir, J. *An introduction to statistical learning with applications in R: by Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani, New York, Springer Science and Business Media, 2013, \$41.98, eISBN: 978-1-4614-7137-7* 2022.

28. Therneau, T. & Atkinson, B. *rpart: Recursive Partitioning and Regression Trees* R package version 4.1-15 (2019). <https://CRAN.R-project.org/package=rpart>.
29. WHO. *World Health Organization Coronavirus (COVID-19) Dashboard* <https://covid19.who.int/>. Accessed: 2023-10-09.
30. Youden, W. J. Index for rating diagnostic tests. *Cancer* 3, 32-35 (1950).