Submission date: 30/08/2021
Resubmission date: 13/12/2021
Camera ready submission date: 16/02/2022

1st round notification: 16/11/2021
2nd round notification: 11/02/2022
Available online: 25/03/2022
Publication date: 25/03/2022

Section: extended version
Extended versions of selected papers from the XVII Brazilian Symposium on Information Systems (SBSI 2021)

Predicting COVID-19 hospitalizations with attribute selection based on genetic and classification algorithms

Miriam Pizzatto Colpo[1,2], Bruno Cascaes Alves[3], Kevin Soares Pereira[3], Anna Flávia Zimmermann Brandão[1], Marilton Sanchotene de Aguiar[1], Tiago Thompsen Primo[1]

[1]Graduate Program in Computing – Federal University of Pelotas (UFPel)
Pelotas, Rio Grande do Sul – Brazil

[2]Information Technology Directorate – Federal Institute of Education, Science and Technology Farroupilha (IFFar)
Santa Maria, Rio Grande do Sul – Brazil

[3]Technological Development Center – Federal University of Pelotas (UFPel)
Pelotas, Rio Grande do Sul – Brazil

{miriam.colpo, bcalves, kspereira, anna.flavia, marilton, tiago.primo}@inf.ufpel.edu.br

Abstract. *The COVID-19 pandemic has been pressuring the whole society and overloading hospital systems. Machine learning models designed to predict hospitalizations, for example, can contribute to better targeting hospital resources. However, as the excess of information, often irrelevant or redundant, can impair predictive models' performance, we propose a hybrid approach to attribute selection in this work. This method aims to find an optimal attribute subset through a genetic algorithm, which considers the results of a classification model in its evaluation function to improve the hospitalization need prediction of COVID-19 patients. We evaluated this approach in two official databases from the State Health Secretariat of Rio Grande do Sul, covering COVID-19 cases registered up to October 2020 and June 2021, respectively. As a result, we provided an increase of 18% in the classification precision for patients with hospitalization necessities in the first database, while in the second one, considering a temporal evaluation with sliding window, this gain was on average 6%. In a real-time application, this would also mean greater precision in targeting resources and, consequently and mainly, improved service to the infected population.*

Keywords. *Feature selection, COVID-19, Genetic algorithm, Machine learning, Hospitalization prediction.*

1. Introduction

COVID-19 is an infectious disease caused by Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), initially identified in China in December 2019. Af-

Cite as: Colpo, M. P., Alves, B. C., Pereira, K. S., Brandão, A. F. Z., Aguiar, M. S. de, & Primo, T. T. (2021). Predicting COVID-19 hospitalizations with attribute selection based on genetic and classification algorithms. iSys: Revista Brasileira de Sistemas de Informação (Brazilian Journal of Information Systems), 15(1), 4:1-4:30. DOI: 10.5753/isys.2022.2187.

ter spreading rapidly around the world, the Coronavirus 2, also called new coronavirus, infected over 79 million people in just one year, bringing over 1.7 million to death [World Health Organization 2020]. Naturally, one of the pandemic sectors most affected is health, which, in addition to the challenges posed by a new disease, has faced the burden of hospital systems and the scarcity of resources. During the COVID-19 pandemic, according to the Pan American Health Organization (PAHO), information systems are essential for the proper performance of primary health care functions [PAHO 2020]. In addition to the agility and numerous benefits of electronic medical records, health information systems integrated with other local and national systems provide a large volume of data with high strategic potential.

Machine Learning (ML) techniques can be applied in the automated analysis of this data to obtain valuable knowledge, which assists in resource management and decision making. In particular, predictive models, based on previous experiences, can be constructed from classification algorithms [Alpaydin 2010], which seek to extract patterns from the data and thus describe and distinguish data classes or concepts [Han et al. 2011]. Although several classification techniques and the most robust ones, such as for ensembles, tend to perform better, in the health context, the interpretability of predictive results is of great importance, especially considering COVID-19, which is a disease whose knowledge is constantly updating. In this sense, Decision Tree (DT) algorithms are attractive options, as, in addition to predictions, they allow us to know the patterns identified in the data and observe their trends. As the name suggests, these algorithms construct hierarchical classification models, with tree-shaped structured patterns [Funchal and Adanatti 2016], in which each inner node represents a test on an attribute; each branch represents a test result, and each leaf node indicates a class label (decision/prediction) [Han et al. 2011].

Although DT models tend to provide good predictive results and are widely adopted, due to their simplicity and interpretability, in different application areas such as medicine, financial analysis, and molecular biology, they are very susceptible to overfitting [Han et al. 2011]. In other words, while learning, these classifiers can incorporate, due to noise or outliers, anomalous patterns from the training data, which do not adequately represent the overall dataset [Han et al. 2011]. To avoid the overfitting problem, we can establish hyperparameters that limit the growth or depth of the tree. In addition, the dimensionality reduction via attribute/feature selection is beneficial for unregularized models, such as DTs [Raschka and Mirjalili 2017].

We can use automatic attribute selection techniques to reduce processing cost and prevent irrelevant or redundant dimensions of the data from impairing the generalizability and, consequently, the performance of classification/prediction models, additionally providing simpler and clearer models [Han et al. 2011]. Attribute selection techniques aim to find the minimum subset of attributes that best contributes to model performance. However, because the evaluation of the entire solution space is unfeasible, heuristics are used to search for an optimal subset, which does not necessarily correspond to the optimal global [Alpaydin 2010]. In this context, we can use genetic algorithms (GA), which are bio-inspired meta-heuristics whose search and optimization process is based on the theory of biological evolution, considering the principle of survival of the individual (or

solution) more apt [Linden 2008].

Thus, considering that predicting and knowing the patterns of hospitalization needed for COVID-19 patients would be of great importance for the development of tools to support decision-making, we propose a hybrid approach to attribute selection, composed of a GA and a DT classifier, to improve the results of an interpretable model for predicting the need for hospitalization of COVID-19 patients. We used a GA heuristic to search for a subset of optimal attributes and DT models to evaluate the aptitude of the subsets of candidate attributes. Notably, although the target variable of our model is hospitalization, we are not proposing to use our approach to decide to admit or not a patient. This decision should always depend exclusively on the clinical situation and the medical evaluation of each patient. However, considering the context of the pandemic, in which the scarcity of resources in health is even more remarkable, an accurate and interpretable model for predicting the need for hospitalization of COVID-19 patients can create opportunities, based on knowledge of patterns and predictions, for better management of resources and services that are not linked to severe neglect of patients. An example of an application would be the prioritization of home follow-up contacts, where there is already a natural lack of assistance due to the health teams' impossibility to contact all patients undergoing home treatment daily.

We initially evaluated our approach in an official database, covering cases of COVID-19 patients registered by the State Health Secretariat of Rio Grande do Sul (in Portuguese, *Secretaria Estadual da Saúde do Rio Grande do Sul* – SES/RS), until October 2020. As a result, we provided an average increase of 18% in the classification precision for patients with hospitalization necessities. We published our approach and its initial evaluation in [Colpo et al. 2021], at the last Brazilian Symposium on Information Systems (SBSI), where we received an honorable mention and the invitation to submit this extended version.

As a differential, we use an additional database, updated with COVID-19 cases registered until June 2021, to expand the previous evaluation and better simulate the reality of future predictions. Considering the dynamic character of the contagion and the disease, in this new assessment, we used different temporal divisions of the training and test data, in order to verify, in the long term, the performance stability of our model to predict the need for hospitalization of COVID-19 patients. In addition, for a fairer comparison between the results of adopting our attribute selection approach, we started to consider automatic optimization of hyperparameters in the model's development. With this additional process of hyperparameters optimization, we intend to include a strategy to avoid the overfitting problem in the scenario of non-adopting attribute selection and boost learning of the minority class in both scenarios (with or without attribute selection). Considering a temporal evaluation with a sliding window, we provide an average increase of 6% in the classification precision for patients with hospitalization necessities, selecting attributes for each temporal division of training. Finally, it is worth mentioning that, in addition to detailing the new database, the methodology, and the results related to this extension of the evaluation, we included in this article details about the pre-processing carried out on the data, which, although important, had been omitted in the previous publication due to space limitations.

We organized this article as follows: in Sections 2 and 3, we present some related works and details of the proposed approach; in Section 4 we describe the technical/methodological decisions that guided the development of this study; and finally, in the Sections 5 and 6, we discuss the results obtained and present the conclusions of this work, respectively.

2. Related Work

Several studies have explored ML techniques to predict diseases and assist health professionals in decision-making. In [Lynch et al. 2017] and [Pradeep and Naveen 2018], for example, authors evaluate different ML algorithms in predicting lung cancer patients' survival, considering that this information has great importance in determining the care or treatments to be adopted. In [Lynch et al. 2017], authors present an ensemble model to predict the number of months patients would survive after being diagnosed. This ensemble was composed of linear regression models, DT, Gradient Boosting Machines (GBM), and Support Vector Machines (SVM). Results presented the best performance when compared to isolated algorithms. Also, in [Pradeep and Naveen 2018] a DT, generated from the C4.5 algorithm, presented the best performance to predict whether patients are more than a year old when compared, in large data sets, to the algorithms *Naive Bayes* and SVM. Also related to lung diseases, in [Heckler et al. 2020] a tool was developed to predict the abandonment of patients from a Pulmonary Rehabilitation Program, considering models based on DT, SVM, and Randon Forest (RF), which is an ensemble composed of DTs.

More directed to the context of COVID-19, [Arvind et al. 2021] and [Burdick et al. 2020] propose predictive models of intubation and mechanical ventilation for COVID-19 patients, aiming to facilitate the identification of high-risk cases and the allocation of hospital resources, such as respirators. For this, both works use ensembles of DTs, considering the RF and XGBoost algorithms, respectively. Although the studies mentioned above, related to cancer patients and pulmonary rehabilitation, consider manually selection of attributes in the pre-processing of the data, based on the previous knowledge of specialists, this approach may not be a good option for the domain of COVID-19, considering its novelty [Arvind et al. 2021]. Thus, automatic attribute selection techniques are a possible alternative for reducing the dimensionality of COVID-19 data to improve classification/prediction results. In the pre-processing data context, we can perform the automatic selection of attributes by methods such as (i) filter – when considering the characteristics of the training data, based on specific metrics, and order the relevance/importance of the attributes without involving ML algorithms; and, (ii) wrapper – when use a sorting algorithm to evaluate the performance provided by each subset of attributes evaluated in the search space during the selection process [Cueto-López et al. 2019]. In the context of predicting the risk of colorectal cancer, for example, [Cueto-López et al. 2019] evaluates the use of attribute selection techniques to improve the performance of different ML models and avoid overfitting. The authors obtained the best results by selecting the wrapper of SVM and Pearson correlation filter, preceding the training of SVM and logistic regression models, respectively. The paper presented in [Monteiro et al. 2020] also addresses the selection of attributes as an

essential technique for improving predictive performance, however, in the context of mortality in intensive care units (ICUs). Among other multivariate statistical analysis methods, the authors analyze principal components to reduce dimensionality.

Although less common, researchers also used genetic algorithms as meta-heuristics to search for subsets of attributes, obtaining good results, such as presented in [Zhou et al. 2021, Maleki et al. 2021, Pawlovsky and Matsuhashi 2017]. In [Zhou et al. 2021], they treated attribute selection as a multi-objective optimization problem and proposed the PS-NSGA (Problem-Specific Non-dominated Sorting Genetic Algorithm). Authors based their approach on the *framework* NSGA-III and sought to minimize the rate of incorrect classifications, calculated from a classifier k-Nearest-Neighbors (kNN); the proportion of selected attributes; and a distance metric, which aims at greater generalization capacity. The method surpassed other evolutionary algorithms and more traditional selection techniques by being evaluated using different disease databases in predictive tasks. Additionally, in the context of predicting the prognosis of patients with breast and lung cancer, respectively, [Pawlovsky and Matsuhashi 2017] and [Maleki et al. 2021] use GA to select attributes and thus improve the results of kNN-based prediction models. Both studies use small databases containing information from a maximum of 1000 patients and address attribute selection as a single-goal optimization problem. However, in the [Maleki et al. 2021] work, GA uses an evaluation function that combines the number of selected attributes and the classification result obtained by kNN, causing both factors to be weighted.

As [Arvind et al. 2021] and [Burdick et al. 2020], this work is inserted in the context of predictions related to COVID-19. However, the hospitalization need of patients is addressed here rather than the need for respiratory support. In addition, although we use automatic attribute selection, as [Cueto-López et al. 2019] and [Monteiro et al. 2020], to improve the predictive model, we propose a hybrid selection approach, which uses a GA in the search for an optimal subset of attributes, addressing the selection process as a single-goal optimization problem, as in [Pawlovsky and Matsuhashi 2017] and [Maleki et al. 2021]. Finally, despite we based our approach in [Maleki et al. 2021] work, in addition to a different context, we considered large databases, with information from more than 200 thousand patients, which motivated different technical decisions. We will describe these methodological differences in Section 4 properly.

3. Proposed Approach

As mentioned before, this study seeks to reduce the dimensionality of patients' data from COVID-19 to allow classification algorithms to concentrate their learning processes on more relevant attributes and provide better results for predicting the need for hospitalization of these patients. For this, we proposed a hybrid approach of attribute selection, composed of a GA, responsible for the search for a subset of optimal attributes; and by a classification algorithm, designed to test the predictive capacity of each subset of attributes and, thus, to promote their evaluations during the evolutionary process of GA, as illustrated in Figure 1.

As a heuristic method based on biological evolution, the GA starts from an **initial population**, in which each individual represents a solution to be evaluated, and leads

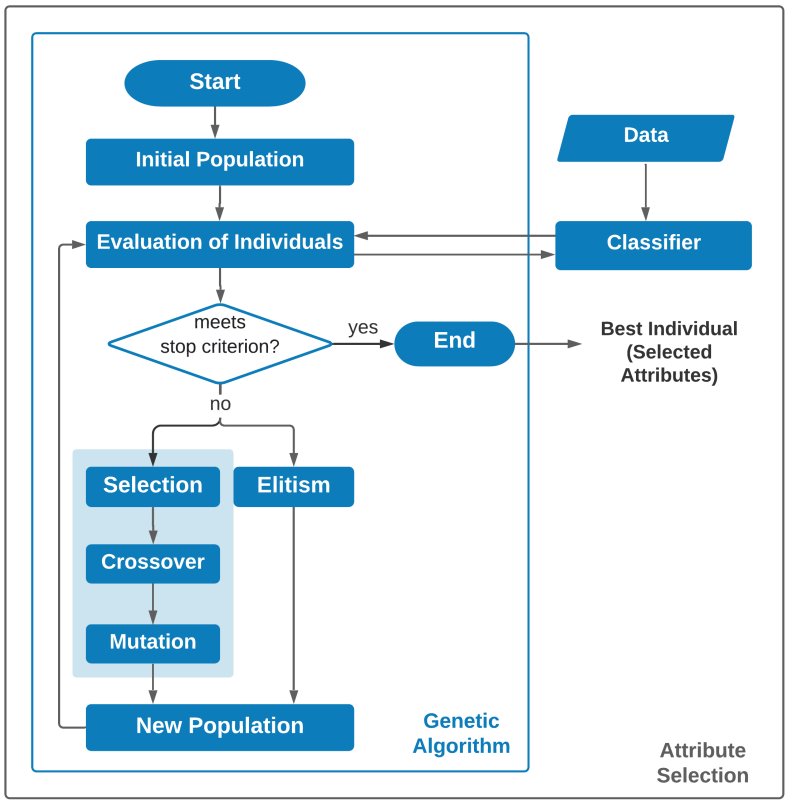

Figure 1. Overview of our hybrid approach to attribute selection.

the search for an optimal solution, which may not be global, based on the principle of survival of the fittest individual [Linden 2008]. The evolutionary process occurs through genetic operators, which are applied to the population, generation by generation, benefiting the fittest individuals until they meet a particular stop criterion (usually a maximum number of generations), and the best solution is found [Zhou et al. 2021]. Basically, with each generation, the population's individuals are evaluated by a function of **evaluation**, which determines the aptitude of each individual, that is, how satisfactory is its solution to the problem treated [Linden 2008]. This aptitude is then considered by an operator of **selection**, in the choice of which individuals will participate, as parents, in the reproductive process. The descending individuals' determination, which will compose the next generation population, is then performed by the **crossover** and **mutation** operators, considering individual probabilities. For example, while crossing two previously selected parents exchange parts of their genetic material to form two new child individuals, the mutation operator arbitrarily incorporates new genetic traits in these children, ensuring genetic diversity to the **new population** [Maleki et al. 2021, Linden 2008].

Notice that the evaluation function significantly influences the evolutionary process since the best-evaluated individuals are more likely to be selected for crossing and, consequently, reproduce and perpetuate their characteristics. Also, when GA considers the **elitism** strategy, the best individuals of a generation are added to the next, ensuring the preservation of the best characteristics and that the performance of the GA never de-

teriorates in the course of evolution [Linden 2008].

In the problem considered here, each individual of the GA corresponds to a specific subset of attributes that is one of the possible solutions for selecting attributes in the search space. We aim to improve the prediction results of COVID-19 patients' hospitalizations, then each individual's aptitude should be evaluated from the impact that its attributes cause in the prediction. Therefore, the proposed approach integrates a **classification algorithm** (classifier) to the attribute selection process, allowing each individual's evaluation to trigger the training and testing of a new prediction model, considering only the data related to the subset of attributes under evaluation. The influence of each subset of attributes on the prediction performance is now considered in its respective individual's aptitude, composing the evaluation function of GA. Although based on the study by [Maleki et al. 2021], mentioned in Section 2, this work, in addition to considering another domain of prediction (need for hospitalization of COVID-19 patients), contains several methodological differences, which we will explain in the next Section.

4. Methodology

Initially, concerning the technologies and tools used, we implemented the proposed approach in the Python language. We used the Pandas[1] and NumPy[2] libraries for data loading and manipulation, and we used the Scikit-Learn[3] package in data pre-processing and in the classification/prediction process. Details about the data used, the attribute selection process' implementation, and the construction and evaluation of predictive models are presented in Sections 4.1, 4.2 and 4.3, respectively.

4.1. Datasets

In this work, we used the SES/RS data, which contains the registered cases of COVID-19 in the Brazilian State of Rio Grande do Sul. Initially, as the data are updated daily, a copy of the database[4] was created and stored, referring to 10/21/2020, for the experiments' later reproducibility. Moreover, as the government considers recovered the patients without hospitalization after 14 days of the onset of symptom manifestation [SES/RS 2020], patients with symptom onset date posterior to 10/07/2020 were removed from this database because they are considered unstable, that is, without definitive hospitalization. Therefore, as shown in Table 1, the records of 219,343 COVID-19 patients were kept in this dataset (Database I), of which 18,832 (8.59%) correspond to hospitalized patients. We divided the resulting database into a training-validation set and another test set, in the usual proportions of 70% and 30%, respectively. We used the training-validation database to select attributes, specifically in the training and validation of the models provided by the subsets of attributes evaluated in the GA; the test base was reserved and used to evaluate the final predictive model after selecting attributes.

In addition, in order to extend the predictive model evaluation and thus better verify the impact of attribute selection, an updated database[5] was collected, referring to

[1] https://pandas.pydata.org
[2] https://numpy.org
[3] https://scikit-learn.org
[4] https://bit.ly/sesrs_covid_data_10212020
[5] https://bit.ly/sesrs_covid_data_07282021

	Database I (10/06/2020)	Database II (06/30/2021)
Hospitalized Patients	18832 (8.59%)	77208 (7.54%)
Non-Hospitalized Patients	200511 (91.41%)	946763 (92.46%)
Total Records/Patients	219343	1023971

Table 1. Quantitative of patients and hospitalizations by database.

07/28/2021. Still considering the stability of the data, we considered only the records of patients with symptom onset until 06/30/2021, that is, until the end of the first semester of 2021. Thus, as shown in Table 1, the records of 1,023,971 patients were kept in this new dataset (Database II), of which 77,208 (7.54%) correspond to hospitalized patients. We can note that, besides covering a more significant period, this new database includes a context of expanding the circulation of the SARS-CoV-2 variants [Faria et al. 2021], which allows evaluating the stability of the predictive model over time and, consequently, over the pandemic evolution. We will present details about the database segmentation in Section 4.1, together with the description of this temporal evaluation.

4.1.1. Pre-processing

We preliminarily removed some duplicated or unsuitable attributes during the data pre-processing:

- The pregnancy indicator attribute also had its information registered in the multi-categorical attribute of conditions/comorbidities, being unnecessary;
- Location attributes (as country of birth, region, city, and neighborhood) in addition to not having a direct relationship with the hospitalization need, could impair the models' generalizability since more populous cities or regions naturally have more records of hospitalizations;
- We considered uninformative the attribute that indicated whether patients are health professionals, as all hospitalized patients had the value "uninformed";
- The attribute of indigenous ethnicity had an empty or "uninformed" value for the vast majority of records, being more suitable to keep only the indigenous indication, in the race/color attribute;
- As the objective is to predict the hospitalization need of the COVID-19 patients, we do not consider attributes related to their evolution and death since this information refers to future events;
- Practically all hospitalized patients presented Severe Acute Respiratory Syndrome. However, as this condition tends to be a post-hospitalization evolution, this attribute was considered inadequate for early predictions.

Additionally, attributes that demonstrated to leak information related to the hospitalization event also needed to be discarded since they would disqualify any attempt to predict:

- We excluded the information source attribute indicating whether the case of COVID-19 from a hospital stay, a health care on sentinel unit, etc.;

- We removed the attribute related to diagnostic criteria because molecular tests were mostly performed in hospitals, especially at the pandemic's beginning.

We treated missing data after removing the inappropriate attributes for the predictive task. Among the remaining attributes, those related to conditions/comorbidities and symptoms have missing values. However, considering that this information, due to its importance, would hardly go unrecorded by health professionals, these absences were understood as inexistence of symptoms or special conditions. Thus, while the multilabel attribute of conditions had its empty values filled with the value "no condition/comorbidity", the categorical attributes of symptoms (dyspnea, fever, cough, throat, and other symptoms) had their missing values filled with "No".

Finally, as the classification algorithms of the Scikit-Learn package only work with predictive attributes of numerical types, transformations were carried out on the non-numeric attributes. The symptom onset date and disease confirmation date attributes were processed and gave rise to two new temporal contextualization attributes: (*i*) days between symptom onset and COVID-19 confirmation, and (*ii*) days between pandemic onset[6] and symptoms onset. We binarized categorical attributes after removing the original date attributes by the one-hot coding technique. To get the desired result concerning the multi-categorical attribute "conditions/comorbidities" (that is, only one binary attribute for each condition), it was also necessary to correct some spelling errors identified in the condition records (some records had condition names with a missing letter, for example).

Thus, in addition to the target attribute (class), which indicates whether or not the patient was hospitalized, 51 attributes were used, related to social data, to indicate gender, age group, race/color, and deprivation of liberty; indicators of pre-existing symptoms and conditions/comorbidities; and temporal contextualization, according to Table 2.

4.2. Attribute Selection

As in this work, one GA individual corresponds to a subset of attributes, in the implementation, as in [Maleki et al. 2021], we represented each individual by a binary vector, in which each position represents a gene (attribute). Its value indicates whether a particular attribute is selected in the subset of attributes corresponding to the individual. In this case, considering that the database has 51 predictive attributes, as specified in Section 4.1, each individual is described as a binary vector of 51 positions.

Additionally, we defined some GA parameters according to the values empirically established in [Maleki et al. 2021]: populations of 20 individuals; crossing and mutation probabilities of 70% and 2%, respectively; and maximum amount of generations equal to 10 (stop criterion). Although GAs generally use a more significant number of generations, we need to consider that, in the attribute selection problem, the evaluation of each generation's individual involves training and testing a machine learning model, which requires a considerable computational cost (e.g., a GA with a population of 20 individuals and ten generations involves the construction and evaluation of 200 models). Therefore, in

[6]Here, we took the symptom onset date of the state's first COVID-19 patient as the onset date of the pandemic.

Table 2. Dataset attributes.

#	Description	Type
0	female	boolean
1	fever symptom	boolean
2	cough symptom	boolean
3	throat symptom	boolean
4	dyspnea symptom	boolean
5	other symptoms	boolean
6	deprivation of freedom	boolean
7	age range 1 to 4	boolean
8	age range 5 to 9	boolean
9	age range 10 to 14	boolean
10	age range 15 to 19	boolean
11	age range 20 to 29	boolean
12	age range 30 to 39	boolean
13	age range 40 to 49	boolean
14	age range from 50 to 59	boolean
15	age range from 60 to 69	boolean
16	age range from 70 to 79	boolean
17	age group > 80	boolean
18	age group < 1	boolean
19	race/color yellow	boolean
20	race/color white	boolean
21	race/color indigenous	boolean
22	uninformed race/color	boolean
23	race/color brown	boolean
24	race/color black	boolean
25	asthma	boolean
26	diabetes	boolean
27	diabetes mellitus	boolean
28	chronic cardiovascular disease	boolean
29	chronic haematological disease	boolean
30	chronic liver disease	boolean
31	chronic neurological disease	boolean
32	chronic kidney disease	boolean
33	chronic heart disease	boolean
34	advanced chronic kidney disease	boolean
35	chronic respiratory disease decompensated	boolean
36	pregnant	boolean
37	high-risk pregnant	boolean
38	immunodeficiency	boolean
39	immunosuppression	boolean
40	no condition/comorbidity	boolean
41	obesity	boolean
42	another chronic pneumatopathy	boolean
43	other conditions/comorbidities	boolean
44	chronic pneumatopathy	boolean
45	chromosomal disease or immuno fragility	boolean
46	puerpera	boolean
47	puerpera up to 45 days of childbirth	boolean
48	down syndrome	boolean
49	days between symptom onset and confirmation	numeric
50	days between pandemic onset and symptoms	numeric

preliminary tests, we also evaluated the use of 20 and 50 generations, with no significant changes in the GAs results, except for the execution time, which was severely impaired.

In contrast to [Maleki et al. 2021], the present work uses a strategy of elitism, where 2 (10%) best individuals of a generation are kept in the next one. In this study, the last individual receives the value 1 in all his genes to represent the complete set of attributes, although the initial population is generally randomly created. Thus, we ensure that only the subsets/individuals that provide the same prediction results as the initial set will be considered better during the evolutionary process. Additionally, in this study, different selection and crossing operators were implemented in the GA to evaluate their results and choose the combination to be adopted definitively. We evaluated the following selection methods:

Roulette wheel selection, in which we simulated a roulette, in which each individual of the population receives a piece proportional to its evaluation, and then selected the individual to whom belongs a certain position, previously selected [Linden 2008];

Tournament selection, in which we selected randomly k individuals of the population and, after, the best rating is chosen [Linden 2008]. In this study, we considered k=2;

Truncated selection, in which individuals in the population are ordered by their evaluation, in a decreasing way, and only the first k individuals participate in the selection process [Linden 2008]. In this work, we established the cut-off point k as 25% of the population (i.e., five individuals), and we used the two-individual tournament in the selection.

We implemented the following crossover methods as described in [Linden 2008]:

One-point crossover, in which a cutoff point is drawn in the representation of the individual, separating the parents into two parts, one to the left of the cutoff point and the other to the right. We generated the first child from genes' concatenation of the left part of the first parent with those of the right part of the second, while we constituted the second child from the remaining parts;

Two-point crossover, in which we have drawn two cutoff points so that the first child is formed by the parts of the first parent external to the cutoff points and by the part of the second parent that lies between the cutoff points, while the second child consists of the remaining parts;

Uniform crossover, in which we have drawn zero or one number for each gene/position of the individual representation. If the drawn value is one, the first child receives the gene from the first parent, and the other child receives the gene from the second parent. Otherwise, we reversed the assignments.

Regarding the classification algorithm integrated into the attribute selection process, we used a DT with the default configuration of the Scikit-Learn library, contrasting the kNN algorithm used in [Maleki et al. 2021]. This difference in classifier choice stems from the data-set used since, unlike [Maleki et al. 2021], in this work, an extensive database is used, with more than 150,000 records destined to the process of selecting attributes, which would make kNN processing very slow. In addition to faster, decision

trees often have good results and have straightforward interpretation, a significant differential to understanding the patterns found by the predictive models [Han et al. 2011]. It is also important to mention that the data used in selecting attributes were divided, stratified, in the proportions of 80% and 20% for training and validation, respectively, of the prediction models related to the subsets of attributes evaluated (individuals of the GA). This decision also considered the size of the databases and the time added by using a more robust separation and validation strategy, such as k-folds cross-validation.

Finally, while [Maleki et al. 2021] works with a balanced data-set and evaluates the subsets of attributes from the Misclassification Rate (MCR) that they provide in the classification; in this implementation, due to the unbalanced nature of the data, it was considered more appropriate to use the *F1-score* [Han et al. 2011] metric in the evaluation. Thus, although inspired by [Maleki et al. 2021], this work's evaluation function becomes a maximization (MaxZ) problem and no longer a minimization (MinZ) task. In the Equations 1 and 2 are presented, respectively, the evaluation function proposed by [Maleki et al. 2021] and the adaptation considered here.

$$MinZ = mcr(1 + \beta \cdot n_f) \tag{1}$$

$$MaxZ = f1_{score}(1 - \beta \cdot n_f) \tag{2}$$

In these Equations, n_f represents the number of attributes selected and β a factor/weight, between 0 and 1, multiplied by the number of attributes selected to penalize individuals who select a more considerable amount of attributes. That is, among two subsets of attributes that provide the same $f1_{score}$, for example, you will get a better evaluation (considered fitter) of the subset that contains fewer attributes. In this work, we used a low value (0.0001) for β in order to assign a soft penalty to the number of selected attributes and thus prioritized the classification result ($f1_{score}$) in the evaluation.

4.3. Classification/Prediction

As a result of the attribute selection process, we obtained the subset of attributes that best contributed to maximizing predictive task performance within the solution space evaluated by the GA. Thus, we removed the attributes that did not belong to this subset from the data, and we trained a predictive model to compare its results with those of the model trained on the entire set of attributes and verify the impact of attribute selection on the predictions' performance. We constructed these predictive models (with or without attribute selection) using the DT algorithm, previously adopted in the attribute selection process. It is essential to highlight that, although more robust models such as ensembles could provide better results, as mentioned in Section 1, our goal is to develop an interpretable model from decision tree algorithm and provide to it, through attribute selection, greater generalizability and, consequently, a better predictive performance. Therefore, our evaluation does not aim to compare different classification techniques but to verify the feasibility and influence of our attribute selection approach on the results of our interpretable model;

Initially, we trained the models using the training-validation dataset and the default parameterization of the Scikit-Learn. Then, the evaluation of these models was performed

on the test dataset, using the standard metrics of precision, recall, f1-score, and accuracy. Besides, we confirmed the results using 10-fold cross-validation. As described in Section 4.1, the training-validation dataset corresponds to 70% of the records in Database I, so that the other 30% composes the test dataset in a stratified division.

However, to better simulate the reality of the future predictions and thus assess the stability of predictive performance over time, considering the dynamic character of the contagion and the disease, the evaluation was extended to consider temporal divisions for training and testing data. As mentioned in Section 4.1, we performed this process using Database II, which is complete and updated, allowing that the analysis considers a more extended time. Figure 2 shows the monthly distribution of COVID-19 infections in Database II, considering the date of symptom onset as reference. As the COVID-19 contagion occurred progressively, the monthly number of cases in the first eight months (until October 2020) is naturally much smaller than the others. Thus, in order to mitigate this disparity and ensure an acceptable volume of data per set, the records were divided into six datasets (A, B, C, D, E, F), covering temporal ranges of different sizes, as indicated in Figure 2. Thus, while datasets A, B, and C cover February to July, August to October, and November to December 2020, datasets D, E, and F sequentially consider the first three bimesters of 2021.

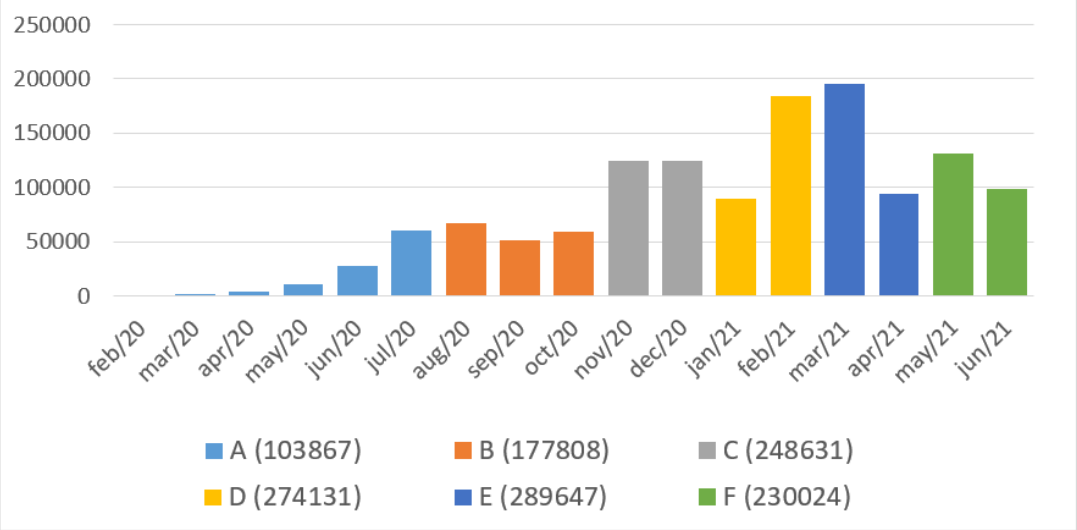

Figure 2. Relation between monthly occurrences of COVID-19 and the composition of datasets A, B, C, D, E, and F, from Database II. In the color legend, each dataset is accompanied, in parentheses, by its quantitative of records.

After the temporal data separation, we used the six sets to carry out nine experiments, varying the training and test sets' composition to create and evaluate different predictive models. In addition to simulating predictions in future data, this evaluation aims to assess the stability of patterns related to the need for hospitalization and, consequently, the predictive performance. As can be seen in Figure 3, the first five experiments consider all datasets in the sequential and iterative expansion of the training set. Notice that, although inspired by the walk-forward validation strategy, in these five experiments,

we did not select as the test set only the set immediately after the training set. Considering the example of the second experiment, which has sets A and B in the training set, instead of only selecting set C as the test set, we keep all the remaining data in the test set. With that, we want to verify whether we maintained the predictive power provided by initial data over the long term. In the last four experiments, in turn, we consider the walk-forward validation with a sliding window width of two training sets. That is, we advance the validation over time, sequentially updating the two sets that make up the training data and using the next set for testing. These experiments allow us to verify how the model behaves when updated and trained only on data closest to each prediction context. This fact is important because the pandemic presented different contexts throughout its evolution, and each of these contexts can affect hospitalization patterns and, consequently, the model's predictive performance.

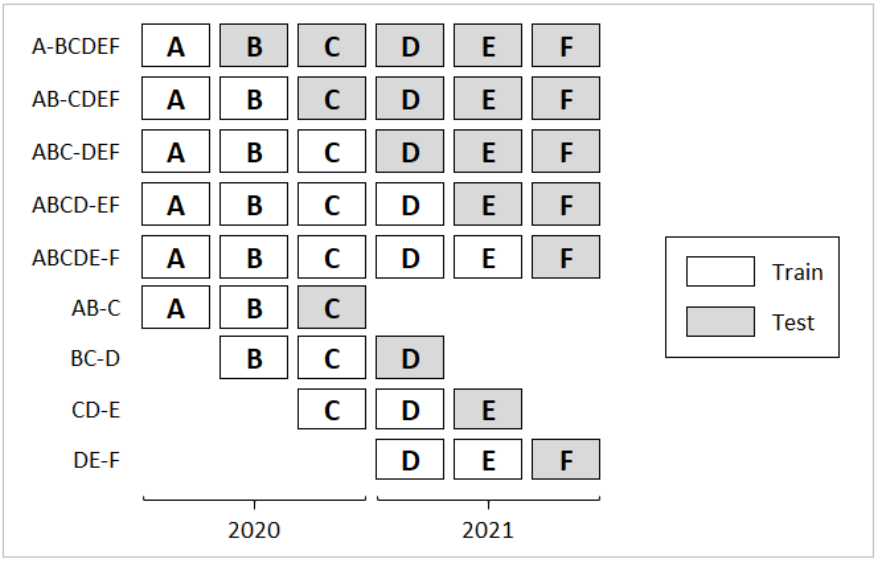

Figure 3. Relation between the temporal divisions (datasets A, B, C, D, E, and F) and the composition of the training and test sets of each experiment performed on Database II.

As with the previous evaluation, we performed each experiment restricting the attributes to the result of the selection performed on its respective training set. That is, we developed models for each temporal division of training from Database II, considering the entire set of attributes or just the subset selected by our genetic approach for each of these partitions. In addition, for a fairer comparison between the results of adopting or not our attribute selection approach, all models developed in this extended/temporal evaluation had their hyperparameters optimized, using the Scikit-Learn GridSearchCV method, with 5-fold cross-validation. In order to avoid the overfitting [Han et al. 2011] and thus guarantee greater generalizability to the models, we optimized the hyperparameters of maximum depth and minimum samples per leaf. In addition, considering the natural imbalance of the data[7], the hyperparameter that assigns different weights to classes during training was also considered, searching to boost the minority class learning. Table 3 shows the values

[7] As presented in Section 4.1, less than ten percent of the records correspond to hospitalized patients.

evaluated in the hyperparameters optimization, with the default values of the DT algorithm highlighted. Table 4 presents the hyperparameter combinations resulting from the optimization process and considered in the construction of each evaluated model. Fields marked with "D" in the table represent the default values of hyperparameters, already informed in Table 3.

#	Hyperparameter	Considered Values
hp 1	max_depth	[**None**, 3, 5, 7, 10, 15, 20, 25]
hp 2	min_samples_leaf	[**1**, 0.001, 0.003, 0.005, 0.01]
hp 3	class_weight (yes, no)	[**(1, 1)**, (2, 1), (3, 1), (5, 1)]

Table 3. List of hyperparameters evaluated in the optimization process.

Training	Without Attribute Selection			With Attribute Selection		
Dataset	hp 1	hp 2	hp 3	hp 1	hp 2	hp 3
A	15	0.001	D	D	0.001	(2, 1)
AB	15	D	D	D	0.001	D
ABC	15	D	D	15	D	(2, 1)
ABCD	15	D	D	15	D	(2, 1)
ABCDE	10	D	(2, 1)	7	D	(2, 1)
BC	15	D	D	10	D	(3, 1)
CD	15	D	D	10	D	D
DE	10	D	(2, 1)	10	D	(2, 1)

Table 4. Hyperparameter setting of the models trained on Database II, without or with attribute selection.

5. Results and Discussions

This section will present and discuss the results of this work. In Section 5.1, we initially describe the results of the experiments performed to choose the GA selection and crossing operators carried out on Database I. Afterwards, we present the attributes selected by our genetic approach, considering Databases I and II, separately. Finally, in Section 5.2, we present the impact of using the selected attributes in predicting the hospitalization need of COVID-19 patients, considering both the initial evaluation presented in the SBSI article and performed on Database I (Subsection 5.2.1), and the extended evaluation performed on the updated data from Database II concerning the temporal aspect (Subsection 5.2.2).

5.1. Attribute Selection

As mentioned in Section 4.2, we implemented and evaluated different genetic operators of selection and crossing in order to choose the best configuration of our GA and, consequently, our attribute selection process. Table 5 displays the results of this validation, performed on Database I. The approach implemented in this work is referenced by GA+DT, accompanied, in parentheses, by the considered selection and crossing operators. To position the results before other attribute selection techniques, we evaluated the classification results after applying the Recursive Feature Elimination with Cross-Validation (RFECV)

and SelectKBest methods from Scikit-Learn [Scikit-learn 2020]. In RFECV, a standard decision tree was used to define the importance of attributes and recursively remove the less important attributes, considering cross-validation of 5 folds. In SelectKBest, which performs based on univariate statistical tests, we considered selecting the best *k* attributes, and a *k*=27 test, which corresponds to the average of attributes returned by the other selectors.

Table 5. Comparison between attribute selectors on Database I.

Attribute Selector	F1-Score	n_f	Time(s)
1. GA+DT (roulette, two-points)	0.895	26	276
2. GA+DT (roulette, one-point)	0.898	21	279
3. GA+DT (roulette, uniform)	0.903	28	265
4. GA+DT (tournament, two-points)	0.897	27	260
5. GA+DT (tournament, one-point)	0.903	38	281
6. **GA+DT (tournament, uniform)**	**0.908**	28	257
7. GA+DT (truncated, two-points)	0.900	42	302
8. GA+DT (truncated, one-point)	0.899	28	250
9. GA+DT (truncated, uniform)	0.889	23	243
10. RFECV	0.892	9	240
11. SelectKBest	0.885	27	3
12. None	0.880	51	2

We can notice that the genetic selector with tournament selection and uniform crossing (line 6) obtained the best result, eliminating 23 attributes and increasing the f1-score by 2.8%, compared to the ranking without any selection strategy (line 12). Besides, it is possible to notice that we achieved the other two best results by combinations that included either the tournament's selection or the uniform crossover (lines 3 and 5), which motivated the choice of these operators to make up our genetic approach, definitively. Considering the most common methods, selection by statistical test (line 11) was naturally faster, almost instantaneous, but showed little contribution in f1-score. On the other hand, the RFECV (line 10) provided the most significant reduction in the number of attributes and improved the classification's performance by 1.2%, but was at a disadvantage of 1.6% compared to the genetic selector (line 6). Although faster, the RFECV took only 17s less regarding the execution time, which confirms the feasibility of the hybrid approach proposed in this work.

Figure 4 shows the convergence graph of the genetic selector with the chosen operators (related to line 6 of Table 5). We can observe that the GA found a subset of attributes that exceeded all other selection methods' performance already in the fourth generation, considering the f1-score. Besides, the GA's performance is continually increasing with elitism, and, including one individual with complete attributes set in the initial population, we have only solutions that improve the classifier's initial performance (that is, without any selection of attributes).

Next, Table 6 shows the attributes that presented the three largest selection recurrences, considering the 11 methods evaluated in Table 5. We can notice that

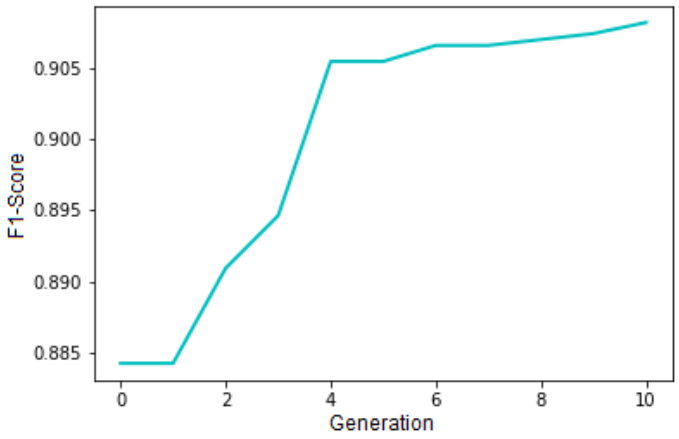

Figure 4. The convergence of the genetic selector, considering the chosen operators (tournament, uniform) and the f1-score (vertical axis), across the generations (horizontal axis).

all methods selected the attributes related to the indications of the following conditions/comorbidities[8]: diabetes mellitus, chronic cardiovascular disease, chronic neurological disease, other chronic pneumopathy, and other diseases/comorbidities. Also, the attributes related to dyspnea symptoms (shortness of breath) and the conditions/comorbidities of asthma and chronic liver disease were no longer selected by only one selection method, while only two selectors did not consider the indication of the age group from 60 to 69 years. This fact means that the vast majority of selectors considered these attributes important for predicting hospitalizations of COVID-19.

Table 6. More frequent attributes in the selections related to Table 5.

Frequency	Attributes
11	27, 28, 31, 42, 43
10	4, 25, 30
9	15

Considering precisely the GA adopted in this work, using selection by tournament and uniform crossing, in addition to all the attributes presented in Table 6, the algorithm selected the following variables for Database I: [0, 1, 6, 9, 10, 11, 17, 18, 21, 29, 32, 35, 36, 38, 41, 44, 45, 46, 47][9]. Among these attributes, we highlight the inclusion of indicators of fever symptom; age groups ranging from 10 to 29 years and elderly over 80 years; and conditions related to obesity, chronic decompensated respiratory disease, pregnant women, and postpartum women.

Also, we trained and ran the genetic selector for all divisions from Database II. We presented the attribute subsets resulting from these selections in Table 7, and we high-

[8]We previously presented the mapping between the number and description of each attribute in the Table 2.

[9]Notice that these attributes, added to those from Table 6, correspond to the total of 28 attributes indicated in line 6 of Table 5.

lighted the attributes that demonstrated the three highest frequencies in these selections in Table 8. Comparing with the attributes selected from Database I and which also presented a higher frequency of selection in the experiments carried out in such database (Table 6), we can observe that the attribute related to the age group 60 to 69 not appears among the most frequently selected for the temporal partitions from Database II (Table 8). On the other hand, attributes related to brown and black races/colors, to symptoms referred to as "Other", and to conditions/comorbidities of diabetes, chronic heart disease, and down syndrome, which was not among the attributes selected from Database I, showed significant frequency among the selections carried out on Database II.

Table 7. Lists of attributes selected for each training division from Database II.

Dataset	Selected Attributes	Attribute Count
A	[0, 1, 4, 5, 6, 7, 9, 15, 16, 18, 21, 22, 23, 26, 27, 28, 30, 31, 32, 35, 36, 38, 40, 41, 42, 43, 44, 45, 46, 48]	30
AB	[1, 4, 5, 8, 11, 12, 13, 14, 15, 16, 17, 18, 19, 21, 22, 23, 24, 25, 26, 27, 28, 30, 31, 32, 33, 34, 35, 36, 37, 39, 41, 42, 43, 44, 45, 46, 47, 48, 49]	39
ABC	[0, 1, 4, 5, 6, 7, 8, 12, 13, 17, 18, 19, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 37, 40, 41, 42, 43, 46, 47, 48]	31
ABCD	[4, 5, 6, 8, 11, 12, 13, 14, 17, 18, 20, 22, 23, 24, 26, 27, 28, 30, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49]	35
ABCDE	[0, 1, 2, 4, 5, 6, 12, 15, 16, 17, 18, 21, 23, 24, 25, 26, 27, 28, 30, 31, 33, 34, 35, 36, 37, 39, 40, 41, 42, 43, 44, 45, 46, 48]	34
BC	[0, 1, 4, 7, 9, 12, 16, 18, 19, 21, 24, 25, 27, 28, 30, 31, 35, 36, 37, 38, 39, 41, 42, 43, 44, 45, 48]	27
CD	[0, 1, 4, 5, 8, 9, 13, 14, 15, 16, 18, 19, 21, 23, 25, 26, 27, 28, 30, 31, 33, 35, 36, 40, 41, 42, 43, 44, 45, 46, 48]	31
DE	[0, 1, 3, 4, 5, 9, 10, 11, 13, 14, 17, 18, 20, 21, 22, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49]	38

Table 8. More frequent attributes in the selections presented in Table 7.

Frequency	Attributes
8	4, 18, 27, 28, 30, 41, 42, 43, 48
7	1, 5, 26, 31, 35, 36, 44, 45, 46
6	21, 23, 24, 25, 33, 40

5.2. Prediction of COVID-19 hospitalizations

As described in Section 4.3, different predictive models were trained and tested in two different evaluation processes. At first, standard DT models were built and evaluated considering Database I. In this case, we performed the evaluation considering the methods of train/test split (in the proportions of 70-30%) and 10-fold cross-validation. After, to better simulate the reality of the future predictions and assess the stability of predictive performance over time, the evaluation was extended to consider temporal divisions for training

and testing data. In addition, to consider Database II, which is more updated and covers a longer period, in this new evaluation, the predictive models had their hyperparameters optimized for a fairer comparison between the results with or without attribute selection, as described in Section 4.3. The results of these two evaluations are presented below, in Sections 5.2.1 and 5.2.2, respectively.

5.2.1. Initial Evaluation

Table 9 shows the results of the train/test split evaluation on Database I, where the left values correspond to the model that does not consider the attribute selection, while the values on the right are resulting from the model trained from the attributes selected by the GA for Database I. In addition to accuracy, we considered the precision, recall, and f1-score metrics for each predicted class (YES or NO for the need for hospitalization) to analyze better the impact caused by the attribute selection. We can notice that the YES class's precision improved by 18.5% with the subset of selected attributes, from 77.63% to 96.13%; that is, of the patients classified as needing hospitalization, 96.13% were hospitalized.

Table 9. Results of the predictive models on test dataset.

	Without Attribute Selection			With Attribute Selection		
	Class Yes	Class No	Average	Class Yes	Class No	Average
Precision (%)	77.63	**98.06**	87.84	**96.13**	97.56	**96.84**
Recall (%)	**79.36**	97.85	**88.60**	73.47	**99.72**	86.59
F1-score (%)	78.49	97.95	88.22	**83.29**	**98.63**	**90.96**
Accuracy (%)	-	-	96.26	-	-	**97.47**

Considering the recall, although it provided a 1.87% improvement in the negative class, the positive class suffered a reduction of 5.89%, meaning that, with the selection of attributes, we identified 5.89% less of the hospitalized patients. However, it is essential to highlight that the YES class is the underrepresented/minority class, corresponding to less than 10% of patients, as mentioned in Section 4.1. Thus, even a slight reduction in hospitalized patients' recovery significantly impacts the recall of the positive class. On the other hand, the 1.87% increase in negative class recall represents that 99.72% of the outpatients began to be correctly identified. This feature means that a much larger number of patients were no longer unduly predicted as requiring hospitalization, considering the proportions of the classes. This reduction is of great importance since, although the model's objective is to detect the patients who will need hospitalization as much as possible, many patients should not be unduly predicted as needing hospitalization (false positives) since this would disperse the focus of patients who need attention.

The results in Table 9 were already included in the SBSI publication. However, the consistency of these results was also verified by 10-fold cross-validation on Database I, as shown in Table 10. It is possible to notice that the values remained similar to those in the Table 9 for all metrics.

Table 10. Results of the predictive models on 10-fold cross-validation.

	Without Attribute Selection			With Attribute Selection		
	Class Yes	Class No	Average	Class Yes	Class No	Average
Precision (%)	77.39	**98.06**	87.72	**96.31**	97.59	**96.95**
Recall (%)	**79.42**	97.80	**88.61**	73.74	**99.72**	86.73
F1-score (%)	78.34	97.93	88.13	**83.38**	**98.64**	**91.01**
Accuracy (%)	-	-	96.23	-	-	**97.49**

In order to analyze how we can interpret some attributes in the prediction, we generated a simplified version of the predictive model that considers the selected attributes, restricting the decision tree's depth to five. Figure 5 presents the visualization of this model, which was generated with the PyDotPlus[10] and GraphViz[11] packages. Although a large number of rules have been disregarded with this restriction (only 9 of the 28 selected attributes appear in the visualization, for example), we can notice that, in general, COVID-19 patients who exhibited the following patterns were considered as needing hospitalization:

P1: With chronic cardiovascular disease or with some condition or comorbidity indicated as "Other", unless they also have a chronic pneumopathy and, even so, do not feel short of breath (dyspnea);
P2: Without chronic cardiovascular disease and conditions indicated as "Other", but with diabetes mellitus;
P3: Without any of the conditions/comorbidities mentioned in the previous rules, but which are asthmatic and present dyspnea;
P4: No chronic cardiovascular disease, no conditions indicated as "Other", no diabetes mellitus, and no dyspnea symptom, but who have a chronic neurological disease.

Considering the decision tree presented in Figure 5, we can notice that of the 13,182[12] hospitalized patients used in the model training, about 60.43% (7966) fit the rules indicated in P1. This fact is because P1 includes both patients who have a chronic cardiovascular disease, represented by the right branch of the root; and those with some condition or comorbidity indicated as "Other", represented by the branch to the left of the root and to the right of the subsequent node "Other comorbidities <= 0.5". Adding the number of true positives (hospitalized patients who were correctly classified) in the leaf nodes of these branches, we obtain 7966 (4956 and 3010[13] for each branch, respectively). In the same way, P2 is represented by the branch to the left of the root, to the left of the node "Other comorbidities <= 0.5", and to the right of the node "Diabetes mellitus <=

[10] https://pydotplus.readthedocs.io

[11] https://graphviz.readthedocs.io

[12] This quantitative can be seen in the "value" field of the root node, which shows the real distribution of the total number of samples among the hospitalization classes ([No, Yes]).

[13] Notice that, in the description of pattern P1, we disregarded the division of the node related to the test "Age range 20 to 29 <=0.5", which, on the right, leads to the light red leaf. This fact is because this division involves few examples and low precision, not significantly interfering in the pattern of the superior branch.

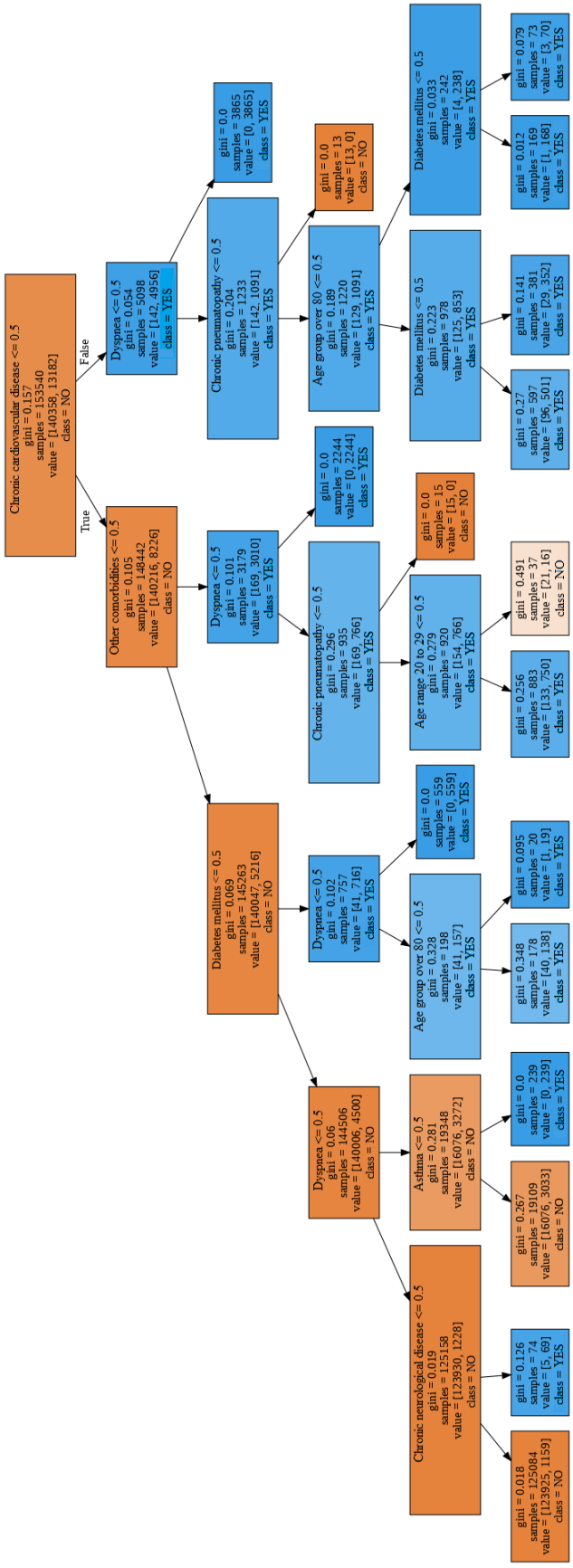

Figure 5. Image of the simplified visualization of the decision tree generated from the training set of Database I. For tree interpretation, it is essential to notice that each internal node contains a test on an attribute and that the branches to the left and right of this node represent, respectively, the confirmation and denial of its test. Considering a node related to the test "Symptom X $<= 0.5$", for example, if a patient does not have symptom X, your record have the value 0 for this binary/boolean attribute and, as 0 is less than or equal to 0.5, the patient is represented in the branch to the left of the evaluated node. Otherwise, the patient would be represented in the branch on the right. Please also notice that a node's "samples" field indicates how many training samples (patients) were tested on it. How many samples resulted from testing the node immediately above (parent). In addition, as no class weights were established, each node's "value" field indicates the real distribution of the tested samples concerning the [NO, YES] hospitalization classes, and the "class" field indicates the majority class among these samples. Finally, the tree's leaves (nodes that have no other descendent nodes) represent the classification results and, thus, indicate the class that the model assigned to the examples represented there. Thus, if the "class" field of a leaf is "YES", the first and second number of its "value" field correspond, respectively, to the True Positives and the False Positives of that classification. Otherwise, if the "class" field of a leaf is "NO", the first and second number of its "value" field correspond, respectively, to the True Negatives and False Negatives of that classification.

0.5", covering 5.43% (716) of the hospitalized patients. Unlike P2, P3 is represented by the leaf node to the left of the node "Diabetes mellitus <= 0.5", and then to the right of the node "Dyspnea <= 0.5" and to the right of the node "Asthma <= 0.5", satisfying 1.81% (239) of hospitalized patients. In turn, unlike P3, P4 corresponds to the leaf node to the left of the node "Dyspnea <= 0.5", and then to the right of the node "Chronic neurological disease <= 0.5", covering 0.52% (69) of the patients hospitalized. Together, these rules correctly described the hospitalization pattern of 68.19% of the patients used as training examples for the positive class (with hospitalization need).

Finally, we can also notice that, even in this simplified version, the model corroborates previous studies when relating the pre-existence of chronic diseases, mainly cardiovascular diseases and diabetes, to the most severe manifestation of COVID-19 [The Novel Coronavirus Pneumonia Emergency Response Epidemiology Team 2020] and, consequently, to the need for hospitalization.

5.2.2. Extended/Temporal Evaluation

Regarding the extended evaluation, we developed and evaluated predictive models for each temporal combination of training and testing data from Database II, considering the entire set of attributes (without selection) and the subsets of attributes selected by our genetic approach for each training set from Database II, as described in Section 4.3. Table 11 presents the results related to the models that considered attribute selection. Notice that we used the symbols ▲ and ▼ to signal, respectively, the temporal experiments with best and worst results for each evaluated metric. See that the model trained and evaluated only on 2020 data, which correspond to experiment 6, had the highest accuracy (more significant than 98%) and similar results to those of the previous evaluation (from Section 5.2.1), having been able to recover 74.78% of hospitalized patients (positive class recall), despite the lower precision of 87.19%. We already expected similarity between these results because, although we did not consider the temporality in separating the training and test data from Database I, the previous evaluation covered records until part of October 2020, which corresponds approximately to the union of sets A and B from Database II. That is, the data used in experiment 6 are close, in time, to those considered in the evaluation on Database I, presented in Section 5.2.1.

Comparing with the results that this same model (trained by the composition of sets A and B) presented when tested on the totality of remaining sets (line 2), it is possible to observe a reduction in the positive class recall. Notice that, although the positive class precision has increased due to a slight gain in the negative class recall, the loss in the positive class recall was more significant, resulting in worse f1-scores and accuracy. These results indicate that this model's predictive performance was not maintained in the long term, especially concerning the ability to correctly retrieve/identify patients with a future need for hospitalization (recall of the YES class). Furthermore, the concentration of worst results in the model trained on data related to the onset of the pandemic and evaluated on all other data (line 1) also shows that, in the dynamic context of the pandemic, the predictive power of a model is not sustained in the long term.

	Precision (%)		Recall (%)		F1-score (%)		Accuracy (%)
	Class Yes	Class No	Class Yes	Class No	Class Yes	Class No	
1. A-BCDEF	▼ 71.80	97.94	74.36	▼ 97.66	▼ 73.06	▼ 97.80	▼ 95.93
2. AB-CDEF	91.81	97.32	66.25	99.52	76.96	98.41	97.02
3. ABC-DEF	95.23	96.83	64.23	99.71	76.71	98.24	96.74
4. ABCD-EF	90.83	96.80	66.97	99.33	77.09	98.05	96.40
5. ABCDE-F	91.23	96.41	▼ 61.11	99.44	73.19	97.90	96.11
6. AB-C	87.19	▲ 98.73	74.78	99.44	80.51	▲ 99.09	▲ 98.26
7. BC-D	94.89	97.54	66.99	99.72	78.53	98.62	97.40
8. CD-E	▲ 97.27	▼ 96.40	63.84	▲ 99.82	77.09	98.08	96.46
9. DE-F	84.73	97.91	▲ 77.83	98.67	▲ 81.13	98.28	96.85
Average 1-5	88.18	97.06	66.58	99.13	75.40	98.08	96.44
Average 6-9	91.02	97.65	70.86	99.41	79.32	98.52	97.24
Overall Average	89.44	97.32	68.48	99.26	77.14	98.27	96.80

Table 11. Results of the experiments considering the attributes selected for each training set from Database II.

	Precision (%)		Recall (%)		F1-score (%)		Accuracy (%)
	Class Yes	Class No	Class Yes	Class No	Class Yes	Class No	
1. A-BCDEF	- 16.19	+ 0.51	+ 6.98	- 1.60	- 3.26	- 0.54	- 0.97
2. AB-CDEF	+ 8.18	+ 0.04	+ 0.22	+ 0.57	+ 3.16	+ 0.30	+ 0.55
3. ABC-DEF	- 2.04	+ 0.39	+ 4.57	- 0.14	+ 2.75	+ 0.13	+ 0.26
4. ABCD-EF	+ 0.81	+ 0.08	+ 0.77	+ 0.06	+ 0.80	+ 0.07	+ 0.12
5. ABCDE-F	+ 10.02	- 1.57	- 17.61	+ 1.17	- 6.75	- 0.22	- 0.46
6. AB-C	+ 10.17	+ 0.12	+ 2.35	+ 0.53	+ 5.86	+ 0.33	+ 0.63
7. BC-D	+ 1.71	- 0.04	- 0.61	+ 0.10	+ 0.18	+ 0.03	+ 0.05
8. CD-E	+ 7.43	- 0.52	- 5.53	+ 0.63	- 1.20	+ 0.04	+ 0.05
9. DE-F	+ 4.71	- 0.26	- 2.95	+ 0.59	+ 0.73	+ 0.16	+ 0.27
Average 1-5	+ 0.16	- 0.11	- 1.01	+ 0.01	- 0.66	- 0.05	- 0.10
Average 6-9	+ 6.01	- 0.18	- 1.69	+ 0.46	+ 1.39	+ 0.14	+ 0.25
Overall Average	+ 2.76	- 0.14	- 1.31	+ 0.21	+ 0.25	+ 0.03	+ 0.06

Table 12. Differences between the results that considered attributes selected for each training set from Database II (Table 11) and those obtained without attribute selection.

This result could suggest that, as the evolution of the pandemic can exert a strong influence on hospitalization patterns, the models would need to be constantly updated, considering more and more training samples to maintain a good/stable predictive performance. However, the results of the first five experiments do not confirm this hypothesis, as they do not demonstrate a pattern/trend of increase in positive class recall as new datasets are added to the training set. On the contrary, the model that concentrated the most significant training data (line 5) had the worst recall in the positive class and unattractive results. The models that had the training data updated from a sliding window (lines 6 to 9) demonstrated, in turn, more stable predictive results, concentrating the best results of all evaluated metrics. This fact demonstrates that the accumulation of training data impairs the model's predictive capacity, being more appropriate to approximate the training data to the prediction context temporally. Furthermore, this fact suggests that the predictive power of the attributes and prediction patterns varied throughout the pandemic so that rerunning the attribute selection and periodically updating the model, considering data that represent a closer context, allows predictors to follow these variations and thus provide adequate and stable results over time.

Table 12 presents the differences between the results obtained in the previous experiments, which considered only the attributes selected for each training partition from Database II and those produced on all attributes, that is, without attribute selection. Notice again that, in general, the attribute selection benefits the negative class recall and, consequently, the positive class precision, the f1-scores, and the overall hit rate (accuracy). Considering the models trained and evaluated from data updated by the sliding window (lines 6 to 9), we can observe that, although the positive class recall may suffer significant reductions with the adoption of attribute selection (-1.69% on the average of the experiments), this reduction is smaller than the gains achieved in precision (+6.01% on the average of the experiments). Given that precision and recall frequently have an inverse relationship, adopting or not the attribute selection naturally depends on the objective of the application. For example, considering the domain of COVID-19, in which the number of infections is significant, as opposed to the size of the health teams, if we used the model to prioritize contacts for home follow-up, precision would have priority. That is, as professionals would hardly be able to contact a large number of patients daily, it would be desirable to optimize as much as possible the service targeting, prioritizing only patients with a genuine and high possibility of needing hospitalization. In this case, our attribute selection approach would provide significant benefits.

Finally, in order to verify possible changes in hospitalization patterns, a simplified version (with the tree depth limited to five) of the most updated model was generated and presented in Figure 6. Just as the visualization presented in Section 5.2.1 refers to a simplified model trained on attributes selected from Database I and on 2020 data, which approximately match datasets A and B; this new graph corresponds to the simplified version of the model related to the DE-F experiment, which was trained on attributes selected for Database II and on 2021 data. Analyzing Figure 6, we can see that the rules that compose the branch to the right of the root and the left of the node "No condition/comorbidity $<=$ 0.5" correctly classify 55.76% (25910)[14] of the hospitalization examples used in train-

[14]Please observe the notes described in the caption of Figure 6.

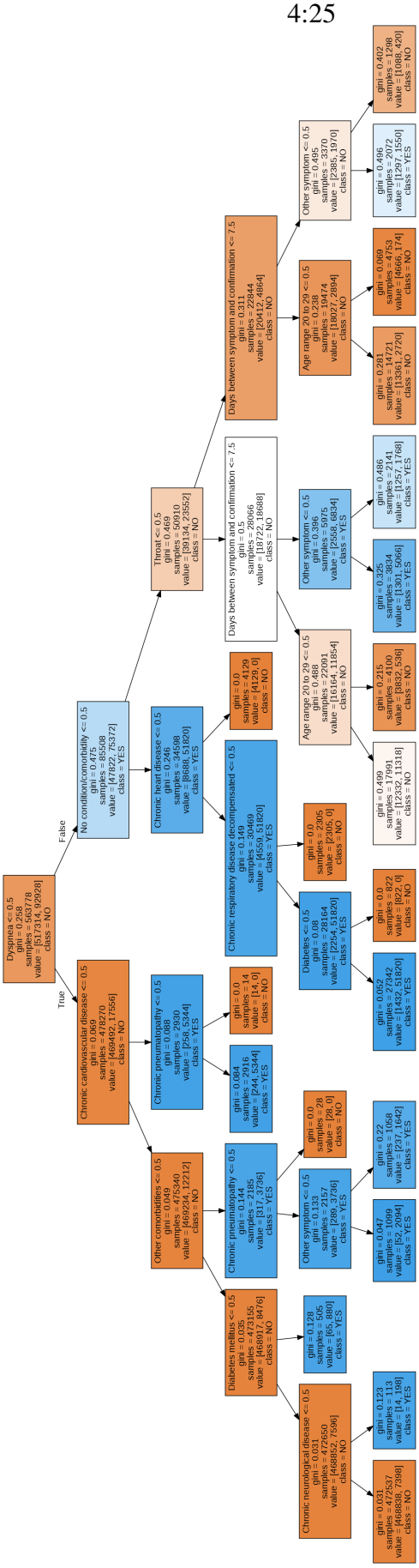

Figure 6. Image of the simplified visualization of the model related to DE-F experiment, which was trained with attributes selected for DE training set from Database II. Note that the weights (2, 1) were considered as hyperparameters for the classes (YES, NO) in the DE-F experiment (as presented in Section 4.3). Thus, the second number of each node's "value" field corresponds to twice the actual number of hospitalization samples. For example, although the sum of true positives in the branch leaves to the right of the root and the left of the node "No condition/comorbidity $<= 0.5$" is 51820, this value corresponds to 25910 hospitalized patients.

ing (46464). In addition, these rules determine needing hospitalization for patients with dyspnea and some condition/comorbidity other than chronic heart disease, chronic decompensated respiratory disease, or diabetes. In turn, the patterns related to the branch to the left of the root correctly describe 10.96% (5079) of the hospitalizations (true positives) and indicate that, in general, patients without dyspnea only need hospitalization if they have chronic cardiovascular disease, any comorbidity referred to as "Other", diabetes mellitus, or a chronic neurological disease. Among these patients, only those with a chronic pneumatopathy, combined with a chronic cardiovascular disease or comorbidity referred to as "Other", do not need hospitalization.

Notice that, despite having a different structure, the previously mentioned rules satisfy the main pattern (P1) of the 2020 model (Figure 5), as they also classify as needing hospitalization the patients with chronic cardiovascular disease or with some condition or comorbidity indicated as "Other", unless they also have a chronic pneumopathy and, even so, do not feel short of breath (dyspnea). However, in the 2021 model, we can observe new patterns, such as those patients with dyspnea but who do not have any condition/comorbidity and neither symptoms related to throat, need hospitalization if more than seven days have passed between the onset of symptoms and the confirmation of COVID-19. Notice that this pattern describes 7.35% (3417) of the positive samples of hospitalization (true positives) but wrongly classifies as hospitalized (false positives) 0.49% of the negative samples. Although this percentage seems small because it evolves the majority class, it corresponds to 2558 patients, which generates a considerable loss in the positive class precision. These more linked and less accurate (leaves with weaker/lighter colors) rules suggest that with the evolution of the pandemic and, fortunately, the vaccination, the hospitalization patterns became more complex. However, the predictive results presented in line 9 of Table 11 are still satisfactory.

6. Conclusions

This work developed a hybrid approach to attribute selection, composed of a genetic algorithm and a classifier, to improve the prediction results of hospitalizations of COVID-19 patients. We used a genetic algorithm to search for a subset of optimal attributes, considering the classification results provided by such attributes in their evaluation function. In addition to considering previous works and contextual particularities, we also based some methodological choices used in this process on experimental validations, such as adopting tournament selection and uniform crossing in the genetic algorithm.

As a result, we identified that the proposed selection method surpassed traditional strategies, such as recursive elimination of attributes and selection based on the univariate statistical test. Additionally, when using the subset of attributes selected by the genetic algorithm in an initial evaluation, considering only data available until October 2020, we perceived a very positive impact on the results of the predictive model, which obtained an average increase of 18% in the classification precision of COVID-19 patients with hospitalization need (positive class). Afterward, by extending the evaluation with data until June 2021 and considering temporal divisions for training and test data, it was possible to verify that the accumulation of training data impairs the model's predictive capacity. However, considering the temporal evaluation with a sliding window, attribute selection

continued to provide a significant gain in the positive class precision, which was 6%, on average. This fact suggests that the predictive power of the attributes and prediction patterns varied throughout the pandemic, so that rerunning the attribute selection and periodically updating the model, considering data temporally closer to the prediction context, allows predictors to follow these variations, thus providing adequate and stable results over time.

Given that precision and recall frequently have an inverse relationship, adopting or not, the selection of attributes naturally depends on the objective of the application. Importantly, we are not proposing that our approach be used to decide whether or not a patient should be admitted. Of course, this decision should always depend exclusively on the clinical situation and the medical evaluation of each patient. However, considering the context of the pandemic, in which the scarcity of resources in health is even more remarkable, an accurate and interpretable model for predicting the need for hospitalization of COVID-19 patients can create opportunities, based on knowledge of patterns and predictions, for better management of services and resources that are not linked to severe neglect of patients. An example of an application would be prioritizing home follow-up contacts, where there is already a natural lack of assistance due to the impossibility of health teams to contact all patients undergoing home treatment daily. In this case, the model would provide better service targeting, prioritizing contact with patients with a high possibility of needing hospitalization. However, it would not bring an uncontacted patient a more significant loss than that already exists today. In other words, in case of a negative evolution of the disease, the patient would continue to go to the hospital to receive appropriate clinical evaluation and treatment.

Finally, as future studies, in addition to the integration of the predictive model into an information system to support decision-making, several improvements can be considered, such as the performance of more exhaustive experiments, including the evaluation of methodological decisions inspired by previous studies (population size and probabilities of mutation and crossing of the genetic algorithm, for example). In addition, we could evaluate data balancing techniques to improve the attribute selection process and predictive model results.

7. Acknowledgement

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001. It was also supported by the Programa Institucional de Incentivo à Qualificação Profissional em Programas Especiais (PIIQPPE) from IFFar. The authors thank CAPES, PPGC/UFPel and IFFar.

References

Alpaydin, E. (2010). *Introduction to machine learning*. MIT Press, Cambridge, 2nd edition.

Arvind, V., Kim, J. S., Cho, B. H., Geng, E., and Cho, S. K. (2021). Development of a machine learning algorithm to predict intubation among hospitalized patients with COVID-19. *Journal of Critical Care*, 62:25–30. doi: https://doi.org/10.1016/j.jcrc.2020.10.033.

Burdick, H., Lam, C., Mataraso, S., Siefkas, A., Braden, G., Dellinger, R. P., McCoy, A., Vincent, J.-L., Green-Saxena, A., Barnes, G., Hoffman, J., Calvert, J., Pellegrini, E., and Das, R. (2020). Prediction of respiratory decompensation in Covid-19 patients using machine learning: The READY trial. *Computers in Biology and Medicine*, 124:103949. doi: https://doi.org/10.1016/j.compbiomed.2020.103949.

Colpo, M. P., Alves, B. C., Pereira, K. S., Brandão, A. F. Z., de Aguiar, M. S., and Primo, T. T. (2021). Attribute selection based on genetic and classification algorithms in the prediction of hospitalization need of COVID-19 patients. In *XVII Brazilian Symposium on Information Systems*, SBSI 2021, New York, NY, USA. Association for Computing Machinery. doi: https://doi.org/10.1145/3466933.3466935.

Cueto-López, N., García-Ordás, M. T., Dávila-Batista, V., Moreno, V., Aragonés, N., and Alaiz-Rodríguez, R. (2019). A comparative study on feature selection for a risk prediction model for colorectal cancer. *Computer Methods and Programs in Biomedicine*, 177:219–229. doi: https://doi.org/10.1016/j.cmpb.2019.06.001.

Faria, N. R., Mellan, T. A., Whittaker, C., Claro, I. M., da S. Candido, D., Mishra, S., Crispim, M. A. E., Sales, F. C. S., Hawryluk, I., McCrone, J. T., Hulswit, R. J. G., Franco, L. A. M., Ramundo, M. S., de Jesus, J. G., Andrade, P. S., Coletti, T. M., Ferreira, G. M., Silva, C. A. M., Manuli, E. R., Pereira, R. H. M., Peixoto, P. S., Kraemer, M. U. G., Gaburo, N., da C. Camilo, C., Hoeltgebaum, H., Souza, W. M., Rocha, E. C., de Souza, L. M., de Pinho, M. C., Araujo, L. J. T., Malta, F. S. V., de Lima, A. B., do P. Silva, J., Zauli, D. A. G., de S. Ferreira, A. C., Schnekenberg, R. P., Laydon, D. J., Walker, P. G. T., Schlüter, H. M., dos Santos, A. L. P., Vidal, M. S., Caro, V. S. D., Filho, R. M. F., dos Santos, H. M., Aguiar, R. S., Proença-Modena, J. L., Nelson, B., Hay, J. A., Monod, M., Miscouridou, X., Coupland, H., Sonabend, R., Vollmer, M., Gandy, A., Prete, C. A., Nascimento, V. H., Suchard, M. A., Bowden, T. A., Pond, S. L. K., Wu, C.-H., Ratmann, O., Ferguson, N. M., Dye, C., Loman, N. J., Lemey, P., Rambaut, A., Fraiji, N. A., do P. S. S. Carvalho, M., Pybus, O. G., Flaxman, S., Bhatt, S., and Sabino, E. C. (2021). Genomics and epidemiology of the P.1 SARS-CoV-2 lineage in Manaus, Brazil. *Science*, 372(6544):815–821. doi: https://doi.org/10.1126/science.abh2644.

Funchal, J. P. d. S. and Adanatti, D. F. (2016). Um estudo sobre a classificação de risco na Área da saúde utilizando Árvores de decisão. *iSys – Revista Brasileira de Sistemas de Informação*, 9(3):9–111. doi: https://doi.org/10.5753/isys.2016.317.

Han, J., Pei, J., and Kamber, M. (2011). *Data mining: concepts and techniques*. Morgan Kaufmann, Waltham, 3rd edition.

Heckler, W. F., Varella, J. d. C., Costa, C. C. d., and Barbosa, J. L. V. (2020). A model to patient abandonment prediction in the pulmonary rehabilitation. In *XVI Brazilian Symposium on Information Systems*, SBSI'20, New York, NY, USA. Association for Computing Machinery. doi: https://doi.org/10.1145/3411564.3411642.

Linden, R. (2008). *Algoritmos Genéticos*. Brasport, Rio de Janeiro, 2nd edition.

Lynch, C. M., Abdollahi, B., Fuqua, J. D., de Carlo, A. R., Bartholomai, J. A., Balgemann, R. N., van Berkel, V. H., and Frieboes, H. B. (2017). Prediction of lung cancer patient survival via supervised machine learning classification techniques. *International Journal of Medical Informatics*, 108:1–8. doi: https://doi.org/10.1016/j.ijmedinf.2017.09.013.

Maleki, N., Zeinali, Y., and Niaki, S. T. A. (2021). A k-NN method for lung cancer prognosis with the use of a genetic algorithm for feature selection. *Expert Systems with Applications*, 164:113981. doi: https://doi.org/10.1016/j.eswa.2020.113981.

Monteiro, F., Meloni, F., Baranauskas, J. A., and Macedo, A. A. (2020). Prediction of mortality in intensive care units: a multivariate feature selection. *Journal of Biomedical Informatics*, 107:103456. doi: https://doi.org/10.1016/j.jbi.2020.103456.

PAHO (2020). Pan American Health Organization. Ficha Informativa COVID-19: A COVID-19 e o papel dos sistemas de informação e das tecnologias na atenção primária. https://iris.paho.org/handle/10665.2/52206, May, 23.

Pawlovsky, A. P. and Matsuhashi, H. (2017). The use of a novel genetic algorithm in component selection for a kNN method for breast cancer prognosis. In *2017 Global Medical Engineering Physics Exchanges/Pan American Health Care Exchanges (GMEPE/PAHCE)*, pages 1–5, Tuxtla Gutierrez, Mexico. IEEE. doi: https://doi.org/10.1109/GMEPE-PAHCE.2017.7972084.

Pradeep, K. and Naveen, N. (2018). Lung cancer survivability prediction based on performance using classification techniques of support vector machines, c4.5 and naive bayes algorithms for healthcare analytics. *Procedia Computer Science*, 132:412–420. doi: https://doi.org/10.1016/j.procs.2018.05.162.

Raschka, S. and Mirjalili, V. (2017). *Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow*. Packt Publishing, Birmingham, UK, 2nd edition.

Scikit-learn (2020). Feature selection. https://scikit-learn.org/stable/modules/feature_selection.html.

SES/RS (2020). Secretaria Estadual da Saúde do Rio Grande do Sul. Painel Coronavírus RS. https://ti.saude.rs.gov.br/covid19/sobre.

The Novel Coronavirus Pneumonia Emergency Response Epidemiology Team (2020). The Epidemiological Characteristics of an Outbreak of 2019 Novel Coronavirus Diseases (COVID-19) — China, 2020. *China CDC Weekly*, 2:113. doi: https://doi.org/10.46234/ccdcw2020.032.

World Health Organization (2020). COVID-19 Weekly Epidemiological Update - 27 December 2020. https://www.who.int/docs/default-source/coronaviruse/situation-reports/20201229_weekly_epi_update_con_20_cleared.pdf, December, 29.

Zhou, Y., Zhang, W., Kang, J., Zhang, X., and Wang, X. (2021). A problem-specific non-dominated sorting genetic algorithm for supervised feature selection. *Information Sciences*, 547:841–859. doi: https://doi.org/10.1016/j.ins.2020.08.083.