InterSciencePlace

International Scientific Journal – ISSN: 1679-9844
Volume 19, Article nº 19, January/December 2024
D.O.I: http://dx.doi.org/10.6020/1679-9844/v19a19
Accepted: 21/06/2024 Published: 29/10/2024

ARTIFICIAL INTELLIGENCE APPLIED TO HEALTH - DATA ANALYSIS FOR CHILDREN'S HEALTH.

INTELIGÊNCIA ARTIFICIAL APLICADA À SAÚDE - ANÁLISE DE DADOS PARA SAÚDE DE CRIANÇAS.

Jackson Henrique da Silva Bezerra
Doctoral Student PGDRA/UFRO. Teacher from the Federal Institute of Education, Science and Technology of Rondônia - Campus Ji-Paraná. E-mail: jackson.henrique@ifro.edu.br

Fabrício Moraes de Almeida
PhD in Physics (UFC), withpost-doctorate in Scientific Regional Development (DCR/CNPq) - Specialization in Software Engineering (FUNIP). Researcher of the Doctoral and Master Program in Regional Development and Environment (PGDRA/UFRO). E-mail: dr.fabriciomoraes001@gmail.com

Fabio Machado de Oliveira
Doctor in Cognition and Language (UENF). E-mail: fabiomac@gmail.com

ABSTRACT

Machine Learning (ML) is a subset of Artificial Intelligence plays an important role in healthcare, providing predictive models created from algorithms and large databases. These models can classify patients for diagnostic or prognostic purposes in various diseases. This research aimed to develop a predictive model for death due to Severe Acute Respiratory Syndrome (SARS) for children aged 0 to 3 years in the North region of Brazil, using data provided by the Brazilian Ministry of Health. An applied research was carried out using the CRISP-DM methodology that guided the entire process of selection, processing, transformation, application of ML algorithms and evaluation of the model. The Random Forest, Logistic Regression, K-Nearest Neighbors and XGBoost algorithms were used through the Weka software, where the model with Random Forest had superior performance. The model was generated with cross-validation and evaluated according to the metrics of sensitivity, specificity, accuracy, precision, F1-Score and AUC-ROC, the latter being the primary evaluation metric. Finally, a software application prototype for using the model was developed in the Java language so that the knowledge generated by the model reaches healthcare professionals.

Keywords: Artificial Intelligence. Machine Learning (ML), Database, Severe Acute Respiratory Syndrome (SARS), Predictive Models.

RESUMO

O *Machine Learning* (ML) é um subconjunto da Inteligência Artificial, tem um papel importante na área da saúde, fornecendo modelos preditivos criados a partir de algoritmos e grandes bases de dados. Estes modelos podem classificar pacientes para fins de diagnóstico ou prognósticos em diversas doenças. A presente pesquisa teve como objetivo o desenvolvimento de um modelo preditivo de óbito por Síndrome Respiratória Aguda Grave (SRAG) para crianças de 0 a 3 anos da região Norte do Brasil, através de dados disponibilizados pelo Ministério da Saúde do Brasil. Uma pesquisa aplicada foi realizada através da metodologia CRISP-DM que guiou todo o processo de seleção, processamento, transformação, aplicação dos algoritmos de ML e avaliação do modelo. Os algoritmos *Random Forest*, *Regression Logistic,* K-*Nearest Neighbors* e XGBoost foram utilizados através do software Weka, onde o modelo com o *Random Forest* teve desempenho superior. O modelo foi gerado com validação cruzada e avaliado conforme as métricas de sensibilidade, especificidade, acurácia, precisão, F1-Score e AUC-ROC, sendo esta última a métrica primária de avaliação. Por fim, um protótipo de aplicação de software para uso do modelo foi desenvolvido na linguagem Java para que o conhecimento gerado pelo modelo chegue aos profissionais da área da saúde.

Palavra-chave: *Inteligência Artificial. Machine Learning* (ML), Banco de dados, Síndrome Respiratória Aguda Grave (SRAG), Modelo Preditivo.

INTRODUCTION

Machine Learning (ML) is a set of rules used to teach computers to automatically “learn” patterns and behaviors from training data (SILVA E, 2022), (SENA, 2021). The main objective of an ML model is to build a computer system that learns from a predefined database and ultimately generates a model for prediction, classification, or detection (PAIXÃO et al., 2022). In practice, ML applications are mainly focused on the use of consolidated databases containing heterogeneous information, for which traditional statistical techniques have limited applicability (PAIXÃO et al., 2022). ML algorithms have already been widely adopted across various fields, including banking systems for fraud detection (LOPES, 2019), internet search engines (CARVALHO, 2012), video surveillance systems (MOITINHO & BENICASA, 2023), data security (HENKE et al., 2018), robotics (RYBCZAK et al., 2024), and in medicine for diagnosis and prognosis (GROSSARTH et al, 2023). In the medical field, with the digitization of medical records, laboratory tests, and imaging exams, there has been a significant growth in databases, which are prime sources for applying ML techniques aimed at disease prevention, early diagnosis, and treatment (PAIXÃO et al., 2022).

ML algorithms can be broadly divided into two categories: supervised and unsupervised learning. In unsupervised learning, the ML model extracts data features and builds a representation without prior knowledge of the data labels; in other words, it heuristically identifies class patterns. This lack of supervision can be advantageous as it allows the algorithm to analyze patterns that were previously not considered (SENA, 2021), (PAIXÃO et al., 2021). In supervised learning, the ML model has knowledge of the data labels, meaning the samples are correctly defined. Training is based on comparing the results predicted by the model with the actual values. This process is repeated until a minimum error is achieved (PAIXÃO et al., 2021). Therefore, if the result of a supervised ML model prediction is a category, the task is called classification, such as predicting a student's grade in a subject within the categories A, B, C, D, and E. However, if the prediction is a specific numerical value, then the task is called regression, such as predicting a student's grade in a subject. ML algorithms can learn through parameter changes (such as linear weights) or learning structures (such as trees) (SILVA E, 2022).

ML algorithms can be broadly divided into two categories: supervised and unsupervised learning. In unsupervised learning, the ML model extracts data features and builds a representation without prior knowledge of the data labels; in other words, it heuristically identifies class patterns. This lack of supervision can be advantageous as it allows the algorithm to analyze patterns that were previously not considered

(SENA, 2021), (PAIXÃO et al., 2021). In supervised learning, the ML model has knowledge of the data labels, meaning the samples are correctly defined. Training is based on comparing the results predicted by the model with the actual values. This process is repeated until a minimum error is achieved (PAIXÃO et al., 2021). Therefore, if the result of a supervised ML model prediction is a category, the task is called classification, such as predicting a student's grade in a subject within the categories A, B, C, D, and E. However, if the prediction is a specific numerical value, then the task is called regression, such as predicting a student's grade in a subject. ML algorithms can learn through parameter changes (such as linear weights) or learning structures (such as trees) (SILVA E, 2022).

In recent years, ML has stood out as an important technological solution in healthcare, enabling the analysis of large datasets to extract knowledge in record time, promoting advancements in improving diagnoses and predicting clinical events, such as cases of Severe Acute Respiratory Syndrome (SARS) (BEZERRA & ALMEIDA, 2024). SARS is a serious medical condition that involves the rapid deterioration of respiratory symptoms, often leading to severe complications and even death. This syndrome can be triggered by various causes, including viral infections such as Influenza A (H1N1) and SARS-CoV-2 (COVID-19), among others, as well as bacterial infections (LEE et al., 2024).

In this regard, predictive models developed using ML can identify patients at greater risk of mortality from SARS, providing support for interventions aimed at reducing deaths (MOULAEI et al., 2022). The knowledge generated through ML can assist in the prognosis of SARS, helping healthcare professionals to better allocate material and human resources in the treatment of patients with a higher chance of death. ML helps predict the severity and progression of diseases like SARS by analyzing large datasets of electronic health records, clinical assessments, and images. These models support decision-making at various stages, from triage to hospital discharge, ensuring that resources such as ICU beds, ventilators, and medical staff are used efficiently to prioritize the most needy patients and improve overall patient outcomes (DEBNATH et al., 2020), (VAN DER SCHAAR et al., 2021). Besides generating models, creating mechanisms to make them available to healthcare professionals is important, as seen in the studies of Aznar-gimeno et al. (2021), Woo et al. (2021), Hu et al. (2021), and Kar et al. (2021).

In this context, the objective of the present work is to demonstrate the practical application of ML in the health area, through the generation of predictive models of death and cure for patients with SARS registered in the 2020 and 2021 SARS databases of the Ministry of Health available on the openDataSUS portal. Maintained by the Health Surveillance Secretariat (SVS), these databases stand out as an important repository of data on patients hospitalized for SARS. The available records are captured by the Influenza Epidemiological Surveillance Information System (SIVEP-Gripe), which tracks SARS cases and deaths in Brazil, caused by viruses such as SARS-CoV-2, Influenza A (H1N1), among others (BRASIL, 2024). Finally, it is noteworthy that the SARS databases on openDataSUS are published under Creative Commons Attribution (cc-by) and Open Data formats, allowing others to share, remix, adapt, and create derivative works (BRASIL, 2024). Another important factor is that all available records are anonymized according to the guidelines of Brazil's General Data Protection Law (LGPD), ensuring that no individual can be identified from the database.

METHODOLOGY

The CRISP-DM methodology is a widely recognized and used framework to guide Data Mining (DM) and ML projects. It consists of a cycle of six non-rigid phases, allowing for forward and backward movement between phases whenever necessary. The result of each phase determines which phase or activity of a particular phase should be performed next (CHAPMAN et al., 2000). The application of the CRISP-DM methodology was carried out according to the adaptation made by Sena (2021), and will be guided solely by the goals and activities of each phase.

The first phase, Business Understanding, focused on understanding the objectives and requirements of the project from a business perspective. At this stage, the risks and technical criteria for the project were also evaluated, along with the potential benefits, final goals, and success criteria. Tools for data analysis, manipulation, transformation, and model creation were also defined during this phase.

In the second phase, Data Understanding, the dataset was examined in depth, considering all relevant aspects. Data was collected for children aged 0 to 3 years. The dataset contained 86 attributes, subdivided into complementary attributes for the main attribute. For example, the attribute "41-Data of vaccination" included six additional fields such as "If < 6 months: the mother received the vaccine" and "If yes, date". The data from openDataSUS was provided in CSV format. To explore the data, MySQL Workbench was used along with SQL (Structured Query Language) programming.

The third phase, Data Preparation, aimed to transform the attributes in such a way as to make the dataset suitable for the application of ML algorithms. After analysis and testing, 100 attributes from the 2021 dataset and 95 from the 2020 dataset were removed, most of which related to internal identification codes and dates, as they were not relevant to the research. New attributes were created from existing ones, such as NU_IDADE_N (patient's age), DIAS_UTI (number of days in the ICU), among others. To enhance the interpretation of the models and facilitate the manipulation of attributes and instances in the ML tool, it was necessary to transform the dataset according to the definitions in the data dictionary. For example, for the attribute TOSSE (cough), the data point "1" was transformed to "Yes", and "2" to "No". After manipulating the data in MySQL, a CSV file of the database was generated via SQL command, which could be read by Weka software. Once loaded into Weka, the dataset was saved in ARFF format, which is Weka's standard. It is worth noting that Weka also allows for the manipulation of attributes and instances. The datasets from 2020 and 2021 were unified to facilitate record selection and manipulation, and an attribute "YEAR" was created to identify the records. Afterward, the datasets were divided according to the target patient groups of the research and loaded into Weka. After this process, the *AttributeSelection* filter in Weka was used to select the best attributes, employing the *CorrelationAttributeEva* and *ClassifierAttributeEval* resources with the *Ranker* method, which selects the best attributes according to the algorithms selected for the project. The *NominalToBinary* filter in Weka was also used to convert nominal attributes into binary numerical attributes in a separate version of the dataset. This conversion was necessary for the use of certain algorithms, such as XGBoost, which do not handle nominal data. After testing with the filters, attributes that did not perform well with the chosen algorithms were discarded.

In the fourth phase, *Modeling*, ML algorithms were studied and applied to the prepared datasets to generate predictive models according to the research objectives. This phase encompasses the entire process of model generation, validation, interpretation, and the selection of the best models. Activities in this phase focused on selecting the appropriate modeling techniques, defining the metrics for model approval, and building the models with tests on the algorithms' hyperparameters. In the fifth phase, *Evaluation*, the models are assessed and validated, analyzing whether the knowledge generated by these models will be used in the deployment phase. These two phases were executed simultaneously, as model generation and evaluation are part of the same process.

The algorithms Random Forest (RF), Logistic Regression (LR), XGBoost (Extreme Gradient Boosting), and KNN (K-Nearest Neighbors) were selected due to the common combination of these four algorithms in studies of this nature, as seen in the works of Moulaei et al. (2022), Schöning et al. (2021), and Kivrak et al. (2021).

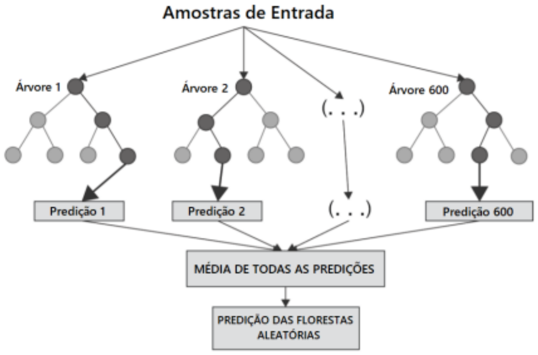
Random Forest

Random Forest (RF) consists of a classifier composed of multiple trees, or a forest of decision trees (SENA, 2021). In this algorithm, decision trees are constructed and represented by two elements: nodes and branches that connect nodes. For decision-making, the flow starts at the root node and navigates through the branches until reaching a leaf node. Each tree node denotes a test of an attribute, and the branches represent the possible values the node can assume. During tree formation, also known as training or learning, the homogeneity of the classes for each node division is considered. Essentially, the algorithm evaluates the information gain of attributes to separate the samples in the training dataset (LIMA et al., 2021). For instance, during the model's construction, three classifiers (trees) are built, and a new instance is labeled by each classifier. If the three classifiers make distinct errors, when one is incorrect, the second and third may be correct, so that the combination of hypotheses through voting can classify correctly. This technique, known as *bagging*, or *Bootstrap* Aggregation, is used in regression or classification models to improve model stability and accuracy (HU et al., 2021), (SILVA & NETO, 2022).

One of Random Forest's main advantages is the ease of measuring the relative importance of each attribute for prediction, automatically calculating this value for each attribute after training; the higher the value, the more important the attribute is. For this, the algorithm uses Gini Impurity (GI), an index for evaluating attributes in separating samples with the same label, seeking class homogeneity to form a node. The index assesses all predictors randomly selected to construct the tree and will choose the one with the highest degree of homogeneity among the samples (LIMA et al., 2021).

Figure 1 demonstrates how a random forest works in the classification process. It is worth noting that the final result is obtained by the average (in the case of regression) or by the majority of votes (in the case of classification) of the predictions of all the trees.

Figure 1 - Example of Random Forest Scheme



Fonte: SILVA E (2022)

Logistic Regression

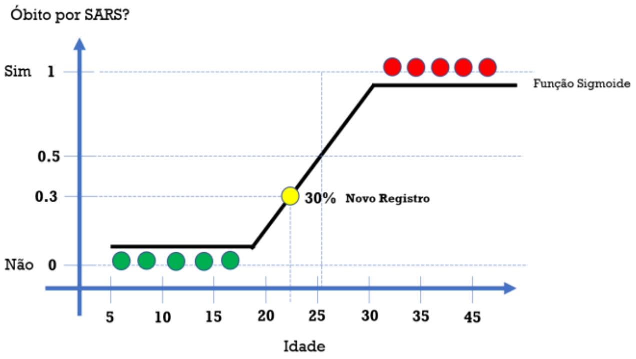
Logistic Regression (LR) is a linear model used for classification. It is also referred to in the literature as logit regression, maximum entropy classification, or log-linear classifier. Binary logistic regression refers to cases of logistic regression where the dependent variable is binary or dichotomous, meaning it can only take on two values (SILVA & NETO, 2022). Logistic regression is used to estimate the association between one or more independent (predictor) variables and a binary dependent variable (outcome). A binary (or dichotomous) variable is a categorical variable that can only assume two distinct values or levels, such as "dead" or "alive," for example. Logistic regression can be used to estimate the probability (or risk) of a specific outcome based on the values of the independent variables. It is important to note that this probability is given as a value between 0 and 1, that is, 1 for "alive" and 0 for "dead" in the aforementioned example (SCHOBER & VETTER, 2021).

The general formula for logistic regression applies the sigmoid function to the linear combination of the independent variables, which allows the linear output to be transformed into a probability between 0 and 1. The following Equation 1 shows the logistic regression formula for binary classification problems:

$$P(Y = 1) = \frac{1}{1+e^{(\beta^0+\beta^1 X^1+\beta^2 X^2+\cdots+\beta_K X_K)}} \qquad (1)$$

where P(Y=1) represents the probability of the event of interest occurring, $\beta^0$is the intercept, β^1, β^2,..., β^K are the coefficients of the independent variables X^1, X^2, ..., X^K, e $X^1$and is the base of the natural logarithm, also called Euler's number which corresponds to the number 2.71 (HOSMER et al., 2013). Figure 2 presents an example of logistic regression classification.

Figure 2 - Example of a Classification Model with Logistic Regression

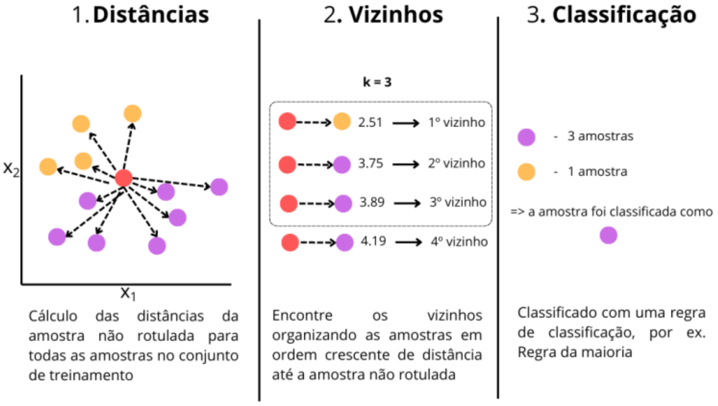


Fonte: Autor

KNN – *K-Nearest Neighbors*

The K-Nearest Neighbors (KNN) algorithm is a widely used machine learning method for binary classification problems due to its simplicity and effectiveness. This algorithm also supports non-binary classification and regression tasks. The fundamental principle of KNN is to determine the class of a data point based on the classes of its nearest neighbors in a multidimensional space. In a binary classification problem, each data point is labeled with one of two possible classes, and the objective of KNN is to predict the class of new data points based on previous observations. To implement KNN, it is first necessary to choose a value for K, which represents the number of neighbors to be considered (MLADENOVA & VALOVA, 2023). Then the algorithm calculates the distance between the data point to be classified and all the data points in the training set. The distance can be any metric measure, such as Manhattan distance, Minkowski distance, or Euclidean distance, the latter being one of the most commonly used (SILVA E, 2022). Once the distances are calculated, KNN identifies the K nearest data points and determines the predominant class among these neighbors (MLADENOVA & VALOVA, 2023). Figure 3 below illustrates this process:

Figure 3 - Classification steps with KNN


1. Distâncias
2. Vizinhos
3. Classificação
x_2
x_1
Cálculo das distâncias da amostra não rotulada para todas as amostras no conjunto de treinamento
k = 3
2.51 → 1º vizinho
3.75 → 2º vizinho
3.89 → 3º vizinho
4.19 → 4º vizinho
Encontre os vizinhos organizando as amostras em ordem crescente de distância até a amostra não rotulada
- 3 amostras
- 1 amostra
=> a amostra foi classificada como
Classificado com uma regra de classificação, por ex. Regra da maioria

Fonte: Mladenova & Valova, 2023 (traduzido)

XGBoost – *Extreme Gradient Boosting*

XGBoost - Extreme Gradient Boosting, is a decision tree-based ML algorithm designed to be highly efficient and scalable. It uses a boosting approach, where multiple trees are built sequentially, each correcting the errors of the previous one (CHEN & GUESTRIN, 2016). XGBoost is an iterative decision tree algorithm with multiple decision trees. Each tree is learning from the residuals of all previous trees. Instead of adopting the majority of voting output results in the Random Forest algorithm, the predicted output of XGBoost is the sum of all the results (WANG et al., 2019). XGBoost creates a model that is the sum of multiple decision trees from the following Formula 2:

$$\hat{y}_i = \sum_{k=1}^{n} \quad f_k(x_i), f_k \in F(2)$$

where F stands for the space of regression trees,f_k corresponds to a tree, then $f_k(x_i)$is the result of tree k,∧ $\hat{y}_i$ is the predicted value of the *i-th* instance x_i (WANG et al., 2019). The main objective of XGBoost is to minimize a regularized cost function that includes both the loss function, which measures the discrepancy between model predictions and actual values, and regularization terms that penalize model complexity to avoid overfitting. This additional regularization differentiates XGBoost from other boosting algorithms, making it more robust and able to generalize better to new data (WANG et al., 2019). The objective function is given by Equation 3 below:

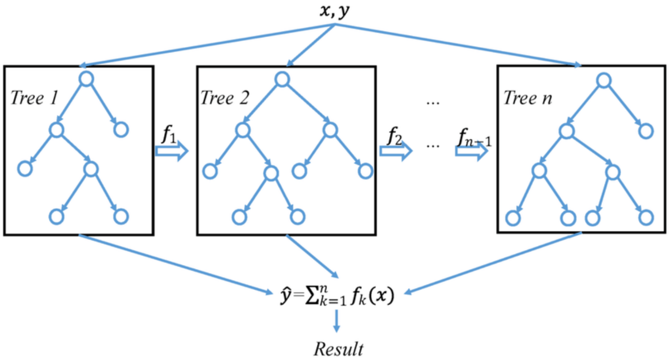
$$Obj(\theta) = L(\theta) + \Omega(\theta)(3)$$

where L(θ) is the loss function that measures the difference between the predictions (i) and the actual values (WANG et al., 2019). Finally, for binary classification the common loss function is the *log-loss* given by Equation 4 below, where $l(yi, \hat{y}_i)$ is the logarithmic loss between the real value and the prediction (i) (WANG et al., 2019).

$$L(\theta) = \sum_{i=1}^{n} \quad l(yi, \hat{y}_i)(4)$$

Figure 4 below illustrates the working process of the XGBoost algorithm, highlighting how it combines multiple decision trees to form a robust and accurate model.

Figure 4 - XGBoost Architecture



Fonte: Wang et al. (2019)

At the top of the image, the input variables x (features) and y (labels) are provided to the model. The process begins with the construction of the first decision tree (Tree 1). The prediction function of this tree is denoted by f_1. Then the second tree (Tree 2) is constructed. It is based on the residuals or errors of the predictions of the first tree and its prediction function is f_2. This process continues successively, with each tree trying to correct the errors of the predictions of the previous trees. The predictions of all trees are combined to form a final prediction. The formula that represents this combination is given by Equation 1 and the final prediction $\hat{y}_i$ is the sum of the predictions of all trees (WANG et al., 2019).

Evaluation Metrics in the Modeling and Evaluation Phase

The *Cross-Validation* process was used to evaluate the performance and overall error of models. Cross-validation is a resampling procedure used to assess machine learning models on a data sample. The procedure has a single parameter called *k*, which represents the number of groups used to divide a given data sample. In 10-fold cross-validation (the standard number of folds), models are trained and tested ten different times, and the average performance metrics (e.g., accuracy, precision, etc.) are estimated at the end of the process (KIVRAK et al., 2021). It is important to note that cross-validation is a widely applied and preferred validation technique in machine learning and data mining due to its difference from the conventional split-sample method. This method helps reduce bias in prediction error, maximizes the use of data for both training and validation without overfitting or overlap between test and validation data, and avoids arbitrary data splitting, which can introduce bias into the model's results (MOULAEI et al., 2022). For the cross-validation of training and testing models, 20 iterations (*folds*) were used, as found in the literature (YU et al., 2021; ZAREI et al., 2022; AN et al., 2020; SUN et al., 2021; MAHDAVI et al., 2021). Thus, the generated models with the seeds that achieved the best performance were documented.

Evaluation metrics for the models were defined based on current literature on the subject. Model performance evaluation is a crucial part of building an effective machine learning model. Several metrics are applied to assess predictive models, with the most common being accuracy, specificity, precision, sensitivity, and criteria from the ROC (Receiver Operating Characteristic) curve. Finally, these evaluation criteria are compared to determine the best predictive model (MOULAEI et al., 2022). To calculate these performance metrics, the confusion matrix of the generated model must be obtained. The confusion matrix is a table used to assess the performance of a classification model by displaying the number of correct and incorrect predictions, organized according to actual and predicted classes. It allows the visualization of correct predictions (True Positives – TP and True Negatives – TN) and errors (False Positives – FP and False Negatives – FN). In the confusion matrix, TN corresponds to the number of negative results correctly classified, TP is the number of positive results correctly classified, FP is the number of negative results incorrectly classified as positive, and FN is the number of positive results incorrectly classified as negative (BÁRCENAS et al., 2022; BOOTH et al., 2021). Table 1 below presents the format of the confusion matrix and the position of correct and incorrect predictions.

Table 1 - Confusion Matrix Model

		Expected Value	
		Death (+)	**Cure (−)**
Real Value	**Death (+)**	TP	FN
	Cure (−)	FP	TN

Fonte: Moulaei et al. (2022)
Note: TP - True Positives, TN - True Negatives,
FP - False Positives and FN - False Negatives

The accuracy, precision, sensitivity and specificity metrics of the models are calculated from the confusion matrix data, as highlighted in Table 2 below.

Table 2 - Performance Criteria Calculations

Performance Criteria	**Cálculo**
Accuracy	(TP + TN) / (TP + TN + FP + FN)
Precision	TP / (TP + FP)
Sensitivity (Recall)	TP / (TP + FN)
Specificity	TN / (TN + FP)
F1-Score	2 * Accuracy * Sensitivity / (Accuracy + Sensitivity)

Source: Moulaei et al. (2022) and Bárcenas & Fuentes-García (2022)
Note: TP - True Positives, TN - True Negatives,
FP - False Positives and FN - False Negatives

Accuracy represents the total percentage of correct predictions of a model. However, this metric is not always the best for evaluating classification models, especially in cases of imbalanced datasets. In situations where one class is significantly more frequent than the other, accuracy can be misleading, failing to reflect the true performance of the model across all classes. This may lead the analyst to believe that the model is good at correctly predicting class A, while making many errors in predicting class B. Therefore, it is important to consider other metrics beyond accuracy, such as precision, sensitivity (or recall), and the F1-score. Precision measures the model's ability to avoid false positives, indicating the percentage of correct predictions among all instances classified as positive. Sensitivity, or recall, indicates the model's ability to correctly identify all positive instances, showing the percentage of correct predictions among all instances that are indeed positive. The F1-score combines precision and sensitivity into a harmonic mean, providing a balanced evaluation of the model's performance, especially in imbalanced datasets (SILVA & NETO, 2022).

Another important metric is the ROC curve and the calculation of the AUC (*Area Under the Curve*). The ROC curve measures the predictive ability of the model through the sensitivity and specificity rates, representing these metrics in a graph. The AUC quantifies the total area under the ROC curve and provides a single metric for the model's performance, independent of a specific decision threshold. This technique is used to visualize, organize, and rank the model based on its predictive performance. Practically speaking, the closer the curve is to the top left corner of the graph, the better the model's performance (SILVA & NETO, 2022). The AUC is the result of integrating all points along the curve's path and simultaneously computes both sensitivity and specificity, serving as an estimator of the overall test accuracy. It provides an estimate of the probability of correctly classifying a subject by chance (test accuracy); for example, an AUC of 0.7 reflects a 70% chance of correct classification. In general, AUC values are interpreted as follows: 0.5-0.6 (very poor), 0.6-0.7 (bad), 0.7-0.8 (poor), 0.8-0.9 (good), and > 0.9 (excellent) (POLO & MIOT, 2020). Finally, it is worth noting that many authors in the literature (KIVRAK et al., 2021; SILVA & NETO, 2022; FERNANDES et al., 2021; VEPA et al., 2021; BÁRCENAS et al., 2022; SUN et al., 2021; BENNETT et al., 2021; ARAÚJO et al., 2022) have used accuracy, sensitivity, specificity, precision, F1-score, and AUC-ROC metrics to evaluate their predictive models for mortality from severe acute respiratory syndrome (SARS). In this context, as noted by Bennett et al. (2021) and Moulaei et al. (2022), AUC-ROC was considered the primary metric, while sensitivity, specificity, accuracy, precision, and F1-score were used as secondary metrics to assess and define the best model.

Analyzing the importance of attributes in predictive models is essential for understanding which factors have the greatest influence on outcomes. The evaluation of attribute importance in a Random Forest model uses the Gini index reduction to determine which variables contribute most significantly to the prediction of results, highlighting the most influential factors in the classification process (MOSLEHI et al., 2022; BÁRCENAS & FUENTES-GARCÍA, 2022). To obtain the Gini index, the R library was used via the Weka interface, as the original Weka library does not directly generate the index. For this purpose, the script was programmed and executed as follows:

```
library(randomForest)
data <- rdata
data_sem_missing <- na.omit(data)
modelo <- randomForest(EVOLUCAO ~ ., data = data_sem_missing)
importancia <- importance(modelo)
print(importancia)
```

The representation of the index in a graph is common in the literature (KUMARAN et al., 2022), (MOSLEHI et al., 2022), (BÁRCENAS & FUENTES-GARCÍA, 2022), (ZHAO et al., 2022) (AZNAR-GIMENO et al., 2021), (HELDT et al., 2021) and facilitates understanding. Therefore, the Gini indices of the *Random Forest* models were demonstrated by graphs..

Balancing Experiment in the Modeling and Evaluation Phase

An imbalance in the data was identified. Moulaei et al. (2022) highlights that one of the main challenges for machine learning algorithms is the problem of imbalanced data, which occurs when classes are not equally represented. As a result, models trained on such data tend to show bias toward the dominant class, potentially leading to a tendency to categorize new observations into the majority class. Upon reviewing the RI studies, it was found that the authors addressed this imbalance in various ways. Azgnar-Gimeno et al. (2021), Moulaei et al. (2022), Heldt et al. (2021), Zarei et al. (2022), Araújo et al. (2022), and Vepa et al. (2021) used the Synthetic Minority Oversampling Technique (SMOTE) to balance the dataset. This technique creates synthetic instances of the minority class based on known patterns, matching the size of the majority class (ARAÚJO et al., 2022; MOULAEI et al., 2022). On the other hand, studies by Li J et al. (2022), Schöning et al. (2021), Booth et al. (2020), Gao et al. (2020), and An et al. (2020) used class weighting techniques to automatically adjust the weights of instances so that each class is equally important during model training (BOOTH et al., 2020; LI J et al., 2022). Finally, authors such as Woo et al. (2022), Yadaw et al. (2020), Bottrighi et al. (2022), Li Y et al. (2020), Yu L et al. (2021), and Bárcenas & Fuentes-García (2022) assumed that the data were imbalanced and did not address balancing. Thus, a balancing experiment was conducted using different techniques to identify possible improvements in model performance and the practical implications of balancing, following current literature. The experiment was performed using the *Random Forest* algorithm.

Table 3 below shows that the results with SMOTE yielded superior performance in sensitivity and F1-score compared to the other models, but with little variation in AUC-ROC performance. However, this performance gain comes at the cost of creating many synthetic instances for the "Death" class, which may represent patterns that do not exist in the real data. Conversely, the class weighting balance performed using the ClassBalancer filter in Weka showed slightly better sensitivity compared to the unbalanced model, but with inferior AUC-ROC performance. In this regard, the decision was made to use the unbalanced data since there were no significant improvements with balancing, following the approach of Araújo et al. (2022), which indicates that recent studies suggest "imbalance is not a problem in itself: correction methods for imbalance can cause poor calibration and even worsen model performance in terms of AUC-ROC." Additionally, the F1-score metric, evaluated in this research, provides a global assessment of the model regardless of the sample size in each class.

Table 3 - Comparison of the Balancing Experiment for Positive Death Outcome

Metrics	Unbalanced	Balanced with SMOTE	Balanced with *ClassBalancer*	Average	Standard Deviation
True Positives (TP)	253	6665	3523	3480	3206
False Positives (FP)	11	35	113	53	53
True Negatives (TN)	7066	7042	3620	5909	1983
False Negatives (FN)	239	223	1213	558	567
Precision	0.958	0.995	0.969	0.974	0.019
F1-Score	0.669	0.981	0.842	0.831	0.156
Sensitivity	0.514	0.968	0.744	0.742	0.227
Specificity	0.998	0.995	0.970	0.988	0.015
Accuracy	0.966	0.981	0.843	0.930	0.076
AUC-ROC	0.950	0.996	0.946	0.964	0.028

Source: Author

In the CRISP-DM methodology, the implementation phase describes the use of knowledge generated by the project within an organization. However, since this is an academic work, the first activity of this phase was adapted to provide the implementation of knowledge through a software application. The software application was developed in Java language with the Apache Netbeans IDE 20 tool for desktop format, that is, it can be installed on any PC device. The weka.jar code library was used to access the model's loading, classification and evaluation features. The choice of the Java programming language was due to the possibility of using the weka.jar library, in addition to the author's experience with the language.

RESULTS

Weka was chosen due to its ability to integrate Python ML libraries and the R software library directly into its interface, making Weka a comprehensive tool with a user-friendly interface. The choice of this tool aligns with the literature on the subject, where Weka was used by authors Bottrighi et al. (2022) and Moulaei et al. (2022) in their research on predictive models for mortality due to severe acute respiratory syndrome (SARS). In addition to the previously mentioned tools, MySQL Workbench by Oracle was used to manipulate and transform the database records. This tool was selected due to its ability to program SQL scripts, allowing for automation of the process. Another key factor was MySQL's ability to handle large data volumes.

The SQL script used for cleaning and transforming the 2020 and 2021 datasets is available in the Zenodo file repository under the DOI – Digital Object Identifier at the link: https://doi.org/10.5281/zenodo.10850628. The list of all excluded attributes can be found in the SQL cleaning script from line 366, marked with the comment "*#cleaning of unselected attribute base.*" The unified database with records from 2020 and 2021 includes a total of 291,775 patients from the Northern region, considered eligible for model application, and is available in the Zenodo repository at https://zenodo.org/doi/10.5281/zenodo.12636544 in ARFF format, which can be read by Weka. After this cleaning and transformation process, the database was made available in the Zenodo repository in ARFF format at https://zenodo.org/doi/10.5281/zenodo.10879240, containing 9471 records. Due to missing data in the "evolution" class attribute, 3204 records were excluded, leaving 7569 records for model generation, with 7077 cases of recovery and 492 deaths. Thus, the final version of the dataset used for the modeling phase contained 40 attributes.

The attribute EVOLUTION was defined as the class, with "Cure" or "Death" as possible outcomes. The models were evaluated by considering the "Death" class as positive, as the goal of the predictive model is to predict the death of patients from SARS. Regarding cross-validation, Weka's "Random Seed for XVal" feature was tested with 20 different seeds for each algorithm and dataset. Accuracy was evaluated, and the standard deviation of the means was less than 0.01% in all tests, indicating no statistically significant differences. The documented models were generated using seed 8 for *Random Forest* (RF), 11 for *Logistic Regression* (LR), 8 for KNN, and 20 for XGBoost. To obtain highly reliable models capable of efficiently predicting the "Death" class, various experiments were conducted to find the best hyperparameters for each analyzed ML model. The following hyperparameters were determined: For *Random Forest*, the number of trees was set to 110; for KNN, the number of neighbors was set to 1, with the *Euclidean Distance* function; for XGBoost, the R library was used via

Weka's interface with the dataset transformed into binary data; and finally, for *Logistic Regression*, Weka's default settings were used.

Métricas do Modelo Preditivo

Table 4 below presents the confusion matrix data of the generated models.

Table 4 - Confusion Matrix

	Predição	
RandomForest	**Death (+)**	**Cure (−)**
Death (+)	261	231
Cure (−)	9	7068
Logistic Regression	**Death (+)**	**Cure (−)**
Death (+)	106	386
Cure (−)	83	6994
KNN	**Death (+)**	**Cure (−)**
Death (+)	327	165
Cure (−)	98	6979
XGBoost	**Death (+)**	**Cure (−)**
Death (+)	139	353
Cure (−)	70	7007

Source: Adapted from Kivrak et al. (2021)

Table 5 below presents the performance metrics of the ML algorithms on the generated models.

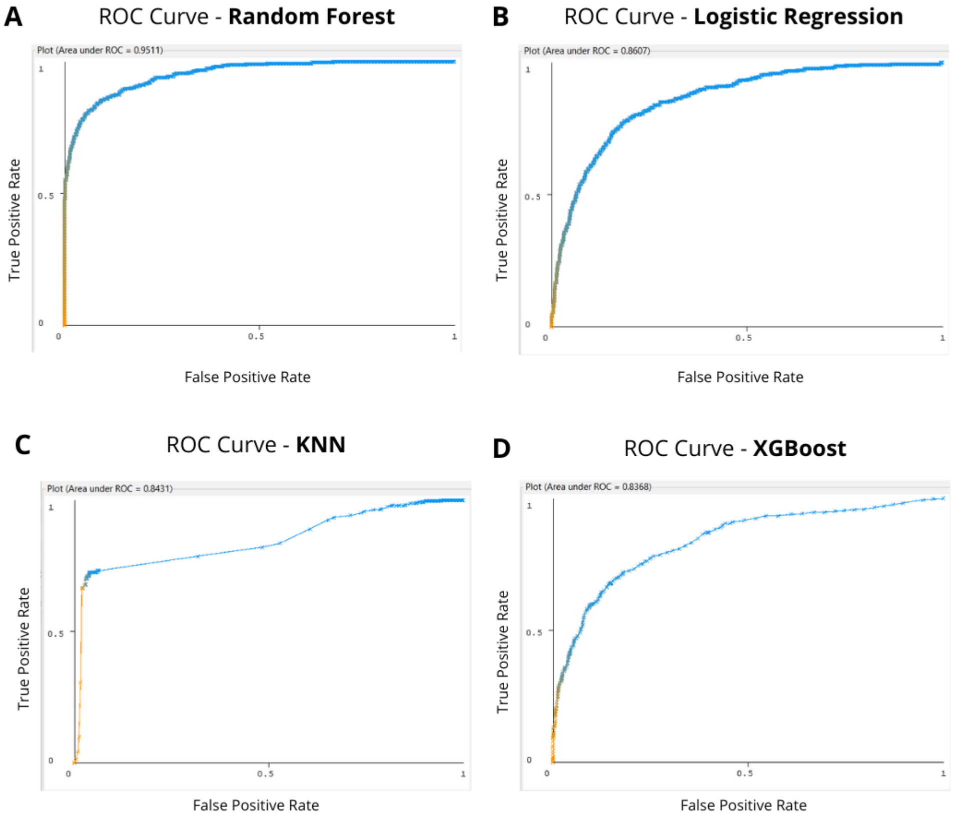
Table 5 - Performance Evaluation of Algorithms

Algorithms	**Sensitivity**	**Specificity**	**Accuracy**	**Precision**	**F1-Score**	**AUC-ROC**
Random Forest	0.530	0.999	0.968	0.967	0.685	0.951
Logistic Regression	0.215	0.988	0.938	0.561	0.311	0.861
KNN	0.665	0.986	0.965	0.769	0.713	0.843
XGBoost	0.283	0.990	0.944	0.665	0.397	0.837

Source: Adapted from Moulaei et al. (2022)

Figure 5 below shows graphs with the ROC Curve and AUC of each algorithm for the purpose of comparing the performance of the generated models, where the superior performance of the model created with the *Random Forest* algorithm can be seen.

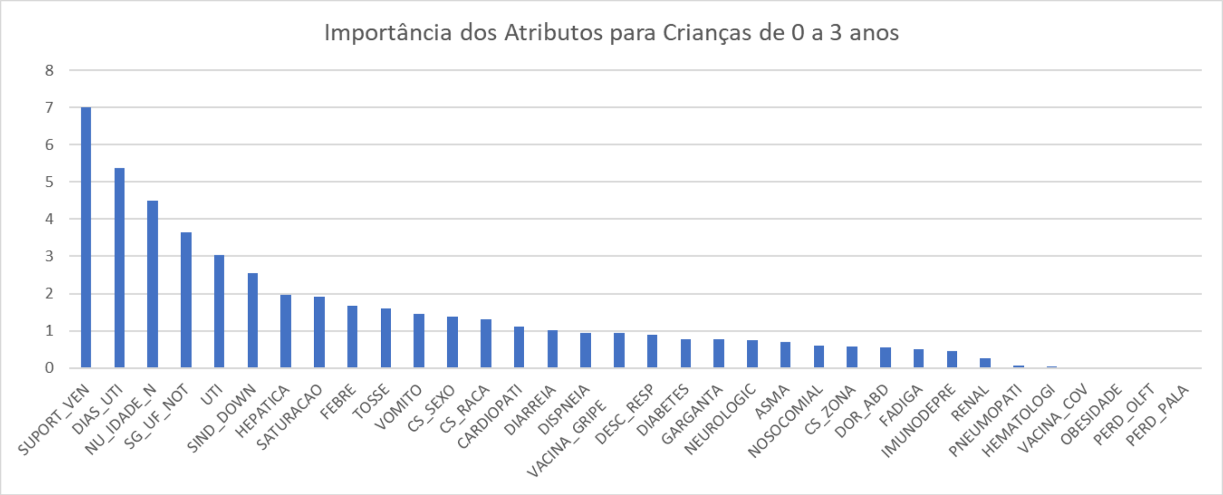
Figure 5 - AUC-ROC of the Models



Source: Adapted from Silva & Neto (2022)

Figure 6 below shows the graph with the most important attributes considered by the model with Random Forest.

Figure 6 - Graph with Gini Index

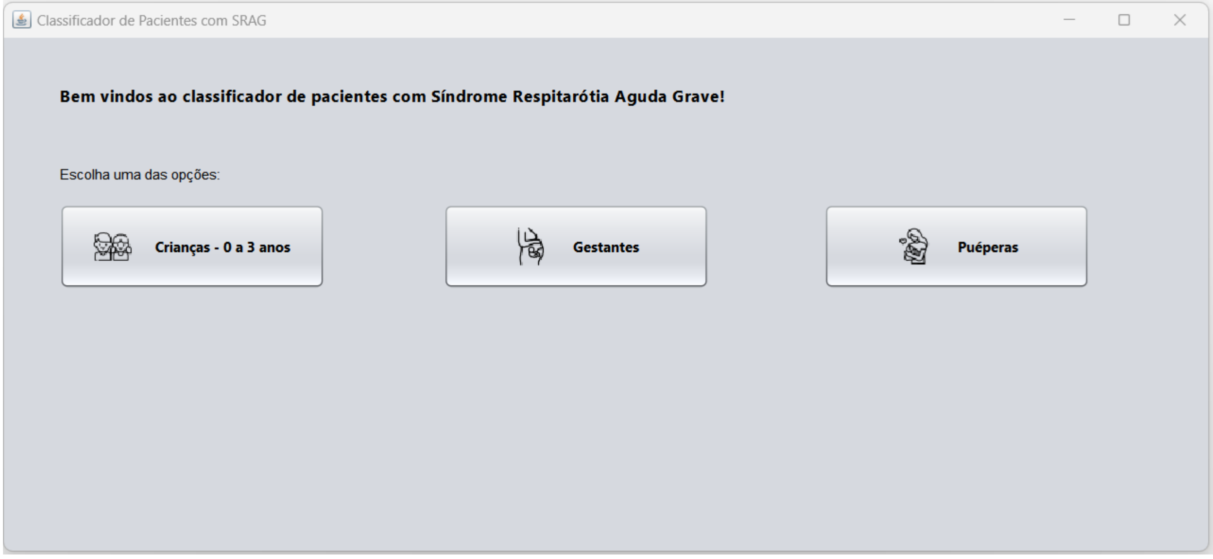


Source: Adapted from Zhao et al. (2022)

Scenario Simulation with the Application Prototype

The following images will show the application's features, which can be downloaded and used by any user with access to a desktop computer. Figure 7 below shows the application's initial menu with the options.

Figure 7 - Application Home Menu



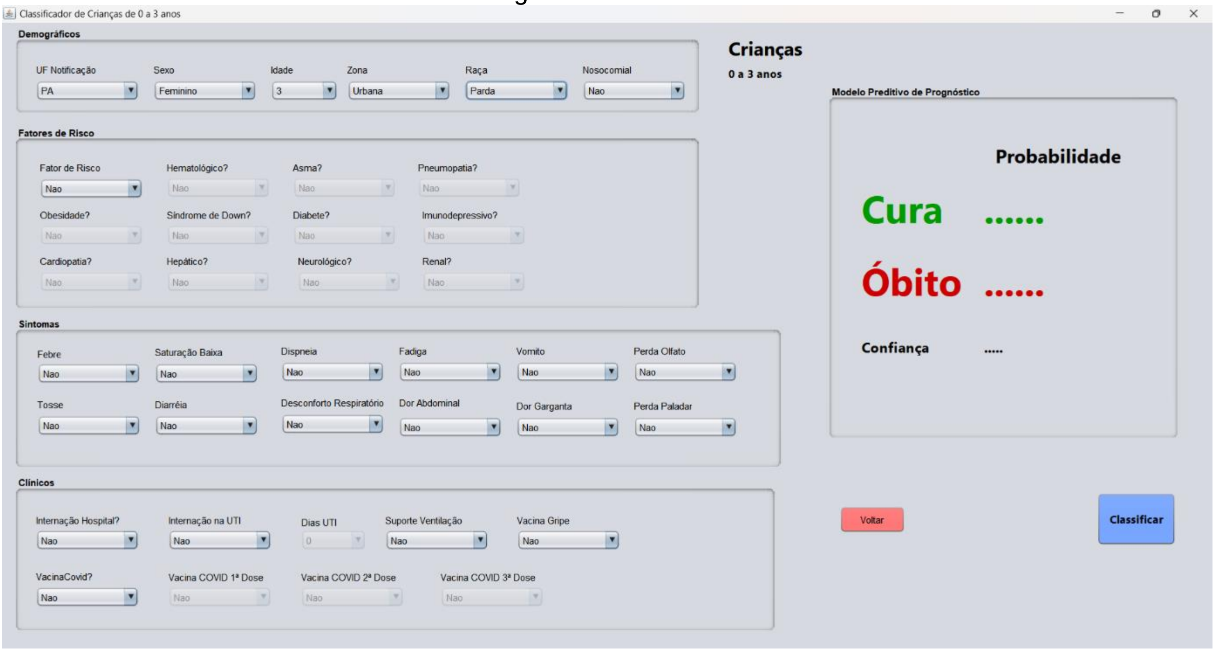
Bem vindos ao classificador de pacientes com Síndrome Respitarótia Aguda Grave!

Escolha uma das opções:

Source: Author

In the initial menu, the user can select three different classifiers, depending on the profile of the patient they wish to classify. It is worth noting that the audiences were defined according to the research objectives. After selecting the desired profile in the initial menu, the system will open the classifier for the group. Figure 8 below shows the classification functionality without filling in the patient characteristics.

Figure 8 - Classifier



Classificador de Crianças de 0 a 3 anos

Demográficos

UF Notificação	Sexo	Idade	Zona	Raça	Nosocomial
PA	Feminino	3	Urbana	Parda	Nao

Fatores de Risco

Fator de Risco	Hematológico?	Asma?	Pneumopatia?
Nao	Nao	Nao	Nao
Obesidade?	Síndrome de Down?	Diabete?	Imunodepressivo?
Nao	Nao	Nao	Nao
Cardiopatia?	Hepático?	Neurológico?	Renal?
Nao	Nao	Nao	Nao

Sintomas

Febre	Saturação Baixa	Dispneia	Fadiga	Vomito	Perda Olfato
Nao	Nao	Nao	Nao	Nao	Nao
Tosse	Diarréia	Desconforto Respiratório	Dor Abdominal	Dor Garganta	Perda Paladar
Nao	Nao	Nao	Nao	Nao	Nao

Clínicos

Internação Hospital?	Internação na UTI	Dias UTI	Suporte Ventilação	Vacina Gripe
Nao	Nao	0	Nao	Nao
VacinaCovid?	Vacina COVID 1ª Dose	Vacina COVID 2ª Dose	Vacina COVID 3ª Dose	
Nao	Nao	Nao	Nao	

Crianças
0 a 3 anos

Modelo Preditivo de Prognóstico

Probabilidade

Cura ......

Óbito ......

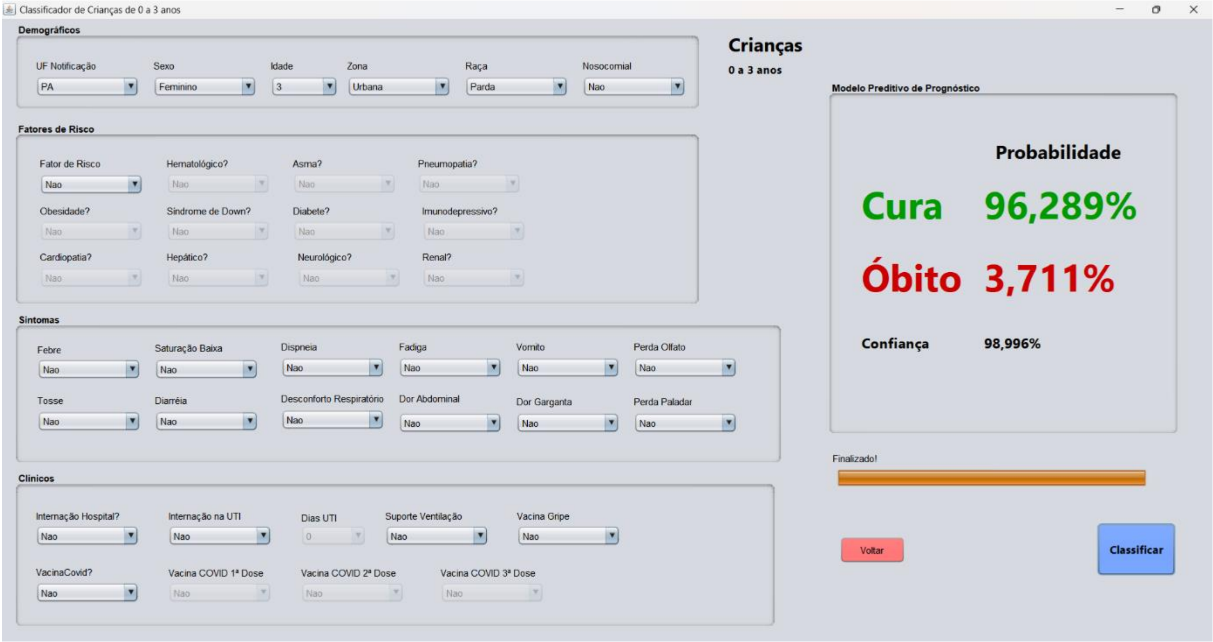
Confiança .....

Source: Author

The user can then enter the characteristics of the patient to be classified using combo boxes and then click on the *Classify* button. Or, they can return to the Home Menu by clicking on the Back button. During the classification process, the system will display a progress bar while the classification is taking place. After the process is complete, the chances of death and cure for the patient will be displayed. The confidence shown is the percentage of accuracy of the model based on the test database, i.e., the accuracy of the model. This value is calculated by the application when testing the model with all records in the database, after training the model using cross-validation.

Figure 9 below shows a scenario simulation, with the change in the functionality state after the classification action, where the probabilities of death and cure predicted by the model for the child are presented according to the characteristics informed. It can be seen that in this scenario the child does not have risk factors, symptoms or has been hospitalized. Therefore, the probability of cure predicted by the model is high.

Figure 9 - Children Classifier with Result of Scenario I Simulation



Classificador de Crianças de 0 a 3 anos

Crianças

0 a 3 anos

Demográficos

UF Notificação: PA
Sexo: Feminino
Idade: 3
Zona: Urbana
Raça: Parda
Nosocomial: Nao

Fatores de Risco

Fator de Risco: Nao
Hematológico?: Nao
Asma?: Nao
Pneumopatia?: Nao
Obesidade?: Nao
Sindrome de Down?: Nao
Diabete?: Nao
Imunodepressivo?: Nao
Cardiopatia?: Nao
Hepático?: Nao
Neurológico?: Nao
Renal?: Nao

Sintomas

Febre: Nao
Saturação Baixa: Nao
Dispneia: Nao
Fadiga: Nao
Vomito: Nao
Perda Olfato: Nao
Tosse: Nao
Diarréia: Nao
Desconforto Respiratório: Nao
Dor Abdominal: Nao
Dor Garganta: Nao
Perda Paladar: Nao

Clinicos

Internação Hospital?: Nao
Internação na UTI: Nao
Dias UTI: 0
Suporte Ventilação: Nao
Vacina Gripe: Nao
VacinaCovid?: Nao
Vacina COVID 1ª Dose: Nao
Vacina COVID 2ª Dose: Nao
Vacina COVID 3ª Dose: Nao

Modelo Preditivo de Prognóstico

Probabilidade

Cura 96,289%

Óbito 3,711%

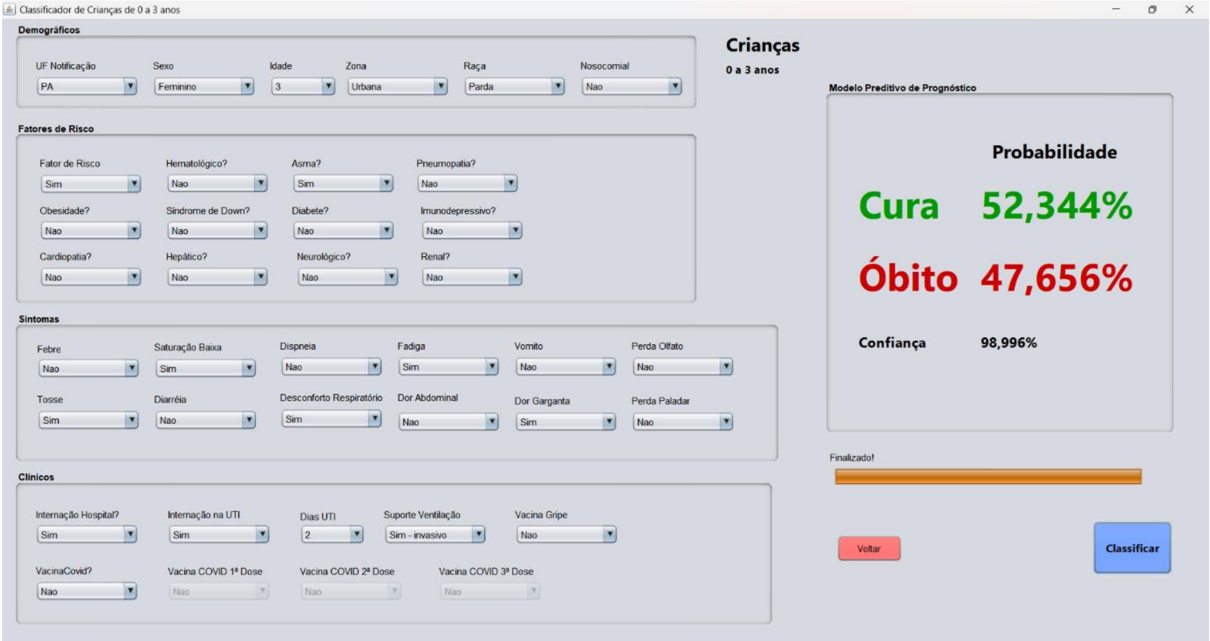
Confiança 98,996%

Finalizado!

Fonte: Autor

The probability is given by the prediction model after classifying the patient based on the set of characteristics informed in the interface before clicking the *Classify* button. In the simulated scenario, the probabilities of Cure and Death change as the patient's characteristics change. Figure 10 below shows the classification for the child in the simulation of a second scenario, where characteristics about risk factors and symptoms common to SARS were inserted.

Figure 10 - Children Classifier with Result of Scenario II Simulation



Classificador de Crianças de 0 a 3 anos

Crianças

0 a 3 anos

Demográficos

UF Notificação	Sexo	Idade	Zona	Raça	Nosocomial
PA	Feminino	3	Urbana	Parda	Nao

Fatores de Risco

Fator de Risco	Hematológico?	Asma?	Pneumopatia?
Sim	Nao	Sim	Nao
Obesidade?	Síndrome de Down?	Diabete?	Imunodepressivo?
Nao	Nao	Nao	Nao
Cardiopatia?	Hepático?	Neurológico?	Renal?
Nao	Nao	Nao	Nao

Sintomas

Febre	Saturação Baixa	Dispneia	Fadiga	Vomito	Perda Olfato
Nao	Sim	Nao	Sim	Nao	Nao
Tosse	Diarréia	Desconforto Respiratório	Dor Abdominal	Dor Garganta	Perda Paladar
Sim	Nao	Sim	Nao	Sim	Nao

Clínicos

Internação Hospital?	Internação na UTI	Dias UTI	Suporte Ventilação	Vacina Gripe
Sim	Sim	2	Sim - invasivo	Nao
VacinaCovid?	Vacina COVID 1ª Dose	Vacina COVID 2ª Dose	Vacina COVID 3ª Dose	
Nao	Nao	Nao	Nao	

Modelo Preditivo de Prognóstico

	Probabilidade
Cura	52,344%
Óbito	47,656%
Confiança	98,996%

Finalizado!

Source: Author

The simulation of scenarios presented previously demonstrates the reduction in the chances of cure and the increase in the chances of death as the patient's characteristics change, highlighting the predictive model's ability to deal with factors related to the degradation of the patient's health. Finally, the application's source code is deposited in the GitHub code repository (private access) and can be accessed and downloaded via the link: https://github.com/jacksonifro/Aplication_Tese_Doutorado.git upon request. To open the application, the Apache Netbeans IDE 20 tool is required. The application's installation setup to be installed on Windows or Linux operating systems is available for download in the Zenodo repository via DOI: https://zenodo.org/doi/10.5281/zenodo.10951429.

DISCUSSION

This study represents an important advance in the creation of classification models capable of identifying patients at higher risk of death from SARS in vulnerable population groups in the North region. Predictive models for classification were developed and compared with four different algorithms: *Random Forest*, *Logistic*

Regression, KNN and XGboost. The models were evaluated according to the metrics of sensitivity, specificity, accuracy, precision, F1-Score and AUC-ROC, the latter being the primary evaluation metric. As highlighted by Polo & Miot (2020), an AUC-ROC greater than 0.90 is considered an excellent performance index for a quantitative data model according to its sensitivity rate (fraction of true positives) and the fraction of false positives (1 - specificity), according to different test cutoff values. Thus, the following discussions consider this threshold for assessing the quality of the model in terms of robustness and reliability.

The model generated with the *Random Forest* algorithm offers robust and reliable performance, achieving an AUC-ROC of 0.951, sensitivity of 0.530, specificity of 0.999, accuracy of 0.968, precision of 0.967 and F1-Score of 0.685. These results indicate an excellent ability to distinguish between classes. Although it was surpassed by KNN in sensitivity and F1-Score with 0.665 and 0.713 respectively, the overall balance of the other metrics makes its performance superior. It is worth mentioning that despite KNN's advantage in sensitivity, its performance is inferior to that of Random Forest and Logistic Regression in AUC-ROC, where it reached only 0.843. Another important point is that despite the imbalance of the classes, it was found that there was no great advantage of KNN over *Random Forest*, with a difference of only 0.028 in F1-Score. In this context, the *Random Forest* algorithm obtained the best overall performance, with the model generated by it being chosen for classifying children aged 0 to 3 years in the application.

These results are in line with the literature on the subject. Heldt et al. (2021) evaluated the performance of the *Random Forest*, *Logistic Regression*, and XGBoost algorithms using a dataset of 619 English patients with demographic, clinical, and laboratory data to predict mortality from SARS-Cov-2. *Random Forest* generated the best model with AUC-ROC of 0.77, against 0.70 and 0.76 from *Logistic Regression* and XGBoost, respectively. In another study (MOULAEI et al., 2022), demographic, clinical, laboratory, and risk factor data from 1,500 Iranian patients hospitalized with SARS-Cov-2 were used. *Random Forest* generated the best model with AUC-ROC of 0.77, against 0.70 and 0.76 of *Logistic Regression* and XGBoost respectively. In another study (MOULAEI et al., 2022), demographic, clinical, laboratory and risk factor data from 1,500 Iranian patients hospitalized with SARS-Cov-2 were used. The results of this study showed that the model developed using the *Random Forest* algorithm performed the best, with an AUC-ROC of 0.99 in predicting patient death, against the AUC-ROC of other compared algorithms such as XGBoost (0.981), KNN (0.967), MLP (0.964), *Logistic Regression* (0.942), J48 (0.921) and *Naive Bayes* (0.920). In a study focused on the Brazilian population, Silva & Neto (2022) used clinical data from 134,639 patients with SARS-Cov-2 registered in the openDataSUS SARS Database between January and September 2021 to evaluate the performance of the *Logistic Regression*, *Decision Tree* and *Random Forest* algorithms in creating predictive death models. In this study, *Random Forest* was superior, achieving an AUC-ROC of 0.75, accuracy of 0.77, precision of 0.76, f1-score of 0.69, and sensitivity of 0.63 for the death class. The *Logistic Regression* algorithm achieved an AUC-ROC of 0.73 and *Decision Tree* of 0.74, being inferior to *Random Forest* in this and other metrics, except for *Decision Tree*, which was slightly superior in precision with 0.78.

The KNN algorithm performed well in this study, achieving an AUC-ROC higher than 0.84 in the models. Bottrighi et al. (2022) obtained an AUC-ROC of 0.81 with the KNN algorithm in a study with 824 Italian patients using demographic data, comorbidities, and symptoms, being surpassed by the JRIP algorithm. The authors Altini et al. (2021) used the KNN algorithm in the comparison with other algorithms using demographic, clinical, and laboratory data from 303 Italian patients, where the algorithm achieved an AUC-ROC of 0.778, being surpassed by the *Decision Tree* algorithm with an AUC-ROC of 0.896.

The *Logistic Regression* algorithm also achieved good performance in the models analyzed in this study, achieving an AUC-ROC higher than 0.86. These results are similar to those found by other studies on predictive death modeling, such as those found by authors Hu et al. (2021) and Reina et al. (2022), who obtained AUC-ROC performance of 0.895 and 0.871, respectively, which is superior when compared to other algorithms such as *Random Forest*, SVM, KNN, and MLP. Authors Murri et al. (2021) and Woo et al. (2021) also achieved superior performance of 0.87 and 0.81, respectively, in AUC-ROC with *Logistic Regression*, but these authors worked with only one algorithm and did not compare it with other studies.

It is also worth noting that the XGBoost algorithm also achieved good performance in the models analyzed in this study, with an AUC-ROC greater than 0.83.

These results are in line with the findings of Aznar-Gimeno et al (2021) who obtained an AUC-ROC of 0.821 with the XGBoost algorithm in a study with 3,623 Spanish patients, outperforming the *Random Forest* algorithm. Also by Bárcenas & Fuentes-García (2022) who achieved an AUC-ROC of 0.899 with XGboost in the study with 220,657 Mexican patients, using demographic, clinical, symptom, and comorbidity data, also outperforming *Random Forest*. Thus, like Kar et al. (2021), where XGBoost outperformed *Random Forest* and *Logistic Regression* in a study with 2,370 Indian patients with clinical and laboratory data. It is worth noting that all the studies cited with XGBoost had as their central objective the creation and comparison of predictive death models.

Based on the Gini indices of the Random Forest model, it was found that the most important metrics for predicting the models in the analyzed data were the attributes SIND_DOWN (Has Down syndrome), HEPATICA (Has liver disease) and SATURAÇÃO (Saturation below 95%), SUPORT_VEN (Ventilation support), DIAS_UTI (Number of days in the ICU), NU_IDADE_N (Patient age), SG_UF_NOT (Notification state) and UTI (ICU admission). These variables play a crucial role in the model's decision, indicating that the need for mechanical ventilation, hospitalization and time in the ICU, and the patient's age are the most determining factors.

Finally, it is noteworthy that when models are made available through a software application that can be used in the hospital environment, this knowledge tends to be more widespread and actually used, not being restricted to literature alone. Thus, given the need to apply theory in practice, an easy-to-use software application prototype was developed so that health professionals could use predictive models in a hospital environment.

Regarding the limitations of this study, the following stand out: the difficulty in generalizing the use of the models for other population groups, such as the elderly, since the models were trained to classify specific groups; the imbalance identified between the death and cure classes, with a much higher number of cured patients than deceased, which may affect the ability of the models to correctly predict the minority class (death), leading to a tendency to overestimate the classification of the majority class (cure); the lack of acceptance tests of the application prototype by health professionals, since the successful implementation of a new technology in the clinical environment can be influenced by a series of factors such as usability and integration with existing systems; and finally, the fact that other ML techniques were not considered for a more comprehensive comparison, even though the study used the algorithms most commonly used in studies of this type.

CONCLUSION

The study provided death prediction models based on the SARS databases of the Brazilian Ministry of Health for children in the northern region of Brazil, as well as software for using these models to assist health professionals in the early identification of severe cases of SARS. The knowledge generated is considered to have the potential to provide health agents with prior knowledge about the prognosis of more severe patients and thus better allocate human and/or material resources for their treatment. This more effective allocation of resources is important in low- and middle-income regions, where these resources are scarce and periodically record an increase in the rates of SARS cases, such as during the period of fires in the northern region.

BIBLIOGRAPHICAL REFERENCES

ALTINI, N., BRUNETTI, A., MAZZOLENI, S., MONCELLI, F., ZAGARIA, et al. **Predictive Machine Learning Models and Survival Analysis for COVID-19 Prognosis Based on Hematochemical Parameters**. *Sensors (Basel, Switzerland)*, *21*(24), 2021. Disponível em: https://doi.org/10.3390/s21248503

AN, C., LIM, H., KIM, D. W., CHANG, J. H., CHOI, Y. J., et al. **Machine learning prediction for mortality of patients diagnosed with COVID-19: a nationwide Korean cohort study**. *Scientific reports*, *10*(1), 2020. Disponível em: https://doi.org/10.1038/s41598-020-75767-2

AZNAR-GIMENO, R., ESTEBAN, L. M., LABATA-LEZAUN, G., DEL-HOYO-ALONSO, R., ABADIA-GALLEGO, D., et al. **A Clinical Decision Web to Predict ICU Admission or Death for Patients Hospitalised with COVID-19 Using Machine Learning Algorithms.** *International journal of environmental research and public health*, *18*(16), 2021. Disponível em: https://doi.org/10.3390/ijerph18168677

ARAÚJO, D. C., VELOSO, A. A., BORGES, K. B. G., CARVALHO, M. D. G**. Prognosing the risk of COVID-19 death through a machine learning-based routine blood panel: A retrospective study in Brazil**. *International journal of medical informatics*. *165*, 104835, 2022. Disponível em: https://doi.org/10.1016/j.ijmedinf.2022.104835

BÁRCENAS, R., FUENTES-GARCÍA, R. **Risk assessment in COVID-19 patients: A multiclass classification approach**. *Informatics in medicine unlocked*, *32*, 101023, 2022. Disponível em: https://doi.org/10.1016/j.imu.2022.101023

BENNETT, T. D., MOFFITT, R. A., HAJAGOS, J. G., AMOR, B., ANAND, A., et al. **National COVID Cohort Collaborative (N3C) Consortium (2021). Clinical Characterization and Prediction of Clinical Severity of SARS-CoV-2 Infection Among US Adults Using Data from the US National COVID Cohort Collaborative**. *JAMA network open*, *4*(7), e2116901, 2021. Disponível em: https://doi.org/10.1001/jamanetworkopen.2021.16901

BEZERRA, J. H. S., ALMEIDA, F. M. **DESENVOLVIMENTO DE MODELOS PREDITIVOS COM MACHINE LEARNING - ANÁLISE DE DADOS PARA SAÚDE DE GESTANTES E PUÉRPERAS**. InterSciencePlace, 19, 2024. Disponível em: https://www.intersciencepIace.org/index.php/isp/article/view/763

BOOTH, A. L., ABELS, E., MCCAFFREY, P. **Development of a prognostic model for mortality in COVID-19 infection using machine learning**. *Modern pathology: an official journal of the United States and Canadian Academy of Pathology, Inc*, *34*(3), 522–531, 2021. Disponível em: https://doi.org/10.1038/s41379-020-00700-x

BOTTRIGHI, A., PENNISI, M., ROVETA, A., MASSARINO, C., CASSINARI, A., et al. **A machine learning approach for predicting high risk hospitalized patients with COVID-19 SARS-Cov-2**. *BMC medical informatics and decision making*, *22*(1), 340, 2022. Disponível em: https://doi.org/10.1186/s12911-022-02076-1

BRASIL. **Ministério da Saúde. SRAG 2021 a 2024: banco de dados de Síndrome Respiratória Aguda Grave**. OpenDataSUS, 2024. Disponível em: https://opendatasus.saude.gov.br/dataset/srag-2021-a-2024

CARVALHO, A. L. C. **Aplicação de técnicas de aprendizagem de máquina na geração de índices para sistemas de busca**. 2012. 101 f. Tese (Doutorado em Informática) - Universidade Federal do Amazonas, Manaus, 2012. Disponível em: https://tede.ufam.edu.br/handle/tede/4517

CHAPMAN, P., KHABAZZA, T., SHEARER, C. **CRISP-DM 1.0: step by step data mining guide.** SPSS, 2000**.** Disponível em: https://www.kde.cs.uni-kassel.de/wp-content/uploads/lehre/ws2012-13/kdd/files/CRISPWP-0800.pdf

CHEN, T., GUESTRIN, C. **XGBoost: A Scalable Tree Boosting System.** *In Proceedings of the* 22nd ACM SIG *KDD International Conference on Knowledge*

Discovery and Data Mining (KDD '16). *Association for Computing Machinery*, New York, NY, USA, 785–794, 2016. Disponível em: https://doi.org/10.1145/2939672.2939785

DEBNATH, S., BARNABY, D. P., COPPA, K., MAKHNEVICH, A., KIM, E. J., et al. **Machine learning to assist clinical decision-making during the COVID-19 pandemic**. *Bioelectronic Medicine*, v. 6, p. 14, 2020. Disponível em: https://doi.org/10.1186/s42234-020-00050-8

FERNANDES, F. T., DE OLIVEIRA, T. A., TEIXEIRA, C. E., BATISTA, A. F. M., DALLA COSTA, G., et al. **A multipurpose machine learning approach to predict COVID-19 negative prognosis in São Paulo, Brazil**. *Scientific reports*, *11*(1), 3343, 2021. Disponível em: https://doi.org/10.1038/s41598-021-82885-y

GAO, Y., CAI, G. Y., FANG, W., LI, H. Y., WANG, S. Y., et al. **Machine learning based early warning system enables accurate mortality risk prediction for COVID-19**. *Nature communications*, *11*(1), 5033, 2020. Disponível em: https://doi.org/10.1038/s41467-020-18684-2

GROSSARTH, S., MOSLEY, D., MADDEN, C., IKE, J., SMITH, I., et al. **Recent Advances in Melanoma Diagnosis and Prognosis Using Machine Learning Methods**. Current Oncology Reports, v. 25, p. 635–645, 2023. Disponível em: https://doi.org/10.1007/s11912-023-01407-3

HU, C., LIU, Z., JIANG, Y., SHI, O., ZHANG, X., et al. **Early prediction of mortality risk among patients with severe COVID-19, using machine learning.** *International journal of epidemiology*, *49*(6), 1918–1929, 2021. Disponível em: https://doi.org/10.1093/ije/dyaa171

HENKE, M., SANTOS, C., NUNAN, E., FEITOSA, E., SANTOS, E., et al. **Aprendizagem de Máquina para Segurança em Redes de Computadores: Métodos e Aplicações**. In: XXVI Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (SBRC). Anais... Manaus: Universidade Federal do Amazonas, 2018. p. 53-74. Disponível em: https://books-sol.sbc.org.br/index.php/sbc/catalog/download/95/419/690?inline=1

HELDT, F. S., VIZCAYCHIPI, M. P., PEACOCK, S., CINELLI, M., MCLACHLAN, L., et al. **Early risk assessment for COVID-19 patients from emergency department data using machine learning**. *Scientific reports*, *11*(1), 4200, 2021. Disponível em: https://doi.org/10.1038/s41598-021-83784-y

HOSMER, D. W., LEMESHOW, S., STURDIVANT, R. X. **Applied Logistic Regression**. 3. ed. Wiley, 2013.

KAR, S., CHAWLA, R., HARANATH, S. P., RAMASUBBAN, S., RAMAKRISHNAN, N., et al. **Multivariable mortality risk prediction using machine learning for COVID-**

19 patients at admission (AICOVID). *Scientific reports*, *11*(1), 12801, 2021. Disponível em: https://doi.org/10.1038/s41598-021-92146-7

KIVRAK, M., GULDOGAN, E., COLAK, C. **Prediction of death status on the course of treatment in SARS-COV-2 patients with deep learning and machine learning methods**. *Computer methods and programs in biomedicine*, *201*, 105951, 2021. Disponível em: https://doi.org/10.1016/j.cmpb.2021.105951

KUMARAN, M., PHAM, T. M., WANG, K., USMAN, H., NORRIS, C. M., et al. **Predicting the Risk Factors Associated with Severe Outcomes Among COVID-19 Patients-Decision Tree Modeling Approach**. *Frontiers in public health*, *10*, 838514, 2022. Disponível em: https://doi.org/10.3389/fpubh.2022.838514

LEE, C. H., BANOEI, M. M., ANSARI, M., et al. **Using a targeted metabolomics approach to explore differences in ARDS associated with COVID-19 compared to ARDS caused by H1N1 influenza and bacterial pneumonia**. Crit Care., v. 28, p. 63, 2024. doi: 10.1186/s13054-024-04843-0.

LI, Y., HOROWITZ, M. A., LIU, J., CHEW, A., LAN, H., et al. **Individual-Level Fatality Prediction of COVID-19 Patients Using AI Methods**. *Frontiers in public health*, *8*, 587937, 2020. Disponível em: https://doi.org/10.3389/fpubh.2020.587937

LI, J., LI, X., HUTCHINSON, J., ASAD, M., LIU, Y., et al. **An ensemble prediction model for COVID-19 mortality risk**. *Biology methods & protocols*, *7*(1), bpac029, 2022. Disponível em: https://doi.org/10.1093/biomethods/bpac029

LIMA, T. P. F., SENA, G. R., NEVES, C. S., VIDAL, S. A., LIMA, J. T. O., et al. **Death risk and the importance of clinical features in elderly people with COVID-19 using the Random Forest Algorithm**. Revista Brasileira de Saúde Materno Infantil, 21(suppl 2), 445–451, 2021. Disponível em: https://doi.org/10.1590/1806-9304202100s200007

LOPES, M. A. **Aplicação de aprendizado de máquina na detecção de fraudes públicas**. 2019. Dissertação (Mestrado em Administração) - Faculdade de Economia, Administração e Contabilidade, Universidade de São Paulo, São Paulo, 2019. Disponível em: https://doi.org/10.11606/D.12.2020.tde-10022020-174317

MLADENOVA, T., VALOVA, I. **Classification with K-Nearest Neighbors Algorithm: Comparative Analysis between the Manual and Automatic Methods for K-Selection**. *International Journal of Advanced Computer Science and Applications* (IJACSA), 14(4), 2023. Disponível em: http://dx.doi.org/10.14569/IJACSA.2023.0140444

MAHDAVI, M., CHOUBDAR, H., ZABEH, E., RIEDER, M., SAFAVI-NAEINI, S., et al. **A machine learning based exploration of COVID-19 mortality risk.** *PloS one*, *16*(7), e0252384, 2021. Disponível em: https://doi.org/10.1371/journal.pone.0252384

MOITINHO, L. C. C., BENICASA, A. X. **Aprendizado de Máquina para o Auxílio à Localização de Pessoas em Ambientes Indoor Monitorados por Câmeras**. In: Concurso de trabalhos de conclusão de curso em sistemas de informação - simpósio brasileiro de sistemas de informação (SBSI), 19, 2023, Maceió/AL. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2023. p. 71-80. Disponível em: https://doi.org/10.5753/sbsi_estendido.2023.229347

MOSLEHI, S., MAHJUB, H., FARHADIAN, M., SOLTANIAN, A. R., MAMANI, M. **Interpretable generalized neural additive models for mortality prediction of COVID-19 hospitalized patients in Hamadan, Iran**. *BMC medical research methodology*, *22*(1), 339, 2022. Disponível em: https://doi.org/10.1186/s12874-022-01827-y

MOULAEI, K., SHANBEHZADEH, M., MOHAMMADI-TAGHIABAD, Z., KAZEMI-ARPANAHI, H. **Comparing machine learning algorithms for predicting COVID-19 mortality**. *BMC Med Inform Decis Mak* 22, 2., 2022. Disponível em: https://doi.org/10.1186/s12911-021-01742-0

MURRI, R., LENKOWICZ, J., MASCIOCCHI, C., IACOMINI, C., FANTONI, M., et al. **A machine-learning parsimonious multivariable predictive model of mortality risk in patients with Covid-19**. *Scientific reports*, *11*(1), 21136, 2021. Disponível em: https://doi.org/10.1038/s41598-021-99905-6

PAIXÃO, G. M. M., SANTOS, B. C., ARAÚJO, R. M., RIBEIRO, M.H., MORAES J. L., RIBEIRO, A. L. **Machine Learning in Medicine: Review and Applicability**. Arq Bras Cardiol. Jan;118(1):95-102, 2022. Disponível em: https://doi.org/10.36660/abc.20200596

POLO, T. C. F., MIOT, H. A. **Aplicações da curva ROC em estudos clínicos e experimentais**. *J Vasc Bras*. 19:e20200186, 2020. Disponível em: https://doi.org/10.1590/1677-5449.200186

REINA, A. R., BARRERA, J. M., VALDIVIESO, B., GAS, M. E., MATÉ, A., et al. **Machine learning model from a Spanish cohort for prediction of SARS-COV-2 mortality risk and critical patients**. *Scientific reports*, *12*(1), 5723, 2022. Disponível em: https://doi.org/10.1038/s41598-022-09613-y

RYBCZAK, M., POPOWNIAK N., LAZAROWSKA A. **A Survey of Machine Learning Approaches for Mobile Robot Control**. *Robotics*. 2024; 13(1):12. Disponível em: https://doi.org/10.3390/robotics13010012

SCHÖNING, V., LIAKONI, E., BAUMGARTNER, C., EXADAKTYLOS, A. K., HAUTZ, W. E., et al. **Development and validation of a prognostic COVID-19 severity assessment (COSA) score and machine learning models for patient triage at a tertiary hospital.** *Journal of translational medicine*, *19*(1), 56, 2021. Disponível em: https://doi.org/10.1186/s12967-021-02720-w

SCHOBER, P., VETTER, T. R. **Logistic Regression in Medical Research**. *Anesthesia and analgesia*, 132(2), 365–366, 2021. Disponível em: https://doi.org/10.1213/ANE.0000000000005247

SENA, G. R. **Modelos Preditivos de Óbito para Pacientes com COVID-19**. Tese de doutorado apresentada ao Instituto de Medicina Integral Prof. Fernando Figueira (IMIP), 2021. Disponível em: http://higia.imip.org.br/handle/123456789/641?mode=full

SILVA, E. A. D. **Algoritmo genético assistido por surrogate para avaliar e descobrir peptídeos contra o SARS-CoV-2**. 2022. 79 f. Dissertação (Mestrado em Ciência da Computação) - Universidade Federal de Uberlândia, Uberlândia, 2022. Disponível em: http://doi.org/10.14393/ufu.di.2022.571.

SILVA, R., SILVA NETO, D. R. DA. **Inteligência artificial e previsão de óbito por Covid-19 no Brasil: uma análise comparativa entre os algoritmos Logistic Regression, Decision Tree e Random Forest**. Saúde em Debate, 46(spe8), 118–129, 2022. Disponível em: https://doi.org/10.1590/0103-11042022E809

SUN, C., HONG, S., SONG, M., LI, H., WANG, Z. **Predicting COVID-19 disease progression and patient outcomes based on temporal deep learning**. *BMC medical informatics and decision making*, *21*(1), 45, 2021. Disponível em: https://doi.org/10.1186/s12911-020-01359-9

VAN DER SCHAAR, M., ALAA, A. M., FLOTO, A., GIMSON, A., SCHOLTES, S., et al**. How artificial intelligence and machine learning can help healthcare systems respond to COVID-19**. *Mach Learn*, v. 110, p. 1–14, 2021. Disponível em: https://doi.org/10.1007/s10994-020-05928-x

VEPA, A., SALEEM, A., RAKHSHAN, K., DANESHKHAH, A., SEDIGHI, T., et al. **Using Machine Learning Algorithms to Develop a Clinical Decision-Making Tool for COVID-19 Inpatients.** *International journal of environmental research and public health*, *18*(12), 6228, 2021. Disponível em: https://doi.org/10.3390/ijerph18126228

WANG, Y. PAN, Z., ZHENG, J., QIAN, L., MINGTAO, Li. **A hybrid ensemble method for pulsar candidate classification**. *Astrophysics and Space Science*. 364. 2019. Disponível em: https://doi.org/10.1007/s10509-019-3602-4

WOO, S. H., RIOS-DIAZ, A. J., KUBEY, A. A., CHENEY-PETERS, D. R., ACKERMANN, L. L., et al. **Development and Validation of a Web-Based Severe COVID-19 Risk Prediction Model.** *The American journal of the medical sciences*, *362*(4), 355–362, 2021. Disponível em: https://doi.org/10.1016/j.amjms.2021.04.001

YADAW, A. S., LI, Y. C., BOSE, S., IYENGAR, R., BUNYAVANICH, S., et al. **Clinical features of COVID-19 mortality: development and validation of a clinical prediction model**. *The Lancet. Digital health*, *2*(10), e516–e525, 2020. Disponível em: https://doi.org/10.1016/S2589-7500(20)30217-X

YU, L., HALALAU, A., DALAL, B., ABBAS, A. E., IVASCU, F., et al. **Machine learning methods to predict mechanical ventilation and mortality in patients with COVID-19**. *PloS one*, *16*(4), e0249285, 2021. Disponível em: https://doi.org/10.1371/journal.pone.0249285

ZAREI, J., JAMSHIDNEZHAD, A., SHOUSHTARI, H. M., HADIANFARD, M. A., CHERAGHI, M., et al. **Machine Learning Models to Predict In-Hospital Mortality among Inpatients with COVID-19: Underestimation and Overestimation Bias Analysis in Subgroup Populations**. *Journal of healthcare engineering*, 1644910, 2022. Disponível em: https://doi.org/10.1155/2022/1644910

ZHAO, Y., ZHANG, R., ZHONG, Y., WANG, J., WENG, Z., et al. **Statistical Analysis and Machine Learning Prediction of Disease Outcomes for COVID-19 and Pneumonia Patients.** *Frontiers in cellular and infection microbiology*, *12*, 838749, 2022. Disponível em: https://doi.org/10.3389/fcimb.2022.838749