# Death risk and the importance of clinical features in elderly people with COVID-19 using the Random Forest Algorithm

Tiago Pessoa Ferreira Lima [1]
https://orcid.org/0000-0002-1148-4288

Gabrielle Ribeiro Sena [2]
https://orcid.org/0000-0002-8430-3599

Camila Soares Neves [3]
https://orcid.org/0000-0001-5477-4296

Suely Arruda Vidal [4]
https://orcid.org/0000-0002-4268-520X

Jurema Telles Oliveira Lima [5]
https://orcid.org/0000-0003-2671-3570

Maria Julia Gonçalves Mello [6]
https://orcid.org/0000-0003-4645-8343

Flávia Augusta de Orange Lins da Fonseca e Silva [7]
https://orcid.org/0000-0003-0528-4164

[1,2,4,5,6] Instituto de Medicina Integral Prof. Fernando Figueira. Rua dos Coelhos, 300, Boa Vista. Recife, PE, Brazil. CEP: 50.070-902. E-mail: gabriellesena8@gmail.com
[3,7] Faculdade Pernambucana de Saúde. Recife, PE, Brazil.

## Abstract

*Objectives: train a Random Forest (RF) classifier to estimate death risk in elderly people (over 60 years old) diagnosed with COVID-19 in Pernambuco. A "feature" of this classifier, called feature importance, was used to identify the attributes (main risk factors) related to the outcome (cure or death) through gaining information.*

*Methods: data from confirmed cases of COVID-19 was obtained between February 13 and June 19, 2020, in Pernambuco, Brazil. The K-fold Cross Validation algorithm (K=10) assessed RF performance and the importance of clinical features.*

*Results: the RF algorithm correctly classified 78.33% of the elderly people, with AUC of 0.839. Advanced age was the factor representing the highest risk of death. The main comorbidity and symptom were cardiovascular disease and oxygen saturation ≤ 95%, respectively.*

*Conclusion: this study applied the RF classifier to predict risk of death and identified the main clinical features related to this outcome in elderly people with COVID-19 in the state of Pernambuco.*

**Key words** *COVID-19, Risk factors, Elderly people, Random Forest*

This article is published in Open Access under the Creative Commons Attribution license, which allows use, distribution, and reproduction in any medium, without restrictions, as long as the original work is correctly cited.

http://dx.doi.org/10.1590/1806-9304202100S200007

## Introduction

From the beginning of the COVID-19 pandemic (coronavirus 2019 disease) until September 27, 2020, Brazil, the largest country in South America and the fifth largest in the world, was already considered the second country in number of deaths from the disease. By mid-October, at least 4,717,991 Brazilians had developed the infection and of these, 141,406 evolved to death.[1] The lethality rate in several states in the Brazilian North/Northeast was much higher than the national average, especially in Pernambuco.[1] Faced with this epidemiological scenario, one of the challenges, besides the vaccine, is the need to guide public health policies for surveillance and control the disease. Through the identification of the main risk factors, for example, it is possible to provide early monitoring of the most vulnerable groups, reducing the chance of evolution to unfavorable clinical outcomes.

Data extracted from patients with COVID-19 are a valuable source of information about both the pathophysiology of the disease and the risk factors associated with death. These data have been widely studied, and it is currently agreed that advanced age and the presence of comorbidities are associated with increased morbidity and mortality.[2] The abundant availability of these data allows the construction of the Learning Machine (LM) algorithms - a branch of Artificial Intelligence - in which it is possible to identify more susceptible people based on individual features. Through methods called Classification, the algorithm learns during a process called training by receiving a set of inputs (clinical characteristics) along with the outputs (outcome). Finally, the algorithm is able to predict an output from inputs not seen during training.

Several LM algorithms are widely used in building predictive models of disease. Random Forest (RF) in particular, has shown higher accuracy when compared to other algorithms.[3] It has the ability to list which attributes contribute to the decision making and is often used as a feature selection technique. Feature selection is considered an essential step in data analysis, as it can reduce the complexity/dimensionality of the problem.[4] An optimized data set leads to a more accurate model and also improves its interpretability.[5] This is especially important in the development of algorithms for clinical screening, as its computational cost should be as low as possible and healthcare professionals are interested in the pathophysiological mechanisms underlying the LM model.

## Basic Concepts

This section presents concepts of MA that are essential for understanding the work.

### Classifier

Given a set of instances, consisting of constructed examples with attribute values as well as the associated class, a learning (or inducing) algorithm generates as output a classifier (also called hypothesis) so that, given an instance with the unknown class, it can label it. Formally, an instance is a pair $\{x_i, f(x_i)\}$, where $x_i$ is the input (set of attributes) and the $f(x_i)$ is output (class or label). Let $X = \{\{x_1, f(x_1)\}, \{x_2, f(x_2)\},..., \{x_n, f(x_n)\}\}$ be a set of $n$ examples, the task of the learning algorithm is to induce a function $h(.)$ that approximates the function $f(.)$. In this sense, $h(.)$ is called a hypothesis about the objective function $f(.)$, or, $h(x_1) \approx f(x_1)$.

### Decision Trees

Decision trees are constructed and represented using two elements: nodes and the branches connecting to nodes. To make a decision, the flow starts at the root of the node, navigates through the branches until it reaches a leaf node. Each node in the tree denotes a test of an attribute, and the branches denote the possible values the node can take. During the tree formation process, also known as training or learning, consideration is given to the homogeneity of the classes for each division of the node. Basically, the algorithm evaluates the information gained of the attributes for the separation of the samples present in the data set destined for training.[6] The Gini impurity (GI) is an index for evaluating attributes in the separation of samples with the same label, that is, the homogeneity of the classes is sought to compose a node. The GI is defined from Equation 2.1, where $p{=}p_1..p_c$ is the proportion of the samples from the $p_c$ to the $m$ node, respectively. The index evaluates all randomly selected predictors to build the tree and will choose the one with the highest degree of homogeneity among the samples. If the $m$ node is pure (homogeneous), then the proportion of the $p_i$ $(m)$ class $i$ to the $m$ node will equal 1 and consequently the index will equal 0. The attribute for division is chosen according to the purity decrement shown in Equation 2.2, where node division of $m$, $P_{esq}$ and $P_{dir}$, are the proportions of the samples in the left and right in the child node, respectively.

$$I_G\,(m) = 1 - \sum^{c}_{i=1} p_i\,(m)^2 \qquad (2.1)$$

$$I_G\,(m) = 1 - \sum\nolimits^{c}_{i=1} p_i\,(m)^2 \qquad (2.2)$$

## Random Forest Algorithm

Let $H = \{h_1, h_2, h_3\}$ be a set or ensemble of three classifiers. One instance $x_i$ will be labeled by each classifier from $H$. If the three classifiers make distinct errors, then when $h_1(x_i)$ is wrong, it is possible that $h_2(x_i)$ and $h_3(x_i)$ are correct, so that combining the hypotheses by voting can correctly classify $x_i$. The random forest algorithm or RF[7] is based on the ensemble strategy. It provides diversity by using the concept of random redistribution of the data. Thus, when building each $h_i \in H$, for a given training $Л$ , set, a subset of data is generated $Л$. In this way, the algorithm generates several decision trees, each trained with a random distribution. A major quality of RF is easy to measure the relative importance of each attribute for prediction. The algorithm implemented in Sklearn,[8] for example, provides an excellent tool for this, which measures the importance of features by analyzing how many nodes in the trees using a given attribute to reduce the overall impurity of the forest. It calculates this value automatically for each feature after training and normalizes the results so that the sum of all the importance equaling to 1. The higher the value, the more important the attribute is. The importance of an attribute is calculated as the total (normalized) reduction of the criteria brought about by this attribute. It is also known as the Gini importance.[8]

## K-FOLD Cross Validation

Cross-validation ($K$-fold cross validation) is a sampling method used to performanalysis of LM algorithms.[9,10] It consists of randomly dividing the ensemble $X$ into mutually exclusive $K$ folds of equal size. The examples in the $K$-1 folds are then used to train the model and the induced hypothesis is tested on the remaining fold. This $K$ process is repeated over and over again, so that all folds are used only once as a test set, as shown in Figure 1 which used $K$=10.

## Performance Metrics

The error rate of a $h$ classifier is denoted by $err(h)$ , obtained from Equation 2.3. This measure compares the class assigned by each example classifier to its true class. If the two classes are equal, $h(x_i)= f(x_i)$ so then $|\,[h(x_i) \neq f(x_i)]| = 1$; otherwise $|[h(x_i) \neq f(x_i)]| = 0$. The accuracy or hit rate is denoted by c and corresponds to the complement of the error rate, as in Equation 2.4.

$$err(h) = \frac{1}{Л} - \sum\nolimits^{Л}_{i=1}|\,|\,h(x_i) \neq f(x_i)| \qquad (2.3)$$

$$acc(h) = 1 - \ err(h) \qquad (2.4)$$

The error and hit rates can be obtained through a confusion matrix, which corresponds to a matrix whose dimension is the number of classes existing in $X$. In a confusion matrix referring to a set of examples with two classes, usually called positive and negative, we have: true positives (TP) which correspond to the example that is positive and was classified as positive; false positives (FP) which are negative examples classified as positive; true negatives (TN) which are negative examples classified as negative; and finally, the false negatives (FN) which are positive examples that were classified as negative. From the confusion matrix, one can then obtain the error rate and the hit rate by means of Equations 2.5 and 2.6, respectively.

$$err(h) = \frac{\text{FN + FP}}{\text{VP +FN + FP + VN}} \qquad (2.5)$$

$$acc(h) = \frac{\text{VP + VN}}{\text{VP +FN + FP + VN}} \qquad (2.6)$$

Another widely used performance metric, AUC (area under the ROC curve), is obtained by generating a plot of sensitivity versus (1-specificity), known as the ROC (receiver operating characteristic) curve, and calculating the area under the curve. Sensitivity is the ratio of true positives to total positive examples, as shown in Equation 2.7. Specificity is the ratio of true negatives to total negative examples, as shown in Equation 2.8. The higher the AUC value, the better the performance of the classifier. AUC values vary over a range [0,1].

$$sensibilidade = \frac{\text{VP}}{\text{VP + FN}} \qquad (2.7)$$

$$\text{especificidade} = \frac{\text{VN}}{\text{FP + VN}} \qquad (2.8)$$

## Methods

We identified 11,375 elderly patients who met the eligibility criteria (age over 60 years) and separated them into a single database. These elderly people were notified in the period from February 13 to June 19, 2020 in the state of Pernambuco, Brazil. The data analyzed came from the Secretary of Planning and Management in Pernambuco (SEPLAG-PE), downloaded on June 20 at: www.dados.seplag.pe.gov.br. All the elderly people who were in home isolation or hospitalized were excluded, since these still did not have the outcome concluded by the end of the period considered. A total of 7486 elderly people remained thereafter, of these 4356 (58.19%) were recovered and 3130 (41.81%) died.

The attributes were considered: sex (male, female), age and clinical features, such as: cough, dyspnea, fever, oxygen saturation ≤95%, presence of cardiovascular, chronic respiratory, chronic renal, diabetes, neurological, neoplasms, alcoholism, smoking. The aim was to build an RF, based on these attributes, and present which are the most important in predicting death in elderly patients with COVID-19 in Pernambuco. The work was implemented in Python[11] language, using the RF algorithm, available in the Sklearn module, according to the documentation available at: https://scikitlearn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html. A Cross Validation with was employed to calculate the performance and importance of the attributes. The methodology flow chart, illustrated in Figure 1, shows how the metrics that are presented in the results were calculated.

## Results

The mean and standard deviation of age was 72.94 ± 9.55 years, with a median of 71.0 years old. The mean age between patients recovered and those who died was 70.95 ± 9.06 and 75.70 ± 9.52 years old, respectively. The female patients corresponded to 3821 (51.04%) the male patients 3665 (48.96%). The overall case fatality rate was 41.81%. The lethality rate by age group, 29.49% being between 60-69 years old, 45.89% between 70-79 years old and 57.65% over 80 years old. In regard to the symptoms presented by the overall group, 4860 (64.92%) had cough, 4403 (58.82%) fever, 3773 (50.40%) dyspnea and 2614 (34.92%) peripheral saturation of $O_2 \leq 95\%$. However, in the group of patients who died, the most relevant clinical manifestation was dyspnea, 2244 (71.69%). In relation to comorbidities, the most frequent in the entire sample were Cardiovascular Diseases 1298 (17.34%), Diabetes Mellitus 1081

**Figure 1**

Flowchart of the Cross Validation methodology using 10 folds.

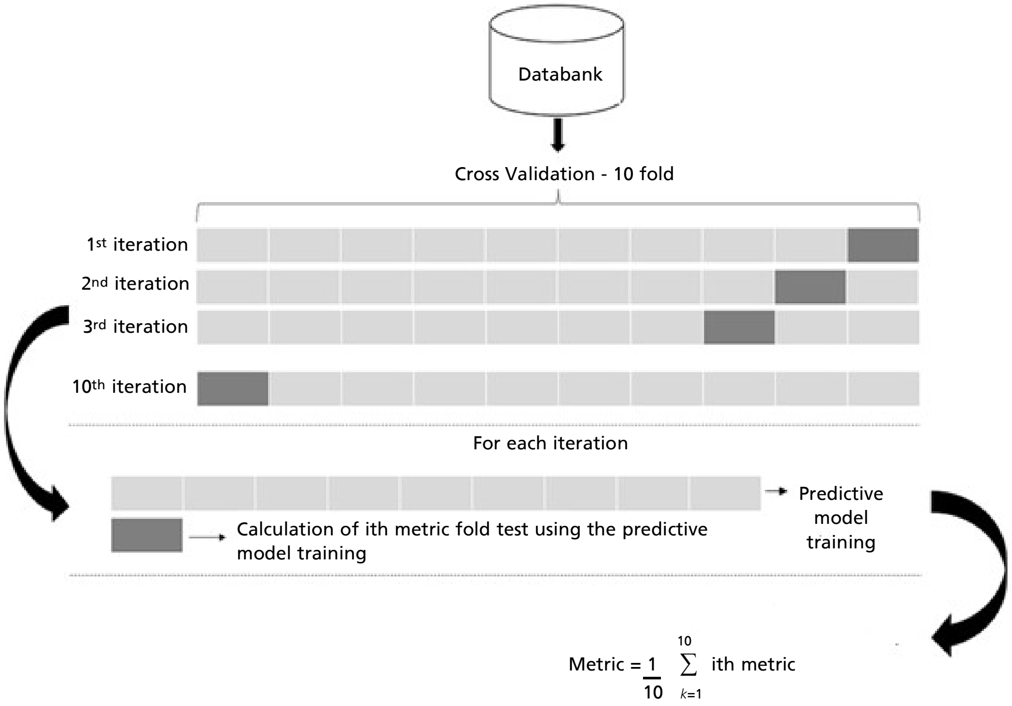

Databank
Cross Validation - 10 fold
1st iteration
2nd iteration
3rd iteration
10th iteration
For each iteration
Predictive model training
Calculation of ith metric fold test using the predictive model training
Metric = 1/10 ∑ k=1 10 ith metric

**Table 1**

Clinical features of elderly people with COVID-19 during the period March 12 to June 19, 2020.

| Cases | Total (N=7486) | | Recovered (N=4356) | | Deaths (N=3130) | |
|---|---|---|---|---|---|---|
| | n | % | n | % | n | % |
| Sex | | | | | | |
| Female | 3821 | 51.04 | 2369 | 62.0 | 1452 | 38.0 |
| Male | 3665 | 48.95 | 1987 | 54.22 | 1678 | 45.78 |
| Age ($\bar{X}$±SD) | 72.94 (±9.55) | | 70.95 (±9.06) | | 75.70 (±9.52) | |
| Age Group (years) | | | | | | |
| 60-69 | 3245 | 43.35 | 2288 | 52.53 | 957 | 30.58 |
| 70-79 | 2314 | 30.91 | 1252 | 28.74 | 1062 | 33.93 |
| 80 or more | 1927 | 25.74 | 816 | 18.83 | 1111 | 35.50 |
| Comorbidities | | | | | | |
| Cardiovascular Disease | 1298 | 17.34 | 79 | 6.09 | 1219 | 38.95 |
| Diabetes Mellitus | 1081 | 14.44 | 426 | 181 | 655 | 20.93 |
| Chronic respiratory diseases | 246 | 3.29 | 30 | 0.69 | 216 | 6.90 |
| Chronic renal disease | 136 | 1.82 | 5 | 0.11 | 131 | 4.19 |
| Neurological disease | 103 | 1.38 | 6 | 0.14 | 97 | 3.10 |
| Neoplasms | 93 | 1.24 | 3 | 0.07 | 90 | 2.88 |
| Smoking | 30 | 0.40 | 2 | 0.05 | 28 | 0.89 |
| Alcoholism | 12 | 0.16 | 0 | - | 12 | 0.38 |
| Signs and Symptoms | | | | | | |
| Cough | 4860 | 64.92 | 2766 | 63.50 | 2094 | 66.90 |
| Fever | 4403 | 58.82 | 2485 | 57.05 | 1918 | 61.28 |
| Dyspnea | 3773 | 50.0 | 1529 | 35.10 | 2244 | 71.69 |
| Saturation< 95% | 2614 | 34.92 | 705 | 16.18 | 1909 | 60.99 |

Data SEPLAG PE.

**Table 2**

Metrics to evaluate the performance of the Random Forest classifier.

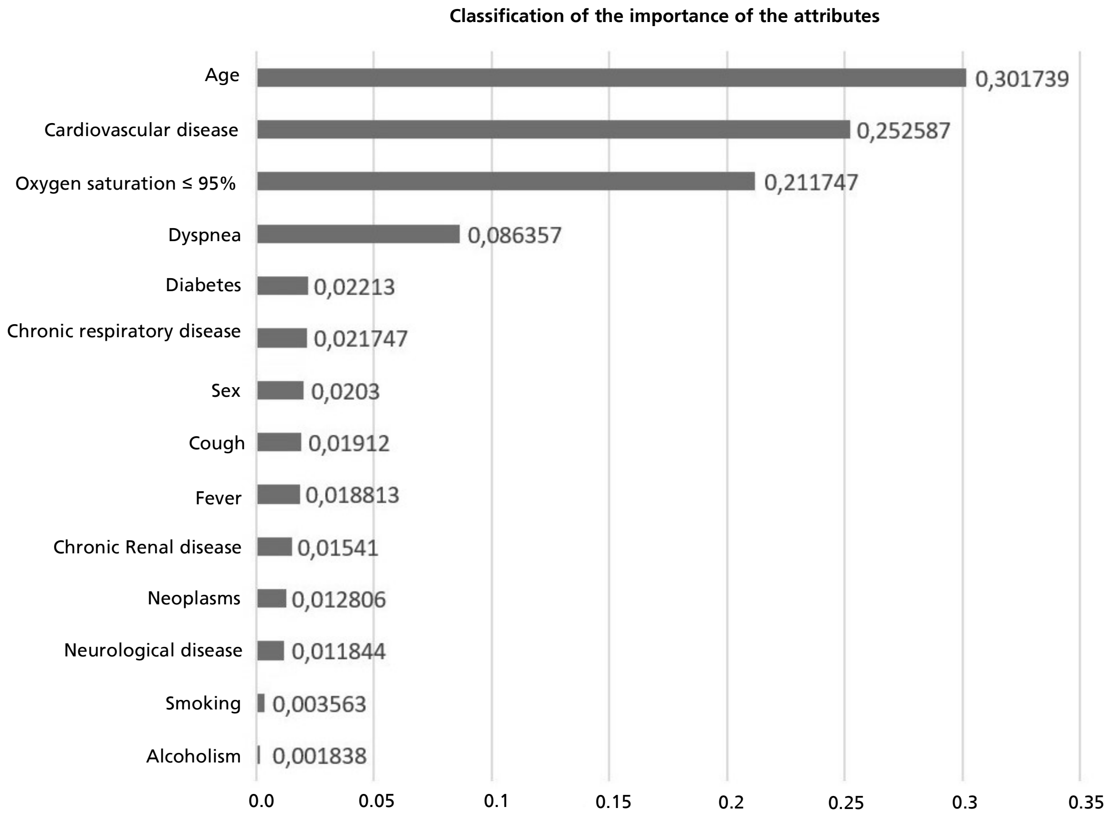
| True Positive Rate | False Positive Rate | Accuracy | Sensitivity | AUC ROC | Outcome |
|---|---|---|---|---|---|
| 0.848 | 0.306 | 0.794 | 0.848 | 0.839 | Recovered |
| 0.694 | 0.152 | 0.767 | 0.694 | 0.839 | Death |
| 0.784 | 0.241 | 0.783 | 0.784 | 0.839 | Weighted average |

(14.44%), and Chronic Respiratory Disease 246 (3.29%). Neoplasms were present in patients, 93 (1.24%) patients were smokers and 12 (0.16%) drinkers as shown in Table 1.

The RF classifier was able to hit the outcome of of patients in the database. To measure the performance of the classification, a confusion matrix was created, and some metrics were adopted, as shown in Table 2. It is possible to see that to predict the outcome of deaths, the RF algorithm showed a sensitivity of 0.784 and an accuracy rate of 0.783, also obtaining an Area under the ROC curve (AUC) of 0.839. Furthermore, the importance of the attributes showed that age (0.302), the presence of cardiovascular disease (0.252) and oxygen saturation less than or equal to 95% (0.212) are the three most important features for the evolution of elderly patients to die of COVID-19, as shown in Figure 2.

**Figure 2**

The importance of the attributes: analyzing how many nodes of the trees, which use a given attribute, reduce the overall impurity of the forest.



## Discussion

Age was the most important attribute related to death, with an importance of 0.302. While the overall lethality rate in Pernambuco at the end of the first three months of the pandemic was 8.25%,[12] the lethality rate for elderly patients in the same period was 41.81%.This value was much higher than the rates found in the literature, which ranged from 5.6% to 28.6%.[13,14] The analysis of lethality by age group also showed higher rates than those presented in Italy, where fatal cases increased mainly after 70 years of age, as 12.5% in the 70-79 years old range, 19.7% in the 80-89 years range, and 22.7% after 90 years.[15] It is worth noting that the high lethality rates found in Pernambuco reflect a period when testing was not widely available.

Several articles also show that the presence of comorbidities is a risk factor for adverse clinical outcomes such as death,[16-21] with cardiovascular disease always being one of the most prevalent comorbidities in the samples analyzed. In this study, the RF algorithm showed that cardiovascular diseases were the second most important feature for predicting death in elderly people with COVID-19, with a value of 0.252. Although, COVID-19 is best known for causing damage to the respiratory system, it is also known that it can compromise or worsen cardiovascular parameters. Furthermore, a retrospective study showed that 33% of deaths of COVID-19 were attributed to cardiorespiratory failure and 7% to isolated heart failure.[22]

The third variable highlighted for death prediction, with an importance value of 0.212, was peripheral oxygen saturation of≤ 95%, in agreement with the current literature.[23] The Ministry of Health even considers the diagnosis of Severe Acute Respiratory Syndrome (SARS) for every individual, of any age, with influenza syndrome and presenting signs of hypoxemia, such as the saturation of $O_2$≤ 95% in room air.[24] Furthermore, studies emphasize that early recognition of hypoxia and administration of oxygen has been shown to reduce mortality for patients with COVID-19.[25]

In conclusion, this study showed that the RF algorithm was able to reveal the most important aspects for predicting death in elderly patients with COVID-19, the three most important aspects are: advanced age, the presence of cardiovascular disease, and evidence of a peripheral saturation of $O_2$<95%. Furthermore, it was possible to see that the algorithm was able to correctly predict the outcome in 78.33% of the patients, obtaining an AUC of 0.839.

## Authors' contribution

All authors contributed equally to the construction of this article.

## References

1. Brasil. Ministério da Saúde. Brasília, DF; 2020 [acesso 5 dez 2020] Disponível em: https://susanalitico.saude.gov.br.
2. Thuler L, Melo A. Sars-CoV-2/Covid-19 em Pacientes com Câncer. Rev Bras Cancerol. 2020;66 (2): e-00970
3. Uddin S, Khan A, Hossain M, Moni M. Comparing different supervised machine learning algorithms for disease prediction. BMC Med Inform Decis Mak. 2019; 19 (1): 1-16.
4. Dash M, Liu H. Feature selection for classification. Intelligent Data Analysis. 1997; 1 (1-4): 131-56.
5. Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res. 2003; 3: 1157–182.
6. Breiman L, Friedman J, Stone C, Olshen R. In: Chapman and Hall. Classification and regression trees. First edition. Wadsworth, New York: CRC Press; 1984.
7. BreimanL. Random forests. Machine learning 2002; 45: 5–32.
8. Scikit-learn: Machine Learning in Python, Pedregosa F, et al. J Mach Learn Res. 2011; 12: 2825-30. .
9. Stone M. Cross-validatory choice and assessment of statistical predictions. J R Stat Soc Ser B Methodol. 2018; 36 (2): 111-33.
10. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Second Edition. USA: Springer; 2009
11. Van Rossum G, Drake FL. Python 3 Reference Manual. Scotts Valley, CA: CreateSpace; 2009.
12. Centro de Informações Estratégicas Vigilância em Saúde Pernambuco. Novo Coronavírus (COVID-19) Atualizações Epidemiológicas SES/PE. Recife, Brasil;2020. [acesso 5 dez 2020]. Disponível em: https://www.cievspe.com/novo-coronavirus-2019-ncov.
13. Zhang L, Zhu F, XieL, Wang C, Wang J. Clinical characteristics of COVID-19-infected cancer patients: a retrospective case study in three hospitals within Wuhan, China. Ann Oncol. 2020; 31 (7): 894-901.
14. Epidemiology Working Group for NCIP Epidemic Response, Chinese Center for Disease Control and Prevention. The epidemiological characteristics of an outbreak of 2019 novel coronavirus diseases (COVID-19) in China. Zhonghua Liu Xing Bing Xue Za Zhi. 2020; 41 (2): 145-51.
15. Edward KB. Coronavirus Disease 2019 (COVID-19) in Italy. JAMA. 2020; 323 (14): 1335.
16. Huang PC, Wang Y, Li PX, Ren PL, Zhao PJ, Hu Y. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet. 2020; 395 (10223): 497-506.
17. Chen PN, Zhou PM, Dong X, Qu PJ, Gong F, Han Y. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. Lancet. 2020; 395 (10223): 507-13.
18. LiuJ, LiuY, XiangP, PuL, XiongH. Neutrophil-to-Lymphocyte Ratio Predicts Severe Illness Patients with 2019 Novel Coronavirus in the Early Stage. 2020 [acesso 5 dez 2020]. Disponível em: https://www.medrxiv.org/content/10.1101/2020.02.10.20021584v1.full.pdf.
19. MaJ, Jing Y,Qian Y, Wu Y. Clinical characteristics and prognosis in cancer patients with COVID-19: A single center's retrospective study. J Infect. 2020; 81 (12): 318-56.
20. GuanW,LiangW, ZhaoY, LiangH, ChenZ. Comorbidity and its impact on 1590 patients with COVID-19 in China: a nationwide analysis. Eur Respir J. 2020; 55 (5): 2000547.
21. Ferreira J, Lima F, Oliveira J, Cancela M, Santos M. Covid-19 e Câncer: Atualização de Aspectos Epidemiológicos. Rev Bras Cancerol. 2020; 66:e-1013.
22. Ruan Q, Yang K, Wang W, Jiang L, Song J. Clinical predictors of mortality due to COVID-19 based on an analysis of data of 150 patients from Wuhan, China. Intensive Care Med. 2020; 46 (5): 846-8.
23. GuanW, NiZ, HuY, LiangW, Chun-quan OU. Clinical Characteristics of Coronavirus Disease 2019 in China. N Engl J Med. 2020; 382: 1708-20.
24. Brasil. Ministério da Saúde. Protocolo de Manejo Clínico do Coronavírus (COVID-19) na Atenção Primária à Saúde (v9). Brasília, Brasil; 2020. [acesso 5 dez 2020]. Disponível em: https://portaldeboaspraticas.iff.fiocruz.br/biblioteca/protocolo-de-manejo-clinico-do-coronavirus-covid-19-na-atencao-primaria-a-saude/
25. Sun Q, Qiu H, Huang M, Yang Y. Lower mortality of COVID-19 by early recognition and intervention: experience from Jiangsu Province. Ann Intensive Care. 2020; 10:33.

Received on October 16, 2020

Approved on February 8, 2021