Analysis of COVID-19 severity prognosis using hospital data

Filipe Loyola Lopes[1,2] [0000−0002−4172−6532], Adriano Simões Gaspar[2] and Ana Carolina Lorena[1][0000−0002−6140−571X]

[1] Aeronautics Institute of Technology (ITA), Praça Marechal Eduardo Gomes, 50, Vila das Acacias, São José dos Campos, Brazil
[2]Federal University of Sao Paulo (UNIFESP), Avenida Cesare Mansueto Giulio Lattes, 1201, Eugenio de Mello, São José dos Campos, Brazil
filipe.loyola@unifesp.br

Abstract. This research assembles a dataset of COVID-19 positive patients suitable for the application of Machine Learning (ML) techniques, allowing to obtain prognosis models that distinguish severe from non-severe patients by taking as input hematological exams performed upon hospital attendance. Six ML techniques were applied to analyze data from 4,320 COVID-19 positive patients, 394 of which evolved to a severe health state, requiring intensive care. The Random Forest classifier showed the best predictive performance among the used algorithms and settings, with an AUC score up to 0.94 ± 0.02. In addition, ten clinical variables revealed to be more correlated to the prognosis by a mutual information score, although some of them had a high fraction of missing values.

Keywords: Machine Learning, Health Data, COVID-19, Health Data Science.

1 Introduction

Hospitals produce large amounts of data every day by collecting laboratory exams and medical records from patients. Important insights can be extracted from such data for decision making, including diagnosis and prognosis [1]. There are some previous studies about COVID-19 severity prediction of clinical aggravation [2]. If the prognosis of severity is anticipated, more appropriate management of the hospital resources can take place. Nonetheless, the presence of missing values and noise on predictive attributes in medical records is quite common, making data preprocessing mandatory [3]. The aim of this research is to assemble a dataset of COVID-19 positive patients suitable for the application of Machine Learning (ML) techniques in the induction of severity prognosis models. The resulting dataset has laboratory tests results from 4,320 patients and some analysis of the variables more correlated to clinical aggravation is also made.

2 Methods

The raw database is from the Hospital Sírio Libanês (HSL) and was made available by the FAPESP COVID-19 Data Sharing (https://repositoriodatasharingfapesp.uspdigital.usp.br/). It contains data from 8,971 patients and 954 types of laboratory tests and hospital sectors where the patients have been. Some filters were applied to this database, as follows: removal of repeated values; removal of patients with no additional exams but the COVID test and with negative results for COVID-19; inclusion of exams performed up to the first three days of attendance, at emergency room; and removal of exams with more than 50 percent of missing values. The resulting dataset was further labeled into four categories: Group 0: 3,393 patients who undergone exams on the emergency room only (non-severe); Group 1: 533 patients who undergone exams at the emergency room and infirmary (non-severe); Group 2: 85 patients who undergone exams from emergency room and intensive care unit (ICU) (severe); and Group 3: 309 patients who undergone exams at the emergency room, infirmary and ICU (severe). Next, five different classification datasets (S1 to S5) were created by different combinations between the previous groups: S1: group 0 vs group 2; S2: group 0 vs group 3; S3: group 1 vs group 2; S4: group 1 vs group 3; and S5: group 0 + group 1 vs group 2 + group 3. Each dataset was divided into 70% for training and 30% for testing. Only the training part was submitted to next steps: normalization, elimination of outliers, replacement of missing values by the median value of the class and random under sampling for balancing. The mutual information technique was applied in order to detect variables more correlated to clinical aggravation. Six ML techniques were applied with 10-fold crossvalidation for each of the classification problems assembled and the area under the ROC curve (AUC) metric was used to measure the predictive results achieved. The ML techniques used are: k-Nearest Neighbors (kNN), Decision Tree (tree), Decision Tree big variation (bigtree), Suport Vector Machine with linear kernel (svmlinear), suport vector machine with radial basis function kernel (svmrbf) and Random Forest. Preprocessing software is available at https://doi.org/10.5281/zenodo.6392307 and application of ML softwares is available at https://doi.org/10.5281/zenodo.6413250.

3 Results and discussion

Data preprocessing resulted in 4,320 patients and 27 variables. Considering only Set 5, which is the larger regarding the number of patients, the ten predictive attributes most correlated with clinical aggravation and its respective number of missing values (in parenthesis) are: 1.Alanine Aminotransferase (ALT/TGP) (1135); 2.C-Reactive Protein (1048); 3.Aspartate Aminotransferase (AST-TGO) (1152); 4.Age (0); 5.Creatinine (627); 6.Urea (778); 7.RDW (330); 8. Potassium (1011); 9.Erythrocytes (272); 10. Neutrophils (308). The AUC ML results from ML techniques with 10-fold cross-validation can be seen in Table 1.

Table 1. Results from six machine learning techniques applied to five sets data.

Algorithms	**S1 (n = 3478)**	**S2 (n = 3702)**	**S3 (n = 618)**	**S4 (n = 842)**	**S5 (n = 4320)**
kNN	0.80 ± 0.09	0.86 ± 0.04	0.72 ± 0.13	0.67 ± 0.05	0.84 ± 0.04
tree	0.75 ± 0.10	0.86 ± 0.04	0.58 ± 0.07	0.77 ± 0.05	0.85 ± 0.05
bigtree	0.72 ± 0.12	0.76 ± 0.06	0.58 ± 0.09	0.71 ± 0.06	0.79 ± 0.05
svmlinear	0.85 ± 0.10	0.87 ± 0.04	0.73 ± 0.10	0.69 ± 0.07	0.84 ± 0.04
svmrbf	0.82 ± 0.13	0.86 ± 0.05	0.66 ± 0.08	0.64 ± 0.10	0.85 ± 0.05
Random Forest	**0.92 ± 0.09**	**0.94 ± 0.02**	**0.76 ± 0.10**	**0.82 ± 0.03**	**0.93 ± 0.03**

The algorithm with the best predictive performance for all datasets was Random Forest. From Table 1 it is possible to observe that the greater the amount of included patients, the better tends to be the AUC results. The mutual information technique results in ten predictive attributes, but some care is necessary about missing values in variables such as ALT, TGP and C-reactive protein, because these variables have a high number of missing values, their values can become biased by data imputation procedures.

4 Acknowledgments

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001. The authors also thank the Brazilian research agency FAPESP for providing hospital data under the FAPESP COVID-19 Data Sharing/BR initiative.

References

1. Alizadehsani, R., Roshanzamir, M., Hussain, S., Khosravi, A., Koohestani, A., Zangooei, M. H., & Acharya, U. R. (2021). Handling of uncertainty in medical data using machine learning and probability theory techniques: A review of 30 years (1991–2020). Annals of Operations Research, 1-42.
2. Wynants, L. et al (2020). Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. BMJ, 369.
3. Rahman, M. M. and Davis, D. N. (2013). Machine learning-based missing value imputation method for clinical datasets. In IAENG transactions on engineering technologies, pages 245–257.