Brazilian Journal of Biometrics, **41** (1), 18–29 (2023)
doi:10.28951/bjb.v41i1.588

ARTICLE

Classification and Analysis of Patients with COVID-19 Using Machine Learning[1]

Glaucia Maria Bressan*[1] , Elisângela Ap. da Silva Lizzi[1]

[1]Department of Mathematics Universidade Tecnológica Federal do Paraná, Cornélio Procópio-PR, Brazil
*Corresponding author. Email: glauciabressan@utfpr.edu.br

(Received: October 20,2021; Revised: April 18,2022; Accepted: July 15,2022; Published: March 1,2023)

Abstract
The rapid spread of the Coronavirus disease (COVID-19) has demanded studies and research works from many areas of knowledge, searching for treatments, vaccines and preventive measures. This pandemic has become a very challenging situation due to its substantial demand for medical infrastructure. In this context, this paper proposes to apply Machine Learning methods to classify and to analyse the outcome of patients with COVID-19 as discharge or death and to describe the profile of patients infected by the coronavirus. The dataset consists of clinical data from Sírio Libanês Hospital, available in the FAPESP repository (2020). Results indicate that, among all tested classifiers, the Naive Bayes algorithm presents better performance and it better represents the phenomenon under study, demonstrating superiority in terms of classification and induction numerical analysis of the epidemiological phenomenon for COVID-19.

Keywords: classification; Machine Learning; COVID-19; patients profile; statistical applied.

1. Introduction

The Coronavirus disease (COVID-19) pandemic officially began in 2019 on the Asian continent, in China, and it has spread widely through the world in a short period of time (Rodríguez-Morales *et al.*, 2020). In Brazil, the Ministry of Health confirmed the first case of coronavirus in São Paulo city, in February 26, 2020. In March 2020, the World Health Organization (WHO) elevated the state of COVID-19 contamination to a pandemic. As of September 02, 2021, infected people worldwide were responsible for more than 218 million cases and more than 4,5 million deaths. In Brazil, there are almost 20 million infected and 581,000 deaths. The Brazil came to occupy the third place in number of infected and dead in the world during the beginning of the year 2021, which corresponds to at least 12.9% of deaths by COVID-19 in the world. The mortality rate of COVID-19 in Brazil is around 2.8% (Dong *et al.*, 2020). With the advance of vaccination in the country, the scenario has improved.

News related to COVID-19 is commented on daily in the media and has caused great concern and impact on Global Public Health. The vaccination of the Brazilian population from priority groups began in January 2021, and there are already 259 million doses applied, among the population that received at least the first dose of the immunizing agent. In relation to those immunized with the second dose, there are 107 million (50.4% of the population).

The symptoms presented by those infected with COVID-19 are common to other respiratory infections and may include fever, cough, sore throat, headache, fatigue, muscle pain, loss of smell and shortness of breath. The clinical condition is diverse, ranging from an asymptomatic condition to acute respiratory syndrome and damage to various systems of the body, and it is possible to observe an increase in inflammatory markers, cardiovascular changes and injuries to the lungs and kidneys (Ministry of Health, 2022).

The overwhelmed hospitals are a direct consequence of rapidly increasing coronavirus cases (Nemati *et al.*, 2020). A characterization of the first 250.000 hospital admissions for COVID-19 in Brazil is presented in Ranzani *et al.* (2021). Authors did a retrospective analysis of patients aged 20 years or older with quantitative RT-PCR confirmed COVID-19 who were admitted to hospital and registered in a nationwide surveillance database in Brazil, between February and December, 2020. Data were analysed from the whole country and its five regions.

The spread of COVID-19 pandemic has motivated the development of studies and researches from all areas of knowledge, searching for treatments, vaccines and preventive measures. Works have been conducted for better understanding the origin and proper treatments of this novel coronavirus (Acter *et al.*, 2020). Epidemic models have been proposed in the literature to describe the dynamics of the COVID-19 pandemic. Each modelling approach seeks to describe a set of variables of interest according to specific objectives. In Grzybowski *et al.* (2020), authors present a case study with three prognostic scenarios for the first wave of the pandemic in the city of Manaus, Brazil. Results show that there are feasible control strategies that could substantially reduce the overload within reasonable time.

Machine Learning methods have been used to address a COVID-19 pandemic in several aspects. Diagnosis and the prediction of virus transmission can be analysed by Machine Learning algorithms, simulations, and digital monitoring (Allam *et al.*, 2020). In the case of classification, segmentation is based on a set of training data, which encodes knowledge about the structure of the groups in the form of a target variable (outcome). In this study the binary outcome is discharge or death by COVID-19. Then, as a result, the classification algorithms used are supervised learning, where a target variable is known and it studies how the input variables assist in this classification (Aggarwal, 2014).

In Lalmuanawma Hussain and Chhakchhuak (2020), authors address recent studies that apply Machine Learning and Artificial Intelligence technology towards augmenting the researchers on multiple angles. It also discusses suggestions conveying researchers on model design, medical experts, and policymakers in the current situation while tackling the COVID-19 pandemic and ahead.

Ahmad *et al.* (2020) provide suggestions to the Machine Learning practitioners to improve the performance of Machine Learning methods for the prediction of confirmed cases of COVID-19 and presents a detailed review of research papers that used Machine Learning to do this prediction. In Nemati *et al.* (2020), by choosing patient discharge time as the event of interest, survival analysis techniques and Machine Learning are used to build predictive models capable of predicting patients' period of stay in hospital, which allows decision makers to be prepared for hospital overloads.

A novel Support Vector Regression method is proposed in Yadav *et al.* (2020), to

analyse different tasks related to novel coronavirus, such as predicting the spread of coronavirus across regions and analysing the transmission and the growth rate of the virus. The approach is evaluated and compared with other well-known regression models on standard available datasets. A mortality risk prediction model for COVID-19 is presented in Gao *et al.* (2020), that uses patients' clinical data on admission to stratify patients by mortality risk, which enables prediction of physiological deterioration and death up to 20 days in advance. The model uses Logistic Regression, Support Vector Machine, Gradient Boosted Decision Tree, and Neural Network and it enables expeditious and accurate mortality risk stratification of patients with COVID-19.

More recently, the study of Hou *et al.* (2021) identifies key independent clinical parameters that predict intensive care unit (ICU) admission and mortality associated with COVID-19 infection. A machine learning algorithm identifies key clinical measures to triage patients more effectively to general admission versus ICU admission and to predict mortality in COVID-19 pandemic. Authors identified the top few variables amongst the large array of clinical variables that were most predictive of the likelihood of ICU admission and mortality.

In the face of the pandemic scenario that plagues the planet currently, the objective of this paper is to apply Machine Learning algorithms to classify the outcome of patients with COVID-19 in a Hospital located in São Paulo city, Brazil. This hospital is private and usually serves patients from all over the country and who have high income, representing a layer of the population with high purchasing power. This paper contributes to classify and to analyse the outcome of patients with COVID-19 as discharge or death and to describe the profile of patients infected by the coronavirus.

The remainder of the paper is organized as follows. Section 2 presents the epidemiological design, the dataset description and the classification methods. In section 3, the results of the classification proposed in this paper are presented and discussed. Finally, in Section 4, the conclusion is presented.

2. Materials and Methods

In this section, the dataset pre-processing is presented, as well as the Machine Learning methods used in this paper to classify the outcome of patients with COVID-19.

2.1 Dataset pre-processing steps

The epidemiological design of this research is classified as descriptive and retrospective epidemiological study with primary data from patients admitted to health units and hospitals.

There are several steps in the process that currently limit the application of machine learning to combat COVID-19 (Alimadadi *et al.*, 2020), such as the linkage of datasets and the interoperability of services. The availability of COVID-19 clinical data, which can be managed and processed in easily accessible databases, is an important current barrier. The data are available in the FAPESP repository (Mello *et al.*, 2020) and this study used information available from Sírio Libanês Hospital, from February to December, 2020. The dataset was structured in electronic spreadsheets and some pre-processing steps were performed, which consists of 4 phases, in order to generate a valid analytical database:

1. Merge among the available datasets in order to make it possible to link information by identifying the unique patient code;
2. Cleaning of duplicate information, with typos and missing information;

3. Organization of the database using two outcomes (discharge or death) and input variables, being: sex, age, length of stay, type of care (outpatient, emergency care, external and internal) and federative unit to which the patient belongs.

4. The final database has 3902 instances with complete patients' information.

After this processing step, all variables were coded numerically, including the qualitative variables, in order to minimize bias in our algorithms due to false interpretation of string variables. In all these stages, computational software support, the R (2020), was used. Therefore, the dataset is organized using numbers to represent the variables.

2.2 Classification methods

In this paper, the input features are the variables described in above and the output variable, which is the outcomes, consists of 2 classes: discharge or death. The Machine Learning methods used in this paper to classify this dataset are briefly described as follows.

The problem of data classification attempts to learn the relationship between a set of feature variables and a target variable of interest, named class. According to Aggarwal (2014), the problem of classification may be stated as: given a set of training data points along with associated training labels, determine the class label for an unlabelled test instance. Then, the classification task consists of two phases. In the training phase, a model is constructed from the training instances and in the testing phase, the model is used to assign a label to an unlabelled test instance.

2.2.1 Support Vectors Machine (SVM)

The SVM algorithm uses a nonlinear mapping to project the original training data into a higher dimension. Within this new dimension, a linear optimal hyperplane capable of separating the data into two classes is computed. According to Han, Kamber and Pei (2012), it is always possible to obtain such hyperplane with an appropriate nonlinear mapping to a sufficiently high dimension, and the hyperplane is obtained by using the so-called support vector and margins (defined by the support vectors). There may be infinitely many hyperplanes that separate the positive and negative instances correctly. A reasonable choice is the one with the largest gap between both classes. This setting, called the Maximum Margin Classifier, may be more resistant to any perturbation of the training data (Aggarwal, 2014).

Although the training time of even the fastest SVMs can be slow, they are highly accurate, due to their ability to model complex nonlinear decision boundaries (Han *et al.*, 2012).

SVM is an approach for controlling model complexity. It chooses important instances to construct the separating surface between data instances. When the data is not linearly separable, it can either penalize violations with loss terms, or leverage kernel tricks to construct nonlinear separating surfaces (Aggarwal, 2014). SVMs can also perform multiclass classifications in various ways, either by an ensemble of binary classifiers or by extending margin concepts.

2.2.2 K- Nearest Neighbours (KNN)

KNN is a supervised machine learning algorithm that can be used to solve classifications problems. The KNN classification approach consists in fixing the number k of samples and letting the width change so that each region contains exactly k samples boundary (Dougherty, 2012). The KNN process starts at the test point and expands a region until it encloses k training samples, labelling the test point x by a majority vote of these samples. If the majority of samples closest to the unknown sample are from a specified class, the sample will be

assigned to that class.

For two classes, the value of k should be odd to avoid a tie, and larger values are more likely to resolve ties. In fact, the larger the value of k, the smoother will be the classification boundary, and smaller values for k results on a more convoluted boundary (Dougherty, 2012).

There is essentially no training involved in the KNN method, being thus considered a lazy learning algorithm. In general, the KNN defers data processing until it receives a request to classify an unlabelled (test) example, classifies it, and then discards any intermediate results. The main KNN advantages are that it is intuitive, analytically tractable and simple to implement boundary (Dougherty, 2012).

2.2.3 Naïve Bayes Networks

Bayesian networks can be defined as a probabilistic graphical model used to represent knowledge about the data domain. These networks learn cause and consequence relationships and can combine a priori knowledge with patterns learned from the data. In addition, the user can interfere in the nodes of the network and insert a knowledge that propagates in the other nodes. The networks are then composed of a structure consisting of a directed acyclic graph and a set of probability tables. The nodes of the network structure represent the variables and the arcs between nodes represent dependency relations between the corresponding variables. When the node labelled as class has arcs directed for each input feature, this network structure represents a Naive-Bayes, which consists of a classifier. The Naive Bayes classifier is based on the Bayes' theorem, and is particularly suited when the dimensionality of the inputs is high (Aggarwal, 2014).

2.2.4 Decision Trees and Random Forest

Decision tree is a method very useful in data mining to extract information from a dataset. The extract information process is done by a TDIDT (Top-Down Induction Decision Tree), which induces a tree structure by splitting data into subgroups more and more uniforms based on "divide and conquer method (Maimon & Rokach, 2014)".

This split process stops when the subset contains just one class or when no more improvement is possible.

Under the root node are the internal nodes, originating from the division of the data set; they constitute the tree branches. At the end of each branch is the terminal node, designed leaves, which represent the most appropriated class for the rule.

Decision tree algorithms construct the tree in two phases: growing and pruning. The growing is a recursive process that establishes the structure of the decision tree according to some splitting criterion, which partitions the domain of one (or more) of the data's attributes, such Information Gain and Gain Ratio. The element with highest Gain Ratio is taken as the root node and data set is split based on the root element values (Han *et al.*, 2012). When the number of instances to be split is below a certain threshold, the splitting stopes. The Information Gain is calculated for all the sub-nodes and the process is repeated until the prediction is completed. Information Gain is an impurity-based criteria that uses entropy measure as the impurity measure (Maimon & Rokach, 2014). To classify a data item, the data item must traverse the tree, beginning at the root. Based on the splitting rule, the data item is sent forward to one of the node's children. This testing and forwarding is repeated until the data item reaches a leaf node (Aggarwal, 2014).

Decision trees are nonparametric in the statistical sense: they are not modelled on a probability distribution for which parameters must be learned. Moreover, decision tree induction is almost always nonparametric in the algorithmic sense: there are no weight parameters which affect the results. Each path from root to leaf generates one rule as follows:

form the conjunction (logical AND) of all the decisions from parent to child.

Random Forest, in turn, is an ensemble learning method for classification and consists of one of the most popular machine learning methods (Breiman, 2001). A Random Forest is composed of Decision Trees, where each Decision Tree is considered as an element of this ensemble, denominated forest. Ensemble classification methods train several classifiers and combine the decision of a set of classifiers by weighted or unweighted voting process to classify unknown examples (Aggarwal, 2014; Breiman, 2001).

An ensemble classifier is generally found to be more accurate than any of the individual classifiers making up the ensemble (Aggarwal, 2014). In order to grow ensembles, often random vectors are generated that govern the growth of each tree in the ensemble. Random forests consider many fewer attributes for each split, for this reason, it can efficiently handle extremely large datasets.

In this paper, all of these considered classification algorithms and methods are implemented using the support of the R (2020) software, specifically the *e1071, rpart, randomforest* and *class* packages.

3. Results and Discussion

This section presents the classification results of patients with COVID-19, obtained from the machine learning methods considered in this paper, in order to classify the dataset into 2 classes: discharge or death by COVID-19. First, a data description is presented, which shows the input and output data used in the classification methods. Then, the results of the classifiers are described and discussed.

3.1 Data description

This section presents a general description of the data used in this paper, regarding the outcome (discharge or death) for the input variables, and the descriptive statistics and summary measures. The quantitative variables age (years) and hospital stay (days) are expressed in Figure 1 by the boxplot distribution graphic. For the age variable, there is greater variability and amplitude for the group who was discharged, and it is possible to perceive that on average these patients are 43.2 years old (Standard Deviation (SD) = 15 years), for the death group, the average age is 75.2 years (SD = 11 years) and there is a smaller and more concentrated variability in older ages. Regarding the variable length of stay in days, there is less variability in the group that was discharged and with an average of 1.6 days (SD = 5.6 days), whereas for the death group, these patients on average spent more time in the hospital with an average of 29.4 days (SD = 29.7), indicating high variability in length of stay in days and with a predominance of longer length of stay linked to death. Therefore, older patients who had a longer hospital stay are predominant in the death outcome.

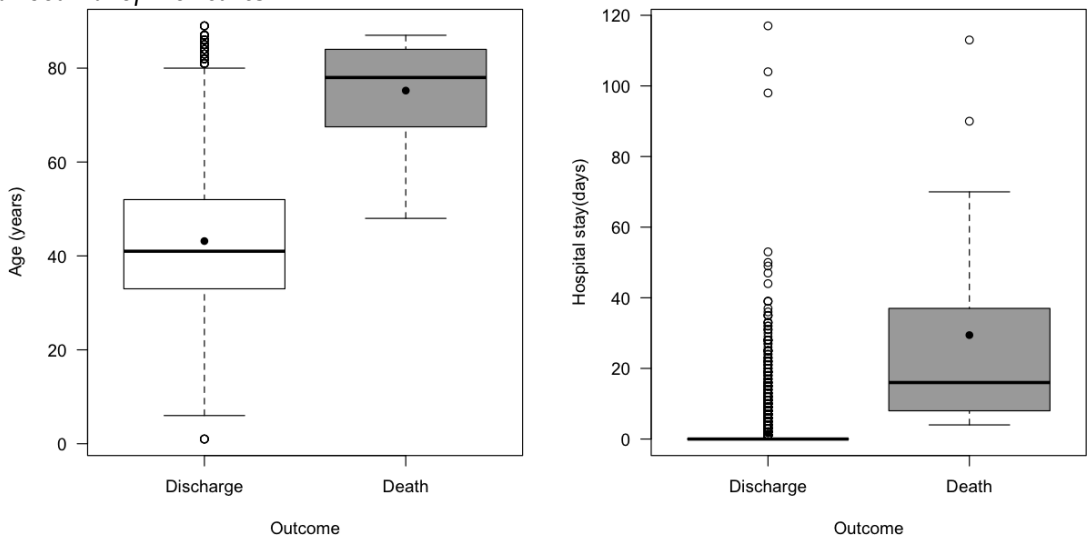

Figure 1. Boxplot aging and hospital stay by outcome (discharge or death)

Table 1 characterizes the input variables: sex, hospital stay (cutoff points), service type and Federative units. It is possible to perceive a balanced distribution in relation to sex, as for the hospital stay (days) there is an inversely proportional distribution between discharge and death, as high as the number of days' increases, the percentage decreases and for death it is the opposite. For the variable Federative units, the highest prevalence is in São Paulo, it is worth mentioning that the data refer to the Sírio Libanês hospital and, with that, can justify this higher numerical quantity.

Table 1. Characterization of categorical variables in relation to the outcome (discharge or death)

		Outcome	
Input Variable		Discharge n(%)	Death n(%)
Sex	Female	1969 (51%)	10 (43%)
	Male	1910 (49%)	13 (57%)
Hospital stay (days)	0 day	2920 (75%)	0 (0%)
	1-7 days	683 (18%)	5 (22%)
	8-14 days	159 (4%)	5 (22%)
	>15 days	117 (3%)	13 (57%)
Service type	Outpatient	1240 (32%)	0 (0%)
	External	1121 (29%)	0 (0%)
	Internal	416 (11%)	22 (96%)
	Emergency Service	1102 (28%)	1 (4%)
Federative Unit	SP=13	3276 (84%)	20 (87%)
	others (Few values dispersed in other units)	596 (16%)	3 (13%)

3.2 Classification results

This section shows the results of the machine learning methods used in this article to classify this data set into 2 classes: discharge or death by COVID-19.

Table 2 shows the levels of accuracy achieved by classifiers, followed by the respective confidence interval (95%) and the kappa coefficient. The kappa is a measure of how closely the

instances classified by the Machine Learning classifier matched the data. Since the dataset is unbalanced, F1-score and the Area Under the Curve (AUC) of the Receiver Operator Characteristic (ROC) curve are also presented. F1-score is a classification error metric used to evaluate the classification algorithms, which considers not only the number of prediction errors, but also the type of errors. Mostly, it is useful in evaluating the prediction for binary classification of data. AUC-ROC indicates how well the Machine Learning binary classifiers are performing. It is the measure of the ability of a classifier to distinguish between classes. The closer to 1 the value of F1-socore and AUC-ROC, the better the performance of the classifiers at distinguishing between output classes.

It is important to highlight that the evaluation of the accuracy through the global accuracy and the coefficients of agreement, provides results with a higher degree of confidentiality among the studied discrimination criteria (Steyerberg *et al.*, 2010).

The highest accuracy values were achieved using SVM and Random Forest methods, followed by KNN and Naïve Bayes methods. However, observing the Confidence Interval (CI), we notice that the values present an intersection between them. In addition, also considering the kappa coefficient, for a better interpretation and confidentiality of the results, the classifier that presents the greatest kappa coefficient (0.4161) is the Naive Bayes, which indicates a moderate agreement. Thus, this method obtained the best results in terms of class discrimination. Observing the F1-score and the AUC-ROC, as the values are close to 1, we can conclude that classifiers are performing well, i.e., the classifiers are distinguishing the two output classes.

Table 2. Measures to evaluate the performance of classifiers

Classifier	Accuracy	CI (95%)	Kappa	F1-score	AUC-ROC
SVM	0.9939	(0.9867, 0.9977)	0.0	0.9969	0.9850
KNN	0.9928	(0.9853, 0.9971)	0.0	0.9964	0.9750
Naive-Bayes	0.9887	(0.9799, 0.9944)	0.4161	0.9993	0.9800
Random Forest	0.9939	(0.9867, 0.9977)	0.0	0.9969	0.9700

Table 3 shows the results of the sensitivity and specificity of each algorithm. The term “not available” means there was no mathematical convergence and then, the numerical value could not be calculated. It is interesting to verify that the class with the highest sensitivity values is the discharge class, since these values are affected by the imbalance of the data sample. Therefore, the class with the highest specificity values is the death class, since it has a lower n. Paralleling the classification with the epidemiological outcome in terms of diagnostic tests, it is known that specificity is the capacity that the diagnostic test / screening has to detect the true negatives, that is, to correctly diagnose healthy individuals (Gordis, 2014). In this study, since the dataset is unbalanced, the algorithm is conservatively constructed to indicate a possible death outcome, as it is more specific in this class. And as the sensitivity is higher in the discharge class, it reflects how effective the algorithms are in correctly identifying, among all the individuals evaluated, those who actually present the discharge outcome.

As there is a tendency to zero and not to calculate the sensitivity and specificity for the death class in SVM, KNN and Random Forest methods, then the Naive Bayes method has shown again a superior performance and it is able to calculate these indicators, presenting superiority in terms of classification and induction numerical analysis of the epidemiological phenomenon for COVID-19 (Aggarwal, 2014; Steyerberg *et al.*, 2010).

Table 3. Sensitivity and Specificity of classes

Classifier	Sensitivity class: discharge	Specificity class: discharge	Sensitivity class: death	Specificity class: death
SVM	0.9939	Not available	Not available	0.9938
KINN	0.9938	0	0	0.9938
Naïve-Bayes	0.9979	0.3077	0.3077	0.9979
Random Forest	0.9939	Not available	Not available	0.9938

Analysing the confusion matrices illustrated in Figure 2 for the test data, after and optimizing the classifiers in the training data, it is possible to observe that the SVM, KNN and Random Forest algorithms present a tendency to classify all data in discharge class. In contrast, the Naive Bayes algorithm provides a better classification, as it managed to divide the classes satisfactorily, even with the sample imbalance that favours the discharge class. It can be said that the SVM, KNN and Random Forest algorithms have obtained biased estimates for the discharge class and are more likely to assign patients in this class. Naive Bayes, on the other hand, managed to adjust itself or obtain more consistent estimates, so it is necessary to weigh the accuracy in relation to the kappa coefficient, sensitivity and specificity. If we focus only on the accuracy, we could choose a classifier totally biased to describe the epidemiological phenomenon in these patients affected by COVID-19.

a) Confusion Matrix for SVM

Class	Discharge	Death
Discharge	970	0
Death	6	0

b) Confusion Matrix for KNN

Class	Discharge	Death
Discharge	969	1
Death	6	0

c) Confusion Matrix for Naive Bayes

Class	Discharge	Death
Discharge	961	9
Death	2	4

d) Confusion Matrix for Random Forest

Class	Discharge	Death
Discharge	970	0
Death	6	0

Figure 2. Confusion matrix of: a) SVM, b) KNN, c) Naive Bayes and d) Random Forest methods

According to Figure 3, using the Friedman hypothesis test (interpretation analogous to the variance test - ANOVA, however used when the assumptions of normality data are violated), that the Naive Bayes algorithm differs from all the others tested. Then, it is possible to infer, observing the described results, that this classifier proved to be robust to the patient's classification in relation to COVID-19. An advantage of this classifier is that the Gaussian probability distribution is used for numerical variables and to categorical variables it is used conditional probability, always respecting the real proportions obtained from the sample. So this advantage enabled

numerical interpretation of the phenomenon of COVID-19 and thus led to a good classification.

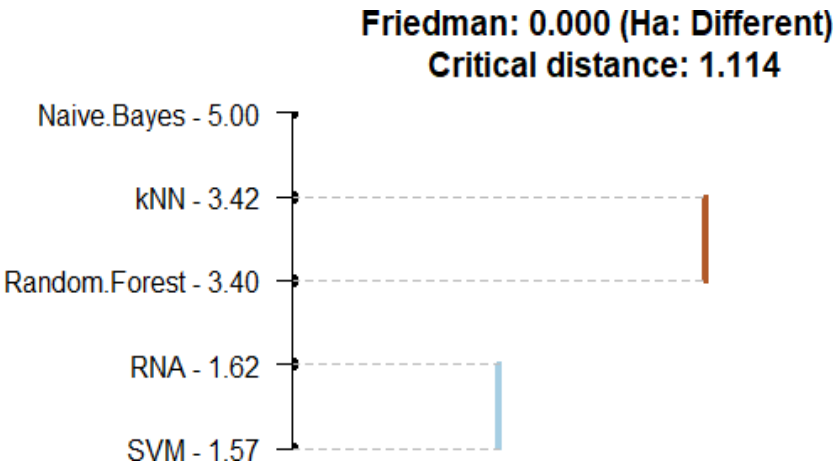

Figure 3. Friedman's hypothesis test chart to compare the algorithms

3.3 Discussion

The Naive Bayes classifier is an algorithm that presents low computational complexity and it is based on the Bayes Theorem (Box & Tiao, 1992). In epidemiology, statistical and numerical methods have come to be widely used with the evolution of evidence-based medicine (Vere & Gibson, 2019; Beckmann & Lew, 2016), technical and systematized information for clinical practice are addressed with widespread use of systematic reviews (Zeng *et al.*, 2020) and statistical and computational methods to support medical practice and treatment choices (Vere & Gibson, 2019).

In this paper, the results show that the domain of the Naive Bayes algorithm presents superior performance and ability to calculate accuracy indicators, such as sensitivity, specificity and kappa coefficient. Thus, this algorithm demonstrated superiority in terms of classification and numerical induction in the analysis of the epidemiological phenomenon for COVID-19. The algorithm enabled the numerical interpretation of the COVID-19 phenomenon and provided a good classification for both outcomes (discharge and death). Thus, Naive Bayes obtained the best results in terms of class discrimination, even with the unbalanced groups in the studied outcomes.

A limitation of this work is related to data and source of dataset. According to Alelyani, Liu and Wang (2011), the underlying characteristics of the data can greatly affect the stability of an algorithm. These characteristics include dimensionality, sample size and different distribution of data in different folds, and the stability problem tends to be data dependent. In addition, the data used in this work are from the FAPESP repository (2020), which means that the data were not obtained directly from an integrated health system, such as the hospital's electronic medical record.

4. Conclusion

This paper proposes the use and the implementation of methods for the classification and analysis of patients with COVID-19, using machine learning algorithms and real data from Sírio Libanês Hospital, available in FAPESP repository (2020). This work contributes to classify the evolution of patients with COVID-19 as discharge or death and to describe the profile of patients infected by the coronavirus.

It is important to mention that the development of features selection algorithms for classification with high accuracy and classification stability is still a challenge, even with computational advancement (Aggarwal, 2014). According to Alimadadi *et al.* (2020), the availability of clinical data related to COVID-19, which can be organized and processed in an open access database, is a barrier. This fact is true especially in Brazil, since the electronic medical record system is not a reality in all health units, showing a weakness of the system in subsidizing structured and unstructured data, in real time, about health problems.

About the applicability of the evaluated Machine Learning methods, we can indicate that the results obtained can be useful in situations of surveillance healthy, management clinical and public health system.

As perspectives for the continuity of this research, we propose the application of machine learning methods in an updated data set, following the evolution of the pandemic, and contemplating the economic profile of patients, considering public and private hospitals. We also propose the possibility of providing interventions aimed at clinical practice and also within the scope of public health policies, with comparisons between the modality of services to healthcare.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Acter, T., Uddin, N., DAS, J., Akhter, A., Choudhury, T.R., Kim, S. Evolution of severe acute respiratory syndrome coronavirus 2 (sars-cov-2) as coronavirus disease 2019 (covid-19) pandemic: a global health emergency. *Science of the Total Environment*. **730**, e138996 (2020).
2. Aggarwal, C.C. *Data classification: algorithms and applications*. (CRC Press, Yorktown Heights, New York, USA, 2014).
3. Ahmad, A., Garhwal, S., Ray, S.K., Kumar, G., Malebary, S.J., Barukab, O.M. The number of confirmed cases of covid-19 by using machine learning: methods and challenges. *Archives of Computational Methods in Engineering*. **28**, 2645-2653 (2020).
4. Alelyani, S., Liu, H., Wang, L. The effect of the characteristics of the dataset on the selection stability. In: INTERNATIONAL CONFERENCE ON TOOLS WITH ARTIFICIAL INTELLIGENCE. *IEEE. Proceedings*. 970-977 (2011).
5. Alimadadi, A., Aryal, S., Manandhar, I., Munroe, P.B., Joe, B., Cheng, X. Artificial intelligence and machine learning to fight covid-19. *American Physiological Society Bethesda*, MD, (2020).
6. Allam, M., Cai, S., Ganesh, S., Venkatesan, M., Doodhwala, S., Song, Z., HU, T., Kumar, A., Heit, J., Coskun, A.F., *et al.* Covid-19 diagnostics, tools, and prevention. *Diagnostics*. **10**, 1-33 (2020).
7. Beckmann, J.S., Lew, D. Reconciling evidence-based medicine and precision medicine in the era of big data: challenges and opportunities. *Genome medicine*. **8**, 1-11 (2016).
8. Box, G.E.P., Tiao, G.C. *Bayesian inference in statistical analysis*. (John Wiley and Sons, Canada, 1992).
9. Brazil. Ministry of Health. Coronavirus Panel Brazil. Available in <https://covid.saude.gov.br>, (accessed in April 26, 2022).
10. Breiman, L. Random forests. *Machine learning*. **45**, 5-32 (2001).
11. Dong, E., Du, H., Gardner, L. An interactive web-based dashboard to track covid-19 in real time. *The Lancet infectious diseases*. **20**, 533-534 (2020).
12. Dougherty, G. *Pattern recognition and classification*: an introduction. (Springer Science & Business Media, California, USA, 2012).
13. FAPESP. *FAPESP COVID-19 Data Sharing/BR*, (2020).

14. Gao, Y., Cai, G.Y., Fang, W., Li, H.Y., Wang, S.Y., Chen, L., Yu, Y., Liu, D., Xu, S., Cui, P.F., *et al*. Machine learning based early warning system enables accurate mortality risk prediction for covid-19. *Nature communications*. **11**, 1-10 (2020).

15. Gordis, L. *Epidemiology*. (Elsevier Saunders, Philadelphia, PA, 2014).

16. Grzybowski, J.M.V., Da Silva, R.V., Rafikov, M., 2020. Expanded seircq model applied to covid-19 epidemic control strategy design and medical infrastructure planning. *Mathematical Problems in Engineering*. e8198563 (2020).

17. Han, J., Kamber, M., Pei, J., *Data mining*: concepts and techniques. (Morgan Kaufmann, Burlington, MA, USA, 2012).

18. Hou, W., Zhao, Z., Chen, A., Li, H., Duong, T.Q. Machining learning predicts the need for escalated care and mortality in COVID-19 patients from clinical variables. *International Journal of Medical Sciences*. **18** (8), 1739-1745 (2021).

19. Lalmuanawma, S., Hussain, J., Chhakchhuak, L. Applications of machine learning and artificial intelligence for covid-19 (sars-cov-2) pandemic: a review. *Chaos, Solitons & Fractals*, e110059 (2020).

20. Maimon, O.Z., Rokach, L. *Data mining with decision trees*: theory and applications. (World scientific, 2014).

21. Mello, L. E. *et al*. Opening Brazilian COVID-19 patient data to support world research on pandemics. *Zenodo*, (2020).

22. Nemati, M., Ansary, J., Nemati, N. Machine-learning approaches in covid-19 survival analysis and discharge-time likelihood prediction using clinical data. *Patterns*. **1**, e100074 (2020).

23. R Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/, (2020).

24. Ranzani, O.T., Bastos, L.S.L., Gelli, J.G.M., Marchesi, J.F., Baião, F., Hamacher, S., Bozza, F.A. Characterisation of the first 250 000 hospital admissions for covid-19 in Brazil: a retrospective analysis of nationwide data. *The Lancet Respiratory Medicine*. (2021).

25. Rodríguez-Morales, A., Macgregor, K., Kanagarajah, S., Patel, D., Schlagenhauf, P. Going global – travel and the 2019 novel coronavirus. *Travel medicine and infectious disease*. **33**, e101578 (2020).

26. Steyerberg, E.W., Vickers, A.J., R., C.N., Gerds, T., Gonen, M., Obuchowski, N., J., P.M., Kattan, M.W. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology*. **21**, 128-138 (2010).

27. Vere, J., Gibson, B. Evidence-based medicine as science. *Journal of Evaluation in Clinical Practice*. **25**, 997-1002 (2019).

28. Yadav, M., Perumal, M., Srinivas, M. Analysis on novel coronavirus (covid-19) using machine learning methods. *Chaos, Solitons & Fractals*. **139**, 110050 (2020).

29. Zeng, X., Zhang, Y., Kwong, J.S.W., Zhang, C., Li, S., Sun, F., Niu, Y., Du, L. The methodological quality assessment tools for preclinical and clinical studies, systematic review and meta-analysis, and clinical practice guideline: a systematic review. *Journal of evidence-based medicine*. **8**, 2-10 (2015).