UNICAMP

UNIVERSIDADE ESTADUAL DE CAMPINAS

Instituto de Matemática, Estatística e Computação Científica

VITOR PEREIRA BEZZAN

Machine Learning Methods applied to COVID-19 patient data

Métodos de Aprendizado de Máquina Aplicados a dados de pacientes de COVID-19

Campinas

2021

Vitor Pereira Bezzan

Machine Learning Methods applied to COVID-19 patient data

Métodos de Aprendizado de Máquina Aplicados a dados de pacientes de COVID-19

Dissertação apresentada ao Instituto de Matemática, Estatística e Computação Científica da Universidade Estadual de Campinas como parte dos requisitos exigidos para a obtenção do título de Mestre em Matemática Aplicada e Computacional.

Dissertation presented to the Institute of Mathematics, Statistics and Scientific Computing of the University of Campinas in partial fulfillment of the requirements for the degree of Master in Applied and Computational Mathematics.

Supervisor: Cleber Damião Rocco

Este trabalho corresponde à versão final da Dissertação defendida pelo aluno Vitor Pereira Bezzan e orientada pelo Prof. Dr. Cleber Damião Rocco.

Campinas

2021

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Matemática, Estatística e Computação Científica
Ana Regina Machado - CRB 8/5467

B469m

Bezzan, Vitor Pereira, 1990-
Machine learning methods applied to COVID-19 patient data / Vitor Pereira Bezzan. – Campinas, SP : [s.n.], 2021.

Orientador: Cleber Damião Rocco.
Dissertação (mestrado profissional) – Universidade Estadual de Campinas, Instituto de Matemática, Estatística e Computação Científica.

1. Aprendizado de máquina. 2. COVID-19. 3. Modelagem matemática. 4. Estatística matemática. I. Rocco, Cleber Damião, 1980-. II. Universidade Estadual de Campinas. Instituto de Matemática, Estatística e Computação Científica. III. Título.

Informações para Biblioteca Digital

Título em outro idioma: Métodos de aprendizado de máquina aplicados a dados de pacientes de COVID-19
Palavras-chave em inglês:
Machine learning
COVID-19 (Disease)
Mathematical modeling
Mathematical statistics
Área de concentração: Matemática Aplicada e Computacional
Titulação: Mestre em Matemática Aplicada e Computacional
Banca examinadora:
Cleber Damião Rocco [Orientador]
Clarissa Lin Yasuda
Sandra Eliza Fontes de Avila
Data de defesa: 13-08-2021
Programa de Pós-Graduação: Matemática Aplicada e Computacional

Identificação e informações acadêmicas do(a) aluno(a)
- ORCID do autor: https://orcid.org/0000-0001-6562-2976
- Currículo Lattes do autor: http://lattes.cnpq.br/1455482906312273

Dissertação de Mestrado Profissional defendida em 13 de agosto de 2021 e aprovada pela banca examinadora composta pelos Profs. Drs.

Prof(a). Dr(a). CLEBER DAMIÃO ROCCO

Prof(a). Dr(a). CLARISSA LIN YASUDA

Prof(a). Dr(a). SANDRA ELIZA FONTES DE AVILA

A Ata da Defesa, assinada pelos membros da Comissão Examinadora, consta no SIGA/Sistema de Fluxo de Dissertação/Tese e na Secretaria de Pós-Graduação do Instituto de Matemática, Estatística e Computação Científica.

This work is dedicated to all who suffered from COVID-19. By the time we wrote this dissertation, more than 600,000+ people had perished in Brazil. May they never be forgotten.

Acknowledgements

Special thanks to all that collaborated to make this work possible. To my supervisor, Prof. Cleber, for all time spent making the necessary changes and turning this into a reality; To my wife Franciele, and my family Iraíma, Rogério, and Otávio for giving me the necessary support.

Resumo

Apresentamos aqui o resultado final de dois artigos que são resultados de uma pesquisa do uso de aprendizado de máquina (*machine learning*) em dados de pacientes acometidos pela doença COVID-19.

No primeiro artigo, nos concentramos em desenvolver um modelo de previsão que fosse capaz de prever a gravidade de um dado paciente de COVID-19 e o número total de dias que este mesmo paciente pudesse ficar internado (em regime ambulatorial ou tratamento intensivo) usando-se para isso dados advindos de exames de sangue. Para tanto, foram utilizadas técnicas de modelagem desenvolvidas recentemente como modelos de *ensemble* de árvores e otimização bayesiana para seleção de modelos entre vários candidatos. Os resultados finais apontam modelos com 0.94 para a área sob a curva ROC para o classificador estimado e 1.87 de erro quadrático médio (uma melhora de 77% sobre o cenário de base) para o regressor estimado.

No segundo artigo, apresentamos o uso de aprendizado não-supervisionado baseado na combinação de uma técnica de redução dimensional (UMAP - *Uniform Manifold Approximation and Projection for Dimension Reduction*) com algoritmo de clusterização DBSCAN em dados oriundos de exames de sangue de pacientes infectados com COVID-19. Com isso, foi possível explicitar diferentes grupos com variadas possíveis manifestações da doença (prevalências entre 2-37%). Além disso, fomos capazes de afirmar que existe evidência observacional (não-causal) de que a doença afeta principalmente a série branca do sangue (associada aos processos de coagulação e imunidade).

Os dois artigos em conjunto permitem estabelecer relações observacionais (não-causais) fortes sobre a real natureza da infecção por SARS-Cov-2 caudador da COVID-19, que parece não ser somente de natureza respiratória, mas multi-sistêmica com implicações nos sistemas imune e plaquetário.

Palavras-chave: COVID-19. Aprendizado de Máquina. Exames de Sangue, Otimização Baysiana.

Abstract

We present here the final result of two articles that are the results of a research on the use of machine learning in data from patients affected by the disease COVID-19.

In the first article, we focused on developing a predictive model that would predict a given COVID-19 patient's severity and the total number of days that the same patient could be hospitalized (on an outpatient or intensive care basis) using data from blood tests. For this, we have used modeling techniques such as tree ensemble models and Bayesian optimization for model selection among several candidates. The final results point to models with 0.94 for the area under the ROC curve for the estimated classifier and 1.87 for the mean squared error (an improvement of 77% over the baseline scenario) for the estimated regressor.

In the second article, we present unsupervised learning based on the combination of a dimension reduction technique (UMAP - Uniform Manifold Approximation and Projection for Dimension Reduction) with DBSCAN clustering algorithm in data from blood tests of patients infected with COVID-19. Thus, it was possible to explain different groups with different disease manifestations (prevalence between 2-37 %). Also, we were able to claim that there is observational (non-causal) evidence that the disease primarily affects the white blood series (associated with clotting and immunity processes).

The two articles taken together make it possible to establish solid observational (non-causal) relationships about the fundamental nature of COVID-19 infection, which appears to be not only a respiratory but of multi-systemic nature with implications for the immune and platelet systems.

Keywords: COVID-19. Machine Learning. Blood tests. Bayesian optimization.

List of Figures

List of Tables

Contents

1 Introduction and concepts

In this introductory chapter I present the motivations and also some concepts used in the articles and give necessary references for readers to proceed ahead on their journey if necessary.

At the end of 2019, right after the first detection of the Sars-Cov2 virus in humans, a process of discovery of knowledge that has never been seen before in the history of science began. What were isolated cases have become a global threat and more than ever all available resources have been allocated to face it.

In Brazil, the first cases started to appear in February. The first case was a patient from São Paulo who had traveled to Italy days before. São Paulo is one of the most important, richest and most populated state among all the states in Brazil, and any epidemic of global proportions tends to hit the state aggressively. The procedure adopted for the containment of the epidemic in Brazil was unsatisfactory at the federal level - and at the time this introduction was included in this dissertation, more than 600,000 people had already perished from COVID-19, the name given to the disease caused by the virus mentioned above. In this text, I bring two studies from my master in applied mathematics developed during this period to contribute to overcome this pandemic.

In the first article, entitled "Predicting special care during the COVID-19 pandemic: A machine learning approach" we seek to use Machine Learning and Bayesian optimization techniques for black-box functions to create a recommendation system that can, with data from blood tests of newly admitted patients, to be able to classify whether a particular patient will require special care and how many days each patient will need such care.

The main logical justifications for following this procedure are detailed in their respective chapters. In this case, several bibliographic sources point to a direct relationship between changes in the measured values of variables arising from blood tests and a positive diagnosis of COVID-19 in a given patient (FERRARI; MOTTA et al., 2020). The use of Machine Learning techniques is therefore necessary because it is to be expected that the data will have multivariate patterns that can be discovered in this way (BRINATI; CAMPAGNER et al., 2020a). The Bayesian optimization process is based on the choice of hyperparameters that are part of the model and that cannot be chosen directly.

We also tried to predict the number of days a given patient would spend in special care. In this context, a decision system based on our models is able to help doctors, hospital administrators and others involved to manage resources more rationally and also increase the quality and satisfaction of patient care. The first article is the topic on Chapter 2.

Based on the logical and theoretical justifications presented in the first article, we expand our intentions with a second article, entitled "Using two-dimensional representations to understand patterns on COVID-19 blood exam data". In this article we have more theoretical pretensions.

We use a dimensionality reduction technique still little used in the medical field known as UMAP (Uniform Manifold Approximation and Projection for Dimension Reduction) which is able to capture global structures without losing local structures in the data (MCINNES; HEALY; MELVILLE, 2020). Although the technique is capable of finding subspaces of arbitrary dimension, we are dedicated to finding two-dimensional structures for a better visualization of readers and greater absorption of the proposed experiments.

Following this dimensionality reduction, we use a clustering technique to find patients who are similar to other patients in the group but different from patients in other groups. To do so, we use a technique called DBSCAN, an old familiar to data scientists - which allows defining the clusters from the distance of two samples, and not from the number of desired clusters. Our second article is on Chapter 3.

It should be made clear that all conclusions reached in both articles are non-causal - that is, we do not focus on establishing cause-and-effect relationships for any of the statements we make. Furthermore, we know that from a medical point of view some results obtained must be tested from a multivariate point of view, which is beyond the scope of the text.

As applied researchers, we stick to more "applied" aspects of our study. However, we invite all interested parties for a frank discussion towards the continuous improvement of what we have developed so far. In this sense, all code is available in our github page.

The published text of the first article can be found here.

The main basic technical concepts used in the studies are presented below.

1.1 Supervised learning

Supervised learning is a branch of Machine Learning dedicated to learn patterns from data using labeled examples. Given a dataset D, it is pretty common to associate with it two subsets. The first one of dimensionality n (commonly known as *columns*), $X \in R^n$, contains the features (also called covariates, variables or independent variables). The second one, $y \in R^m$, contains the targets (also called labels or dependent variables). For most settings, each pair (X, y) can be considered one "training example". We will focus on the specific univariate case of $m = 1$, as the overall justification we will present works well with some modifications for the case $m > 1$.

The main task for this type of setting is to find a suitable function $\hat{f}(X)$ that relates X and y from a pre-defined "family of functions" $f(X)$ in order to minimize prediction error ε. Common choices for f can be linear models, decision trees or neural networks (Eq. (1.1)).

$$y = \hat{f}(X) + \varepsilon \tag{1.1}$$

Supervised learning problems can be further divided in two subcategories - regression and classification. On a regression problem, y usually takes values on R^n (or some compact subset of it). On a classification problem, y is usually constrained in the $[0, 1]$ interval that represents the probability each instance has of being part of one specific class. In this context, ε is usually substituted by a loss function that should be minimized in order to find the best $\hat{f}$.

A common choice for a loss function (where N is the number of examples in the dataset) is the "mean squared error" (MSE) (Eq. 1.2), usually used in regression problems. For classification, is very common to use Equation (1.3), which is known as cross-entropy (as it measures how well one class can be predicted using the information about the other).

$$MSE = \frac{1}{N}\sum^{N}(y - \hat{f}(X))^2 \tag{1.2}$$

$$J = -\sum^{N} y \log(\hat{f}(X)) + (1 - y)\log(1 - \hat{f}(X)) \tag{1.3}$$

On our articles we will focus in a specific family of models f called boosting trees, which consist of several decision trees, and a set of rules to estimate each tree in sequence, in most cases, a new weighted version of the dataset is used to estimate new trees - these weights are proportional to the mean loss function for the training examples being considered. For more information about boosting, please see (FREUND; SCHAPIRE, 1997).

A decision tree is a function that associates a constant value to a specific partition R_i of the space X (divided into R_1, R_2,..., R_M). Its functional form is by consequence very simple function (Eq. 1.4), where I is the indicator function for the set and θ is related to further

restrictions applied on the tree, also called *hyperparameters*. To fit an specific tree, greedy algorithms are used to find the optimal partitions R_i. See (HASTIE; TIBSHIRANI; FRIEDMAN, 2009) for more details.

$$\hat{f}(\theta, X) = \sum^{M} c_M I(X \in R_M) \tag{1.4}$$

Special attention needs to be devoted to find θ. As they are not related at all with X or y, a different procedure to find them is necessary - and this procedure not only finds sufficiently good values for θ, but also validates the loss measurement for our function $\hat{f}$. The procedure consists of creating two subsets of our dataset D, called "training" and "testing" datasets. The model is fitted (estimated/adjusted) on the "training" dataset and all relevant statistics are calculated on the "testing" dataset (never used to fit any function). By using this procedure, we are capable of selecting the best θ for a given family of functions f. For more information about validation of models and cross-validation, please see (EFRON, 1983).

1.2 Unsupervised learning

In the supervised setting, our dataset D was comprised of (X, y) example pairs. When considering X and y as random variables that are related via a function $\hat{f}(X)$ of some pre-defined family of functions, we are tempted to write the expression below (where ρ denotes the probability density function of a random variable)

$$\rho(X, y) = \rho(X|y)\rho(X) \tag{1.5}$$

so our estimated function $\hat{f}(X)$ can be associated with $\rho(X|y)$, where we intend to learn how X "affects" y. When we consider only the $\rho(X)$ term, we land then on the setting of unsupervised learning, where we want to learn about X without the "help" from y or a loss function. Summarizing, unsupervised learning consists of learning some set of parameters λ that describe $\rho(X)$.

The choice of λ and its subsequent evaluation depends entirely on the user, as there are no specific techniques or loss functions associated to the problem. Common problems in unsupervised learning are:

- **Clusterization** - Finding groups of X where its elements are similar to each other, but different between groups;
- **Anomaly detection** - Finding training instances that are anomalous (different) from the majority of others on the dataset D;

- **Dimensionality reduction** - Finding a low-dimensional representation of data while retaining the most information possible.

On the articles presented here, we use clustering and dimensional reduction techniques. On the clustering side, we focus specifically on DBSCAN that consists of finding core points, directly reachable points and reachable points based on some distance specified by the user (ε). This technique was first introduced in (ESTHER; KRIEGEL et al., 1996).

On the dimensionality reduction side, we use a technique called UMAP with its roots in Riemmanian geometry and algebraic topology (beyond the scope of this text) which tries to model a global structure while preserving local structures. Please see (MCINNES; HEALY; MELVILLE, 2020) for more details.

1.3 Bayesian optimization

On finding the parameters θ for a specified supervised learning task, the common procedure usually is to break it in a grid, and testing the overall loss function for all elements of the grid. But this approach have two inherent flaws.

The first flaw is the fact that selecting the grid itself is not a well-established procedure; There are a lot of *ad-hoc* procedures, but no standard approach exists. The second flaw resides on the fact that θ can have several dimensions, thus hampering even more the grid selection because the grid elements will suffer from the curse of dimensionality (they will be mostly located on the edge of the space).

To alleviate these problems, instead of using traditional grid-searching methods to minimize the loss function $L(\hat{f}(X), y|\theta)$ we use Bayesian optimization, which considers L as a black-box function and uses the following processes:

- Create a prior using sampled points from the loss function, sampled using some strategy as Sobol sequence numbers or Gaussian sequence numbers;
- Update the prior based on the evaluations of the functions, locating areas with lowest-/biggest values;
- Define next evaluation points based on the posterior distribution obtained.

Using this technique, we are capable to optimize a black-box function, without derivatives. For more details, please check the seminal article (MOCKUS, 1974).

2 Predicting special care during the COVID-19 pandemic: A machine learning approach

More than ever, COVID-19 is putting pressure on health systems worldwide, especially in Brazil. In this study, we propose a method based on statistics and machine learning that uses blood lab exam data from patients to predict whether patients will require special care (hospitalization in regular or special-care units). We also predict the number of days the patients will stay under such care. The two-step procedure developed uses Bayesian Optimisation to select the best model among several candidates. This leads us to final models that achieve 0.94 area under ROC curve performance for the first target and 1.87 root mean squared error for the second target (which is a 77% improvement over the mean baseline)—making our model ready to be deployed as a decision system that could be available for everyone interested. The analytical approach can be used in other diseases and can help to plan hospital resources in other contexts.

2.1 Introduction

The COVID-19 pandemic is a considerable challenge for Brazil and many other countries around the world. The disease is putting tremendous pressure on health care services and there is no strong consensus on what measures are the most effective in terms of dealing with it. There are various independent reports that indicate a high occupancy rate in intensive care units with facilities to support patients who have severe respiratory tract failure and related conditions, thus creating a unique opportunity to solve this problem with scientific rigor helping to improve this difficult situation. The disease is spreading quickly, and social distancing measures are being phased out in several countries despite recommendations on the contrary issued by the World Health Organization (WHO) and the Centers for Disease Control and Prevention (CDC) (SANCHE; LIN et al., 2020).

As pointed out by (PEIFFER-SMADJA; MAATOUG et al., 2020), the massive amount of data acquired from several sources should be put into fair use for intensive training of machine learning algorithms to better understand the disease, the patients, and possible prognosis, enabling informed decision-making. Our main motivation is to unify subjects, such as Machine Learning, Optimization, Hospital Planning and applied AI to serve the purpose of using hospital resources responsibly and improve the quality of care provided to patients. We propose an analytical approach that leverages the most recent discoveries in each one of these areas and uses laboratory blood test data to estimate the probability of one given patient to require special-care treatment, also estimating the number of days the same patient will be under such care. Our aim is to create the basis of a decision system that can be used by anyone interested in replicating and estimating such outcomes, with the capability to expand the proposed method to deal with other diseases when needed.

We used data available in (FAPESP, 2020), which joins laboratory test data from the Sírio Libanês Hospital, Albert Einstein Israeli Hospital, and Fleury Laboratories (all located in the city of São Paulo, Brazil). These data comprise several different laboratory tests performed on patients (mostly blood tests). This preference for blood tests is not coincidental: most of them are well-standardized and usually inexpensive to perform, accessible in most situations, even for developing countries.

This article is organized as follows. In Section 3.2, we examine some of the most relevant literature present in machine learning with a healthcare perspective. In Section 3.3, we present our analytical approach used to create ML models to predict special care probability and extend the same techniques to predict how many days any given patient will spend under such care - focusing on the overall applicability and explainability of the models trained. The overall numerical results are then presented for both targets in Section 3.4, considering the candidate models and the final selected optimized ones. Finally, we present our conclusions, limitations and possible extensions that should follow for other diseases and situations where our approach could be useful.

2.2 Literature Review

This literature review will focus on shedding light on recent efforts using ML and decision systems from a healthcare perspective. Some specific references concerning COVID-19 will be analyzed. Moreover, we will also focus on new, interesting and emerging applications for other diseases and situations to clarify research in the subject and compare this article with others in the same field.

Using statistical methods in healthcare for a large number of individuals comprising a great number of data points dates back to the 1950s. The Framingham Heart Study was established, showing correlations between doctors' health measurements (including some laboratory test results) and heart diseases, diabetes, and obesity. See (MAHMOOD; LEVY et al., 2014) for a historical perspective and (BERTSIMAS; OHAIR; PULLEYBLANK, 2015) for a statistical point of view. This study is considered one of the finest and earliest examples of how statistics and decision systems could be implemented to help governments and policymakers make well-informed decisions that have a huge impact on a specific individual's quality of life and overall survival rate.

After the 1950s, with the advent of faster computers that have high-level programming languages and frameworks, several studies arose under the ML and decision systems umbrella. From medicine to economics and social sciences, these studies helped people and governments to make more scientifically informed decisions with *really huge* and diverse data coming from different sources. From now on, we will focus on recent developments.

Recent examples of ML being used to detect and diagnose different types of diseases using test data appear in other contexts. In (GUNčAR; KUKAR et al., 2018), classifiers can be observed that are applied to detect hematological disorders and are sometimes better than hematologists themselves. They are frontiers that algorithms, in general, are reaching leading to substantial implications.

In (ALSHEREF; GOMAA, 2019), the authors use laboratory data on patients also to detect blood diseases. In their approach, they select several candidate models within minimal pre-treatment of data to understand which algorithm behaves better. In the present study, we expand our reach by proposing a second optimization procedure on the selected algorithm type to improve the specificity-sensitivity characteristics of the final optimized model. Please see the scheme in Figure 10 for more details.

Blood test data are also being used to detect more complex types of diseases. There is a particular interest in several areas, in which (LIU; OXNARD et al., 2020) is an excellent example. They aim to detect more than 50 types of different cancers by analyzing different DNA signatures, showing a 99.3% specificity rate. This article can be seen as an improvement in the field of "liquid biopsies," reducing the need for patients to undergo complicated procedures to be given a diagnosis.

There are other diseases where ML algorithms-aided diagnosis could play a significant role. For example, (WU; YEH et al., 2019) applies random forests for the final selected model to predict fatty liver disease and create an indicator to separate high-risk patients from low-risk ones, effectively allowing customization in treatments and improving overall outcomes. Considering other perspectives, there is also a substantial number of studies using algorithms that do not rely on laboratory data to predict outcomes (for example, deep learning to learn from medical images). A useful review on this topic is provided by (FATIMA; PASHA, 2017), where heart disease applications, dengue fever, hepatitis, and diabetes are explored.

Analyzing the interface in decision systems, we can cite (BEELER; DBEIBO et al., 2018) as an application of ML-backed classifiers to understand the potential of bacterial infection in a given patient in a hospital setting. Special attention is given to prioritizing hospital resources and early detection of bacteremia, an infectious disease caused by microorganisms that propagate much like COVID-19. On the same topic, we can also cite (COLUBRI; SILVER et al., 2016), an article showing the creation of a decision system given to hospitals to predict the outcomes of Ebola in West African patients (Ebola is a highly contagious virus that demands special care of patients, resembling COVID-19).

There is also a wide range of books on these topics. In (JAIN; CHATTERJEE, 2020), various ML applications can be observed in different areas spanning disease diagnostics with laboratory data, image recognition methods, unsupervised learning and the Internet of Things.

Interest in these topics is becoming more substantial as time passes and technology advances. Conferences and meetings are being held in several places. One notable example is the *Machine Learning for Healthcare* (MLHC, 2020) conference, which took place virtually in 2020 due to the COVID-19 pandemic.

Specifically linked to COVID-19, there are several reports on the use of ML to detect the disease using laboratory data. In (BRINATI; CAMPAGNER et al., 2020a), the authors trained classifiers that attained an 82%-86% accuracy while keeping high levels of specificity and sensitivity, therefore increasing the general applicability of the method selected. There is also an example in (DUTTA; BANDYOPADHYAY, 2020) of deep learning-based methods used to estimate the overall epidemiological parameters for the disease considering stacked Long Short-term Memory (LSTM) models and polynomial neural networks.

Some novel and fresh approaches are emerging from the need to diagnose patients using any data available. In (ELAZIZ; HOSNY et al., 2020), a novel feature generation approach can be observed in X-ray images combined with optimization techniques and high-performance computing used to create a classifier for patients with 96-98% accuracy. On an even more unusual front, text data is being used to diagnose patients in (KHANDAY; RABANI et al., 2020).

Considering that COVID-19 is itself a relatively novel subject, extensive reviews for articles relating it with ML algorithms are only beginning to emerge. One of the first examples is

addressed in (LALMUANAWMA; HUSSAIN; CHHAKCHHUAK, 2020).

There are two main differences between this article and the ones cited earlier. The first one is the target itself: instead of predicting the presence/absence of COVID-19 in one give patient, we attempt to explore the probability of this patient requiring special care at hospital (and the number of days required under special care). The second main difference is the number of algorithms: instead of focusing on one or two algorithms, we firstly considered several, and then we select the best algorithm class overall to perform the Bayesian Optimisation. Table 1 summarizes the findings in this section and positions our study among them.

Table 1 – Review of machine learning for disease prediction.

Reference	Algorithm	Key Results
(BERTSIMAS; OHAIR; PULLEYBLANK, 2015)	Logistic Regression, Random Forests	0.72 AUC
(COLUBRI; SILVER et al., 2016)	Model Ensembles	0.80 AUC
(WU; YEH et al., 2019)	Random Forests	0.92 AUC
(BEELER; DBEIBO et al., 2018)	Random Forests	0.82 AUC
(ALSHEREF; GOMAA, 2019)	Several	0.69 - 0.97 AUC
(GUNčAR; KUKAR et al., 2018)	Random Forests	59% - 80% Precision
(LIU; OXNARD et al., 2020)	Several	99.3% Specificity
(BRINATI; CAMPAGNER et al., 2020a)	Random Forests, SVM and others	92% - 95% Sensitivity
(DUTTA; BANDYOPADHYAY, 2020)	LSTM	62% - 87% Accuracy
(ELAZIZ; HOSNY et al., 2020)	DNNs	96% - 98% Accuracy
(KHANDAY; RABANI et al., 2020)	Naïve Bayes	96.20% Accuracy
(JAIN; CHATTERJEE, 2020)	Several	-
(LALMUANAWMA; HUSSAIN; CHHAKCHHUAK, 2020)	Several	-
(FATIMA; PASHA, 2017)	Several	-
This article	xgBoost + Bayesian Optimization	0.94 AUC

2.3 Method

This section addresses all the groundwork used in this study. Firstly, we present some medical basis, showing some results and references linking blood test results and their respective impacts on COVID-19 patients. We also offer the algorithmic reasoning behind all the techniques involved and why we selected them.

2.3.1 Medical Basis

As COVID-19 is a virus, it is coherent to assume that it causes changes in patients´ blood tests. The article (LIPPI; PLEBANI, 2020) brings a structured review on the parameters that show abnormalities in blood tests to a given patient when contracting COVID-19. Table 2 contains an excerpt of the main tests that show significant changes in laboratory test results for the patients analyzed in this study.

There are also consistent abnormalities described in (FERRARI; MOTTA et al., 2020), mainly dealing with white-blood cells, platelets, C-reactive protein, AST, ALT, GGT,

and LDH parameters. This study concludes that some cutoffs for these tests could be applied as an alternative to RT-PCR tests when necessary and pave the way for automated tests using ML when more patient data becomes available.

In (YUAN et al., 2020), the patients were separated using the overall gravity of the infection, which could be used as a proxy for special-care treatment. This study's main results point out significant changes comparing the patients with established reference values and within different infection gravity groups. The most relevant values obtained were for the white-blood-cell count, LDH, C-reactive protein and others. Moreover, the article concludes by stating that the virus could be related to a state of hyper-coagulation in critically-ill patients, exposing a possible interaction between COVID-19 and laboratory blood test results. Knowing these facts, we propose an extension to use the same test data jointly with hospital outcomes to predict whether the same given patient will also need special care - effectively anticipating the use of valuable medical time and resources. We also model the number of days each patient will be in special care using the same data.

Table 2 – Main abnormalities found in COVID-19 patients, according to (LIPPI; PLEBANI, 2020).

Lab Exam	COVID-19 Effects
Albumin	Decrease
Reactive C-Protein – PCR	Increase
Eritrocytes	Increase
Haemoglobin	Decrease
Leukocytes	Increase
Neutrofils	Increase
Lymphocytes	Decrease
TGP-ALT	Increase
TGO-AST	Increase
Lactate Desidrogenase - LDH	Increase
D Dimer	Increase
Bilirrubin	Increase
Creatinin	Increase
Troponin I	Increase
Procalcitonin - PCT	Increase
Protrombin	Increase

2.3.2 Machine learning procedure

Even without analyzing the available data, it is expected from the domain of science data that three things should be present: **sparsity**, as some laboratory tests are not performed for all patients, revealing many gaps (NAs) in the dataset. Moreover, one should expect **unbalancing**, as not all patients will require special care (only a small number of them will need it). The last thing expected is **non-linearity and interaction**. As every patient will have a different set of

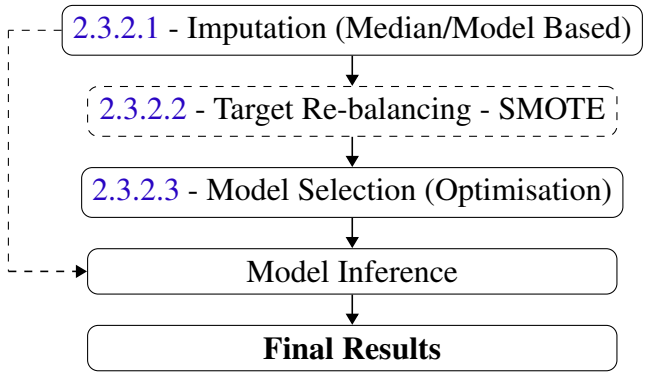

Figure 1 – Steps in second part for our targets. Black continuous arrows are for training phase and dash one for prediction phase. Dashed step is not applied in number of days target.

variables, the final combination and composition will express the outcome distinguished for each patient.

We will focus primarily on Sirio Libanês Hospital data, which includes patient outcomes and dates of admission and discharge, making it possible to analyze the number of days each patient stays in special care and associate it with laboratory test data. All data is taken for each patient, and a pre-processing step is carried out to relate the first test ever recorded for the patient, therefore we preserve the time dependency relevant to the problem. Later test should not constitute reliable data as they introduce temporal leaks.

To model the situation correctly, we propose (for both targets) a two-part procedure that addresses all issues cited above. The first part comprises an initial exploration of data to understand its particular shape and properties, focusing on age and blood white-cell components, as discussed earlier. After that, we explore the usage of *off-the-shelf* algorithms with little to no customization to better understand which candidate suits best - considering the baselines for each model (a coin for the classifier and the average training value for the target number of days in special care), as well the overall capacity to accept different *hyperparameters* to increase the fitness of the model. We also consider the training time and complexity trade-offs of all algorithms as a secondary but important factor.

Once the selected class of model is chosen, we follow the procedure outlined in Figure 10, composed of data imputation, re-balancing, and estimation steps. The following subsections will deal with practicalities and possible choices showing the pros and cons for each one of the steps to pave the way to establish a precise method that can be used in other similar situations.

2.3.2.1 Imputation strategies

To process the data **sparsity**, we have three options with different assumptions, and each one implies model dynamics that are discussed in the next paragraphs. A sparsity treatment

similar to ours can be found in (MAZUMDER; HASTIE; TIBSHIRANI, 2010), a seminal article in the field.

The first one retains the sparsity, i.e., not applying any technique to deal with the completion of variables. There are two disadvantages to this - the first one is that most models do not handle sparsity very well. Some of them even fail altogether during the training phase as they depend on a dense matrix for parameter estimation (a significant part of the *"classical statistical"* models fall in this category). The second major issue is that models, in general, need some variance to "learn" the most relevant variables in a dataset. When a dataset is substantially sparse, some variables lose their "protagonism" and may become irrelevant even whether they are essential considering the application domain. The main advantage of using this approach is that data can be used *as it is*, without resorting to pre-processing and cleaning.

The second major option relies on model-based variable completion, such as the ones presented in (KUMAR, 2016) and (TROYANSKAYA; CANTOR et al., 2001). Most of these procedures consist of Singular Value Decomposition variants, commonly used in biological and medical applications. These model-assisted matrix completion algorithms introduce interaction terms that can be very useful whether the number of patients is high enough in the dataset. This technique's main disadvantage is the care needed to find the optimal values for each of the hyperparameters in each of the algorithms, in turn consuming more time and computation resources. This is a barrier to implementing it for a huge dataset. However, there are some developments in running the algorithms more efficiently and parallelly distributed.

The third and more straightforward way is by inputting some known statistics of the sample as the default value for each variable. The most common values used for this are the mean and median (using the points with observations). Overall justification for this procedure relies on the fact that assuming that there are more healthy patients than unhealthy ones (or more patients that do not require special care), the mean and median for a sample describes a healthy population as the number of samples increase, helping models to identify abnormal values. The main disadvantage remains that some tests can be prescribed more for unhealthy (or healthy) patients, therefore, skewing the mean to be used as input, generating some sample bias.

In this study, we choose the second and third options interchangeably in different parts of the analysis - with a particular preference to use the third one, simplifying the calculations.

2.3.2.2 Data re-balancing

We should expect from the data that not all patients require special care. Moreover, it is likely that only a few of them will. In machine learning, this type of problem is known as **unbalancing** between classes. By having only a few samples of one specified occurrence, the model cannot generalize well, considering the few examples giving a low specificity/sensitivity model. Here accuracy is not essential because a model that responds to the predominant class

will generally present a good value for accuracy. The Receiver Operating Characteristics (ROC) statistics can also be affected by this situation to a minor extent.

Some studies have attempted to understand the overall effect of unbalancing on classifiers of different types. For example, (NGUYEN, 2019) tries to understand the widespread impact in several publicly available datasets and even proposes changes in calculating performance metrics that are more adequate to these situations. This is undoubtedly an improvement to the original problem, but we will use another alternative that is more automated and depends less on human interaction.

Manual techniques such as undersampling of the majority class or oversampling of the minority class through bootstrapping were usually considered in the past for some studies and practical applications, with mixed results and poor reproducibility when new data arrives for model updates. To avoid this, here we will use the Synthetic Minority Oversampling Technique as described by (NGUYEN, 2002), a technique to combine the minority class oversampling and synthetic example generation with majority class undersampling, augmenting the area under the ROC curve statistics, making the model more sensitive to the minority class.

2.3.2.3 Model Estimation and Optimization

When selecting models for a specific application, several aspects should be considered. The most relevant is the overall "*capacity*" of the algorithm - how a particular algorithm learns about different patterns existing in data without over-fitting to it. Most algorithms regulate this capacity by the change of hyperparameters controlling various aspects. Finding optimal hyperparameters is a matter of discussion in scientific debates as ML has gained traction as an everyday tool, as pointed out by (FEURER; HUTTER, 2019), and is still a growing field for discoveries. Well-known libraries among data scientists for computational ML implement different strategies (see (PEDREGOSA; VAROQUAUX et al., 2011) for a good example). Most of them are based on grid searches of several parameters. Moreover, there are two major disadvantages doing this. The first and more obvious one is in the process itself, requiring a high number of evaluations in the cross-validation process, directly proportional to the number of folds. The second is less apparent and more critical which refers to the search space that needs to be crafted and selected (considering all relevant parameters for the problem).

While most techniques cannot deal well with the second disadvantage (crafting the search space), there is a possible improvement usually requiring fewer evaluations in our cross-validation procedure with its roots in optimization and statistics. Here we propose Bayesian Optimization as in (MOCKUS, 1994) to select model hyperparameters achieving optimal performance within the selected grid. Our procedure will be very similar to the method described in (SNOEK; LAROCHELLE; ADAMS, 2012). The parameters we optimize will be discussed in the Results section for the selected algorithm.

Other algorithms and heuristics can be considered in this optimization problem. There

are articles considering this in different contexts; good examples are (LORENZO; NALEPA et al., 2017), (QOLOMANY; MAABREH et al., 2017) and (LALWANI; MISHRA et al., 2021), which consider some variations on heuristics from traditional particle swarms with different hyperparameter selections to more intricate heuristics such as gravitational search algorithm. There is a recent example of heuristics that was applied to a biological context in (HAN et al., 2021). We consider applying heuristics in future revisions of our technique with new datasets. The authors opted for a Bayesian Optimization approach because our previous experience with the algorithm helped us to validate our results quickly.

2.3.2.4 Brief discussion about feature selection

A good statistical point-of-view in feature selection for biometrical applications can be seen in (HEINZE; WALLISCH et al., 2018). A ML approach can be seen in (CAI; LUO et al., 2018) and (GUYON; ELISSEEFF, 2003). We opted not to use feature selection methods in our analysis for two main reasons. The first one is increased algorithm complexity and running time. The second one is that we want for the algorithm to select the best variables based on the optimization process. In Section 3.4, we detail the hyperparameters we used in our selected algorithm. We selected L_1 and L_2 regularization parameters to be optimized, and values for these parameters tend to shrink feature contribution, effectively working as a coupled feature selection mechanism inside our procedure, resembling the inner workings of LASSO (TIBSHIRANI, 1996).

2.4 Computational Results

Here we present the computational results of our work, divided into three parts. First, in 2.4.1, we analyze some data features of our problem, examining some variables already mentioned in other sections. In section 2.4.2, we use several algorithms with default parameters to select the best algorithm type to use together with Bayesian Optimization considering the hyperparameters to be tuned and their overall performance. In 2.4.3, we introduce the optimized models for both targets and discuss their results.

2.4.1 Data

Our dataset consists of laboratory test data collected from 9633 patients from the Sírio Libanês Hospital, who sought treatment in several different departments during the COVID-19 pandemic in Brazil. All patients from this list had a COVID-19 test (we included both positives and negatives), and 674 (7%) of them required special care treatment (hospitalization in common, semi-, or intensive care units). Among the ones requiring special treatment, the mean number of days needed for each patient was 1.52 days with a high variation, considering a standard deviation of 6.92 days.

There are 165 different types of laboratory test results (which in turn helps to understand the aforementioned **sparsity**). Considering demographics, the age and gender is available for each patient. Age will be analyzed further ahead in more detail.

We first show our exploratory analysis results in Table 3 considering some statistics for the dataset variables (for the ones with most coverage). We also show the two-sample Kolmogorov-Smirnov (KS) statistic value for each one considering special care target values as a class variable to understand the overall statistical difference between distributions that can arise between classes.

Table 3 – Variable metrics for the ones with most coverage within dataset (146 variables omitted).

	Mean	Std	Min	IQR	Max	Coverage (%)	KS Statistic
Sex	0.46	0.50	0.0	1.0	1.0	100.0	0.00
Age (*years*)	42.48	13.99	15.0	17.0	87.0	99.0	0.00
MCH (*pg*)	29.16	2.26	18.0	2.0	38.0	18.0	0.17
Hematocrit (%)	39.61	5.48	15.0	6.0	62.0	18.0	0.00
CMCH (*pg*)	33.09	1.23	27.0	2.0	37.0	18.0	0.00
Erythrocytes (*million/mm^3*)	4.06	0.80	1.0	1.0	7.0	18.0	0.06
Leukocytes (*/mm^3*)	6258.91	3541.01	100.0	3015.0	55110.0	18.0	0.00
RDW (%)	13.22	2.51	11.0	2.0	38.0	18.0	0.02
Hemoglobin (*g/dL*)	12.97	1.99	5.0	2.0	21.0	18.0	0.00
Platelets	205748.36	78948.08	7000.0	95000.0	529000.0	18.0	0.00
Neutrophils (%)	61.71	14.57	1.0	19.0	97.0	18.0	0.00
Eosinophils (*/mm^3*)	81.96	112.61	0.0	100.0	950.0	18.0	0.00
Monocites (%)	9.24	4.49	0.0	5.0	43.0	18.0	0.00
Eosinophils (%)	1.04	1.72	0.0	2.0	14.0	18.0	0.00
Lymphocytes (%)	25.75	12.38	0.0	16.0	84.0	18.0	0.00
Basofils (%)	0.07	0.30	0.0	0.0	4.5	18.0	0.19
Neutrophils (*/mm^3*)	4132.13	3142.68	20.0	2550.0	53730.0	18.0	0.00
Lymphocytes (*/mm^3*)	1463.58	841.17	20.0	920.0	14350.0	18.0	0.00
Basofils (*/mm^3*)	24.15	25.71	0.0	20.0	410.0	18.0	0.00
Monocites (*/mm^3*)	575.24	420.51	10.0	310.0	9170.0	18.0	0.00
Platelet Volume	9.85	0.92	8.0	1.0	13.0	18.0	0.10
Creatinine (*mg/dL*)	0.51	0.86	0.0	1.0	11.0	16.0	0.00
Urea (*mg/dL*)	34.71	18.32	10.0	14.0	201.5	16.0	0.00
Potassium (*mEq/L*)	3.54	0.55	2.0	1.0	6.5	15.0	0.00
Sodium (*mEq/L*)	138.42	3.05	121.0	3.0	152.0	14.0	0.00
ALT (*U/L*)	37.26	38.03	6.0	25.0	521.0	13.0	0.00
AST (*U/L*)	35.76	45.41	9.0	16.0	1140.5	13.0	0.00
DHL (*U/L*)	488.87	345.04	201.5	166.0	8958.0	11.0	0.00

Note: On Table 3, we use the KS statistic as measurement of difference for classes. A next step, should we consider to evaluate differences in groups (like white-series or red-series) is to use multivariate tests like *Hotelling T-squared* or *Wilks* Λ.

As pointed out in (BONANAD; GARCíA-BLAS et al., 2020), age seems to be a critical factor overall considering COVID-19 and the sample of the population we are considering.

As it is the only continuous demographic variable, we display the class histogram for age with adjusted kernels in Figure 2. We see a very distinct separation between classes arising for each one of the groups. Moreover, this pattern by itself is not substantial in terms of making any assumptions or conclusions about our targets.

In Figure 3, we see the histograms and adjusted kernels for selected white blood cell components count, which superficially represents immunological responses for each one of the patients in data and also mentioned as necessary by other authors investigating samples coming from similar conditions, as mentioned earlier. By close inspection, we see that separation for the variables considering the classes is not evident using only univariate reasoning, which again points to the necessity to use multivariate and non-linear algorithms.

This brief analysis shows a perfect match for ML applications: We have sufficient patient data, with no identifiable univariate patterns relating to our target, thus opening up the possibilities of multivariate analysis and algorithms recognizing several different types of trends and interactions (the aforementioned **non-linearity**).

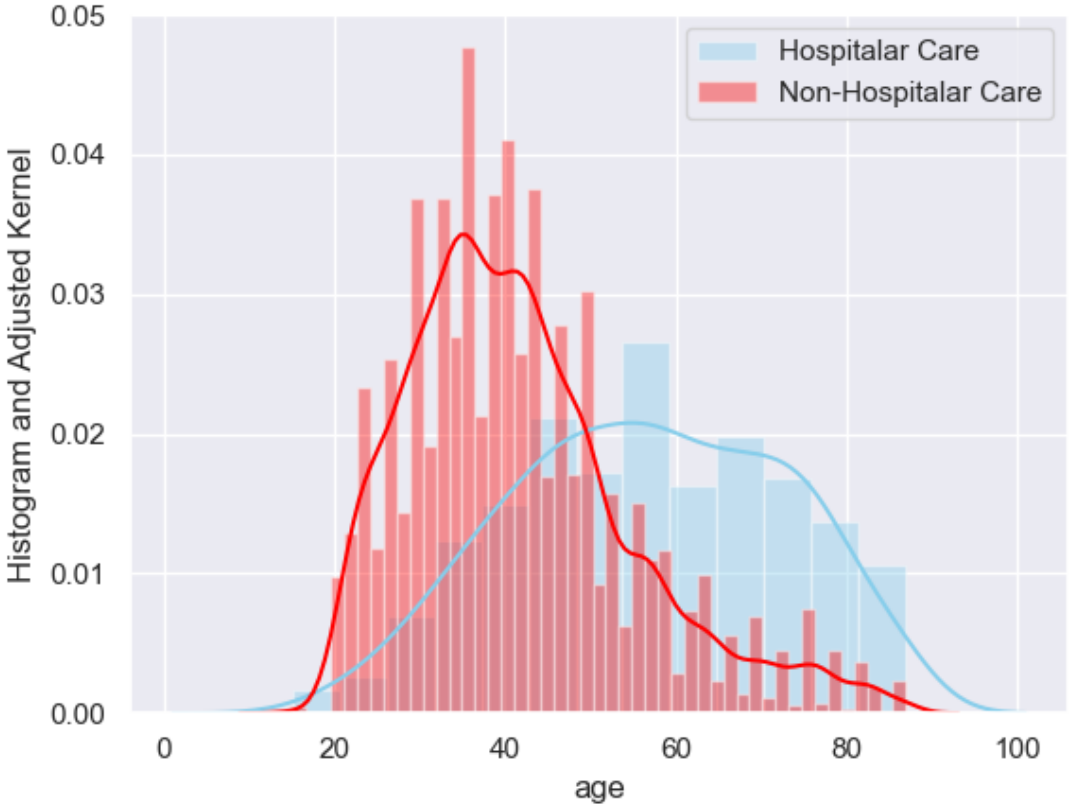

Figure 2 – Histogram and adjusted kernels for age, divided using the special care target.

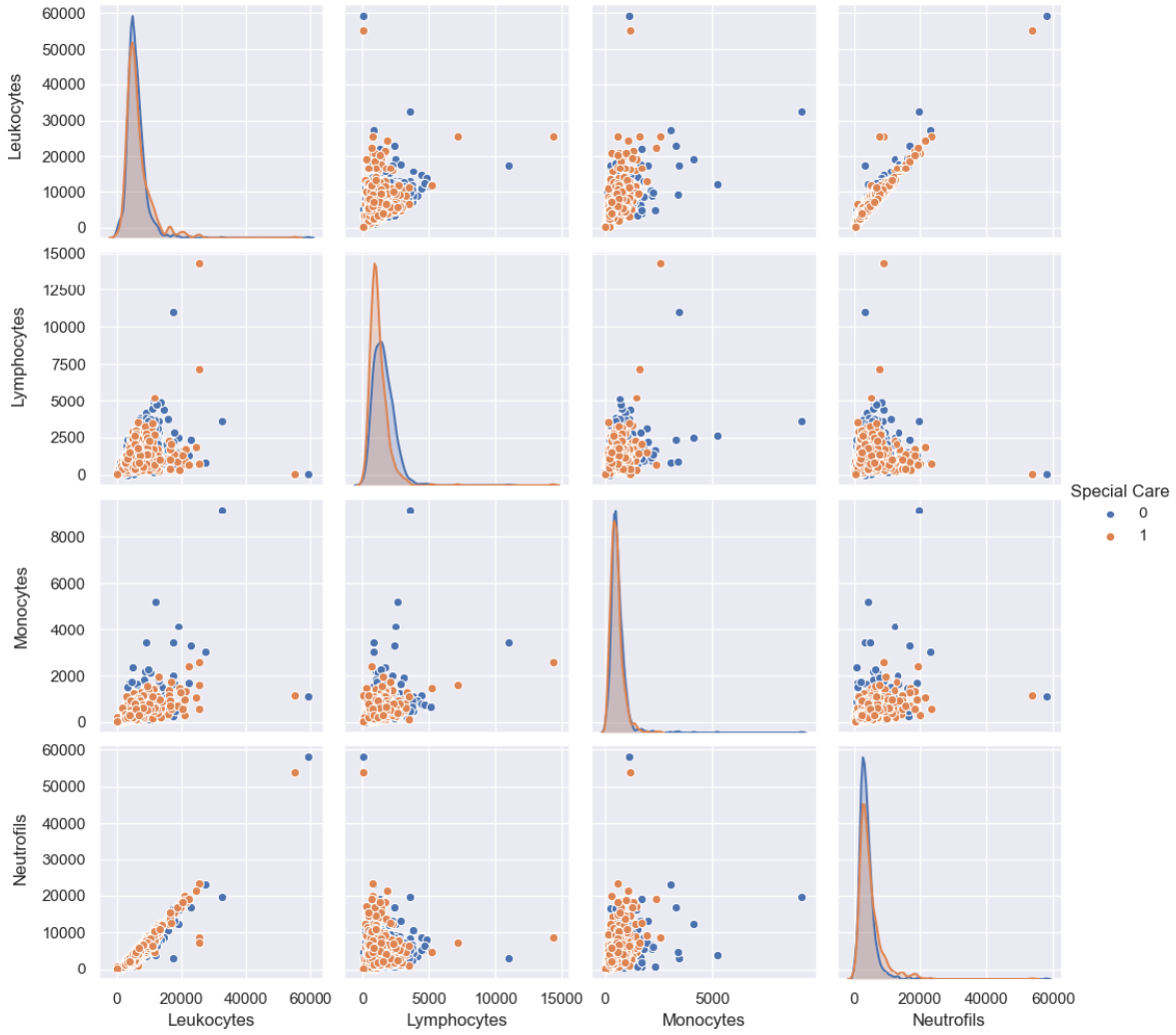

Figure 3 – Histogram and adjusted kernels for white-cell blood components, divided using the special care target.

2.4.2 Preliminary Models

To begin our modeling, we used several ML algorithms without tuning the parameters to select the best algorithm type to be optimized later. Our tests considered Naïve Bayes, Decision Trees, AdaBoost, Support Vector Machines (SVD), Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Logistic Regression (and regularized ones such as Ridge Regression and Least Absolute Shrinkage and Selection Operator (LASSO)), Orthogonal Matching Pursuit (OMP) and other algorithms based on ensembles of trees, such as Extra Trees (GEURTS; ERNST; WEHENKEL, 2006), Random Forests (BREIMAN, 2001), xgBoost (CHEN; GUESTRIN, 2016), and LightGBM (KE; MENG et al., 2017). All results were obtained using Python 3.7 as our programming language. To obtain the following results, data were treated as-is, i.e., without any treatment or imputation strategies.

Model type selection for further optimization should consider three critical practical aspects, emphasizing the first two. The first one is predictive power - we want an algorithm that

predicts well and does not overfit our data while capturing the multivariate effects that we expect. The second aspect involves the number of hyperparameters available to tune the model. The more parameters, the more opportunities we have to improve our algorithm predictive power while keeping the generalization capacity. The third reason is the training time - which, although less important, generates problems when the datasets are large enough and which can be considered even within our context because the algorithm requires several full training passes through our data when considering the optimization process. Table 4 presents results considering algorithms for the special care target and all relevant metrics. The baseline for this model is a coin with a ROC AUC value of 0.5. Table 5 presents results and relevant metrics for the number of days under special care target. The baseline here is the mean value of the training set.

Table 4 – Results from preliminary models on special care target (Top 10 of all models tested). Chosen algorithm for optimisation is highlighted.

Model	Balanced Accuracy	ROC AUC	F1 Score	Time Taken (s)
Bernoulli Naïve Bayes	0.90	0.90	0.92	0.14
QDA	0.88	0.88	0.91	0.22
Gausssian Naïve Bayes	0.85	0.85	0.95	0.15
xgBoost	0.85	0.85	0.96	1.31
LightGBM	0.82	0.82	0.96	0.47
AdaBoost	0.82	0.82	0.96	0.92
SVC	0.81	0.81	0.95	2.52
Random Forest	0.81	0.81	0.96	1.14
Baging	0.80	0.80	0.96	0.78
Decision Tree	0.80	0.80	0.96	0.23

Table 5 – Results from models on number of days of special care needed (Top 10 of all models tested). Chosen algorithm for optimization is highlighted.

Model	R-Squared	RMSE	Time Taken (s)
xgBoost	0.70	2.15	1.28
Vanilla Gradient Boosting	0.68	2.22	1.98
Random Forest	0.66	2.31	7.67
Bagging	0.64	2.38	0.92
LightGBM	0.60	2.49	0.36
Extra Trees	0.60	2.50	9.65
Histogram Gradient Boosting	0.60	2.52	4.75
Huber Regression	0.45	2.94	1.70
LinearSVR	0.44	2.96	3.14
Decision Tree	0.43	2.99	0.22

The final selected algorithm is xgBoost for both targets. The primary rationale for this is the characteristics mentioned above: high predictive power, hyperparameter tuning, and

overall training time. We could also select LightGBM interchangeably as the results were very close (and the algorithms are similar). Moreover, it was faster. Between the two algorithms, our previous experience with xgBoost motivated us to choose it. Algorithms such as the Naïve Bayes one stand out as they have almost no hyperparameters to tune and were unconsidered, even performing very well in the preliminary analysis.

2.4.3 Optimized Models

Having selected the final algorithm type to use, we must define which hyperparameters to use in Bayesian Optimization and which strategy to deal with sparsity and unbalancing. Table 8 shows all parameters considered in the Bayesian Optimization and its respective intervals and descriptions. All optimization is performed using Ax (BAKSHY; DWORKIN et al., 2018), a platform created inside Facebook that streamlines all optimization processes and makes it possible to use integer hyperparameters, which are not available in other solvers.

Table 6 – Parameter grid and intervals used in Bayesian Optimisation procedure.

	Interval	Description
eta	$[0.01, 1]$	Learning rate (shrinkage applied in weights calculation)
gamma	$[0, 100]$	Minimum loss reduction to split a node in tree
max_depth	$[1, 9]$	Maximum depth of each tree in training process
subsample	$[0.5, 1]$	Number of features used to train a tree
lambda	$[1, 100]$	L_2 regularization term using in training
alpha	$[0, 100]$	L_1 regularization term using in training
n_estimators	$[10, 200]$	Total number of trees

For a classification model to be useful, we need to analyze Receiver Operating Characteristic (ROC) curves and Precision-Recall (P/R) curves, which can be a different format considering the variable distribution. Figure 4 summarizes the ROC curve and Figure 5 summarizes the P/R curve. Using the median as imputer in our tests gave us the best results overall for the special care target.

It can be observed that our optimization improved the ROC statistic by selecting a new set of hyperparameters different from the defaults. By doing that, we guarantee that we have the best model while keeping model generalization capabilities.

From a hospital perspective, False Positives (the abscissa from our ROC plot) constitutes the most lost resources. They are patients that do not need any special care, but the model indicates the opposite, and we should keep them on a minimum level. We see by close inspection of the curves that this is satisfied, and the model is indeed useful for classifying patients using blood-test samples. At the best threshold value for cutoff, we obtained 0.94 for ROC AUC and 0.77 for P/R AUC.

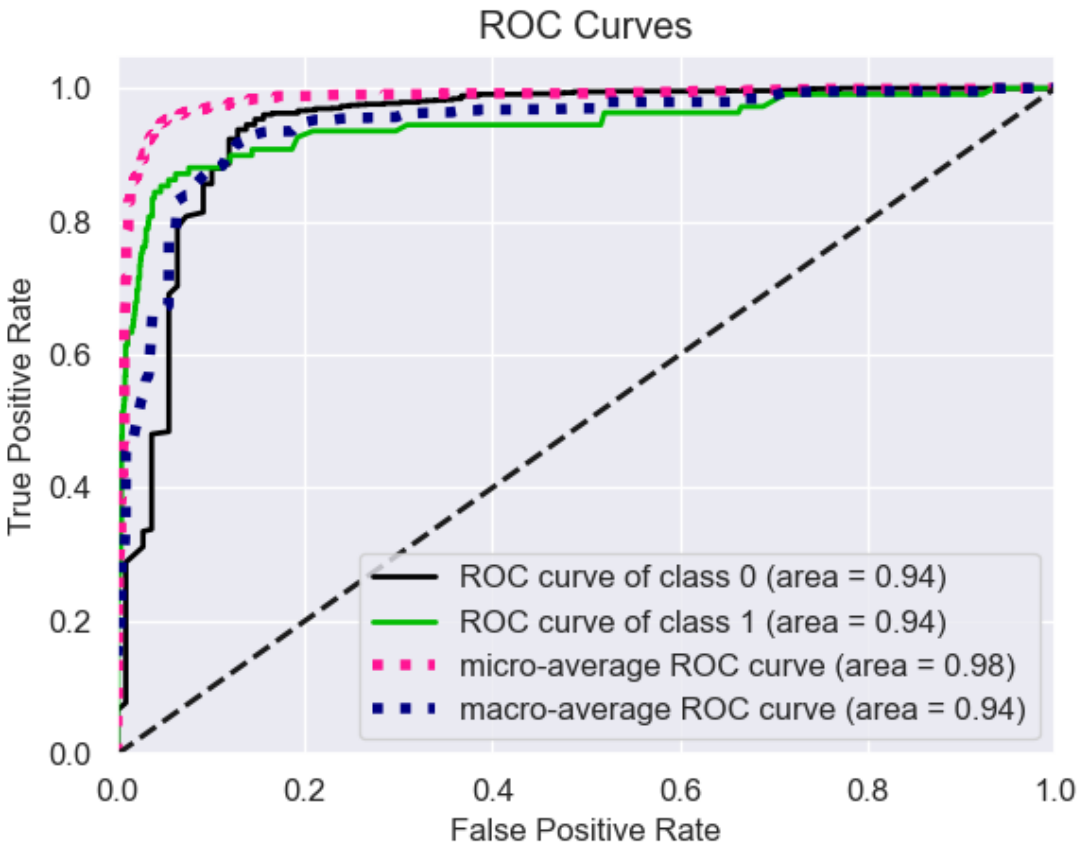

Figure 4 – ROC Curve for special care target, both classes.

Moreover, as we used ensembles of trees to make predictions, one thing that arises naturally is a variable importance plot. To obtain this plot, we used Shap (LUNDBERG; LEE, 2017), which creates this plot using a game-theoretical approach to calculate the variable importance for row and data levels. In Figure 6, it can be observed that some of the variables presented as important (mentioned in Section 3.3) in (FERRARI; MOTTA et al., 2020) and (YUAN et al., 2020) are indeed some of the most relevant in our model, which are in line with the expectations (This plot should not be seen as indicating any direct causal relationships as our data is not experimental, but observational).

Results for the days under special care were similar in performance achievements. Table 7 summarizes the findings and compares them with the baseline for this model, the mean value of days spent in special care for the training set. Best results were obtained using no imputer at all (using model-based input gave us the worst results in comparison), defying some preconceptions we had from the start. This effect is explained in (HASTIE; TIBSHIRANI; FRIEDMAN, 2009): adding variables to boosted or bagged regressors can make the model worse. Using imputers, we forced the model to be non-sparse, giving protagonism to all variables at once, amplifying this condition. The condition for classifiers is the opposite: adding variables to boosted or bagged models always increases the performance (but the improvement could be

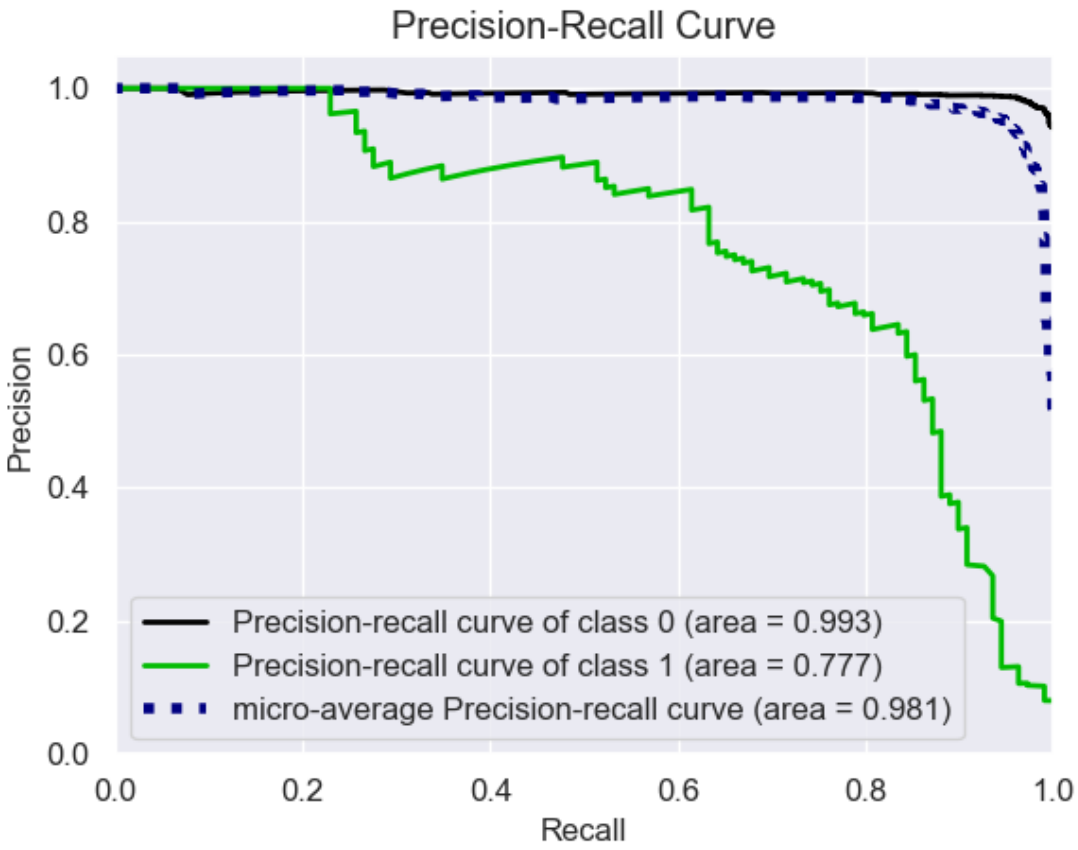

Figure 5 – Precision-Recall Curves for special care target, both classes.

marginal).

Although our model is capable of making good predictions as guaranteed by statistical tests, in Figure 7 we see a tendency to *overshoot* and *undershoot* the results caused by the very nature of the model (splits in trees have a very poor tendency in addressing extreme situations as the capacity to extrapolate wanes as we go to the ends of our interval). A more in-depth discussion on model improvement can be found in Section 3.5.

Table 7 – Results for days under special care target, baseline and percentual improvement over baseline.

	Model	Baseline	Improvement (%)
RMSE	1.87	3.96	77.78
MAE	0.41	1.27	67.96
R-Squared	0.78	0.00	–

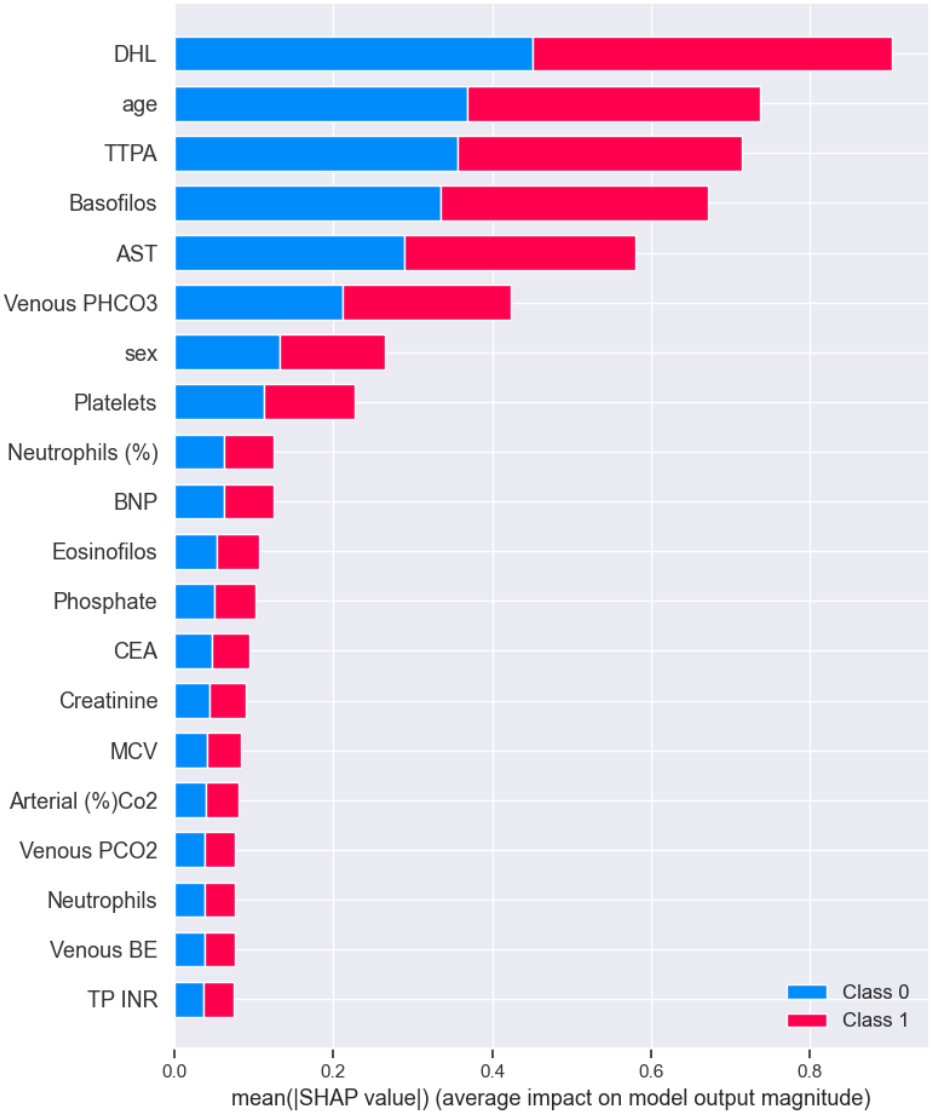

Figure 6 – Variable importance plot for special care target, both classes.

2.5 Limitations and Possible Extensions

So far, with our classification model we have only dealt with 0/1 outcomes. But what happens if we want to order our patients according to their risk (risk being associated with a measure of probability ranging from 0 to 1)? The algorithm used to learn our special care from data is not well suited for this specific task. In (NICULESCU-MIZIL; CARUANA, 2005), this effect is described as the algorithms having difficulties making predictions near the frontiers of the [0,1] interval because the variance of the base trees drives the result away from the edges in a way to minimize the overall cost function. To diagnose this problem, one can calculate the overall Brier score (BRIER, 1950) for a given model or make a calibration plot. To solve this issue, we could apply Platt's method (SMOLA; BARTLETT, 2000), which essentially adjusts a Logistic Regression on a different fold during the model training phase or use an Isotonic Regression (WU; WOODROOFE et al., 2001), again on a different fold during model training.

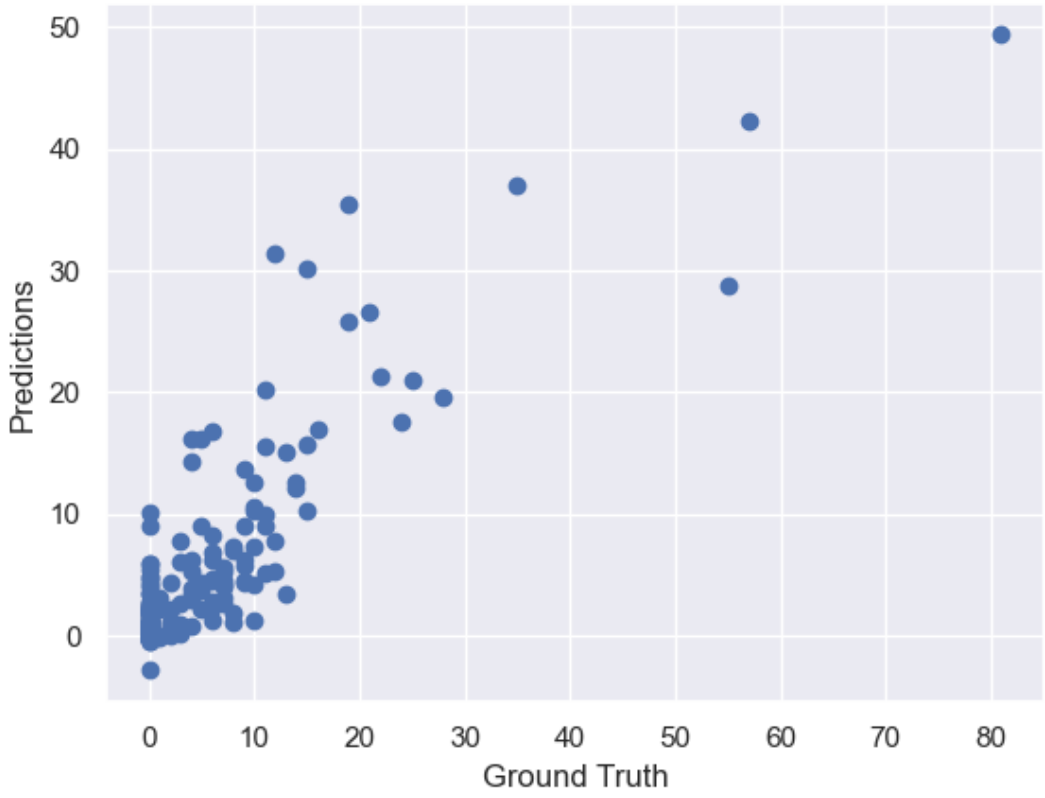

Figure 7 – Scatterplot for days under special care target.

However, more data for patients is required to perform that in a meaningful way.

To deal with negative predictions arising in the number of days under special care targets, we must first understand that the model used to make the predictions is not restricted in any form about the prediction interval itself. All of its predictions lie within the real line $\mathbb{R}$, but we know that our values are at least limited by 0. A recent way to deal with this is emerging in disciplines such as Finance and Banking, presented in (SIGRIST; HIRNSCHALL, 2019) where ensembles of trees are trained to perform the Tobit regression. The overall maturity for the packages is increasing fast, posing as an exciting development as ensembles of trees have very high predictive power in general and several hyperparameters that can be optimized using Bayesian Optimization in the same process.

To deal with *overshooting* and *undershooting* for our number of days under special care targets, several possibilities are arising from traditional statistics worth exploring such as the Zero Inflated Negative Binomial (ZINB) models (HALL, 2000) in which the target distribution comprises a very high proportion of zeroes, such as our target. The result for this type of model usually consists of a probability attached to a counter, probability measuring the overall chance of a given patient needing special care, and the counter giving the number of days the same patient will spend under such care. The major drawback for this from the model is the predictive

power (especially for the probability part), where standard packages use only linear terms (which introduce needs on data pre-processing, such as multicollinearity removal or variance inflation factors analysis) and no ensembles to make predictions. A viable but not tested alternative could be mixing two "worlds," trying different sets of variables on the dataset guided by Bayesian Optimization, and then applying a ZINB model for each one, averaging the results. The counting model in this situation is discrete, also solving the issue with non-integer predictions.

2.6 Final remarks

The growing necessity to predict hospital resources' needs guided the exploration of novel methods to create and plan policies accessible for everyone. More than ever, the COVID-19 pandemic is pushing health systems to the limit. Having this in mind, we developed an analytical approach based on mathematical models and algorithms adopting the most recent techniques available in the fields of statistics and machine learning using public data available online.

We obtained promising results in this study. The estimated 0.94 area under the ROC Curve combined with 0.77 P/R statistic proves that the analytical approach can indeed be used in a decision system for hospitals, governments, and health providers alike to guide their resource allocation with minimal requirements as we use test data that is available and affordable. The target for the number of days under special care certainly needs refinement but is adequate in our view. Other interesting results are also in line with other studies conducted by researchers all around the world.

Our biggest contribution was standardizing a method to create decision systems/ML models that can be applied to several different diseases, with low processing requirements, using cheap datasets that can be collected and analyzed easily. Our method also allows for suitable customization in the methods used and also for other infectious diseases.

3 Using bi-dimensional representations to understand patterns on COVID-19 blood exam data

Blood tests have an essential part in everyday medicine and are used by doctors in several diagnostic procedures. Still, this data is multivariate – sometimes, diseases like COVID-19 could have different manifestations and outcomes. This study proposes a method of extracting useful information from blood tests using UMAP - Uniform Manifold Approximation and Projection for Dimension Reduction combined with DBSCAN clustering and statistical techniques. The analysis indicates several clusters of infection prevalence varying between $2-37\%$, meaning that our procedure is indeed capable of finding different patterns. A possible explanation is that COVID-19 is not just a respiratory infection but a systemic disease with critical hematological implications, primarily on white-cell fractions, as indicated by relevant statistical tests p-values in the range of $0.03-0.1$. The novel analysis procedure proposed here could be adopted in other data-sets of different illnesses to help researchers to discover new patterns of data that could be used in various diseases and contexts.

3.1 Introduction

COVID-19 (Coronavirus Disease) came under intense scrutiny worldwide throughout 2020-2021. Some countries are already seeing hospitalization rates going down due to mass vaccination campaigns, social distancing, and lockdown measures. In a sad turn of events, Brazil is one of the biggest economies still witnessing death and hospitalization rates which were high at the beginning of 2021 (especially in the North region), according to Johns Hopkins Coronavirus Resource Center (JHCRC, 2021). More than ever, humanity should use all the tools available to understand the infection scientifically.

In this study, we propose an exploratory data analysis using the bi-dimensional representation generated by UMAP - Uniform Manifold Approximation and Projection for Dimension Reduction (MCINNES; HEALY; MELVILLE, 2020), followed by a DBSCAN clustering and posterior usage of statistical tests on the clusters obtained to reveal (non-causal) links between different parameters of blood-test data and their diagnostic counterparts.

Data was obtained from the Albert Einstein Hospital in Sao Paulo, Brazil (HIAE, 2020). The data consist of patients´ blood tests, providing information about whether or not a given patient has COVID-19 and if the patient needed special care or not (hospitalization in standard, semi, and intensive care units).

This article is organized as follows: in Section 3.2, we examine the most up-to-date literature regarding Machine Learning, which is used to model blood test results and explain their results. In Section 3.3, we present the method used to perform the two data experiments, revealing their results in Section 3.4. In Sections 3.5 and 3.6, we outline the results and discuss the limitations and possible implications of this study.

3.2 Literature Review

In this section, some studies from the literature are presented that inspired and laid the authors' foundations to create their analysis and perspective. Some interesting results were obtained without using machine learning (ML) and should be encouraged as a first-line open-access tool available to most researchers. In (FERRARI; MOTTA et al., 2020), it can be observed that statistically significant differences were found using two-way tables based on blood test data from a hospital in Italy, which is a quick and cheap solution to detect infections. A new study is (LIAO; ZHOU et al., 2020), which is much more focused on hematological data and sheds light on significant statistical differences and possible risk factors associated with different patients. One specific meta-analysis, including the results of 35 other studies (BAO; LI et al., 2020), indicated factors that contribute the most to non-severe patients to develop severe diseases.

Using blood tests with machine learning seems to have gained traction since the beginning of the pandemic. Theoretical justification and groundwork for supervised ML techniques

can be observed in several articles. In (BRINATI; CAMPAGNER et al., 2020b), attention is paid to possible combinations of models that could be used with results varying between the values of 0.6 - 0.9 area under the Receiver Operating Characteristics Curve (ROC) to detect infected patients. In (BATISTA; MIRAGLIA et al., 2020), similar results were obtained using the same dataset we adopted in this study. Amalgamating the results of these articles and some others, there is (ALJAME; AHMAD et al., 2020) which uses ensembles and achieves 99.88% accuracy in predicting infections.

Other articles with similar but not identical purposes are available. (BEZZAN; ROCCO, 2020) uses several ML models on a dataset provided by the Sírio Libanês Hospital, in Brazil, to predict special-care probability and the number of days under special care, obtaining a value of 0.94 area under the ROC curve for the first target. In (BARBOSA; GOMES et al., 2021), we see a prime example of how a system could be implemented to detect COVID-19 in a given patient. This study also stands out as it uses a small sample and optimization techniques to find the most important variables for the problem.

As an example of unsupervised learning techniques, an article that can be mentioned is (KUKAR; GUNČAR et al., 2020), which uses a model to predict infection and compares COVID-19 manifestations with other diseases using t-distributed stochastic neighbor embedding (similar to the purpose of UMAP), concluding that blood parameters of those affected with severe COVID-19 resemble more bacterial than viral infections, which was a very surprising result.

The main difference between this study and the others mentioned above is the fact that we are not pursuing the creation of a fully supervised model. Instead, we aimed to test the "manifold hypothesis" on this data to check the existence of different groups where the manifestations of the disease could be different, providing researchers a whole new set of techniques to apply in other datasets in a similar context.

The use of clustering techniques is widespread in medical sciences in general. In a first class of articles, patient characteristics are used to unveil some hidden data structure present for diagnosing or understanding the disease's progression, such as (MCLACHLAN, 1992) and (SKERMAN; YATES et al., 2009). Another class of articles tends to use more comprehensive statistical analysis with clustering to separate manifestations and possible patterns arising in a more specific group of patients as in (PAUL; SAYED, 2010) and (ALASHWAL; HALABY et al., 2019).

Although our article offers non-causal inference, it is relevant to point out sources (OLTEAN; GAGNIER, 2015) that mix up causal inference and clustering in a medical setting, something we believe that should be further explored if any other dataset allows us to do so.

3.3 Method

The procedure behind our analysis primarily consists of two phases. In the first phase, we project high-dimensional laboratory exam data into a two-dimensional subspace using UMAP (tuning two hyperparameters), making the dataset more amenable to clustering techniques. In the second step, we cluster the data representation using DBSCAN (SCHUBERT; SANDER et al., 2017) to find any patterns that may arise. The number of clusters obtained is a consequence of the hyperparameter tuning method used. Here, we used DBSCAN as a clustering alternative because the number of clusters is not specified upfront. By doing that, we assume more neutrality when analyzing the data structure.

The "overall quality" of fit for a specific combination of hyperparameters is measured without resorting to the target´s current value, using the silhouette coefficient for a given arrangement (J.ROUSSEEUW, 1987). We then compare different arrangements using this metric, selecting the one with the maximum value overall. Table 8 summarizes all hyperparameters used in the cluster tuning procedure.

Table 8 – Parameter grid and intervals used in the clustering procedure.

Parameter	Interval	Description
neighbors	$[1, +\infty)$	Balance between local and global data representation
spread	$[0, +\infty]$	Minimum distance allowed between points in representation
eps	$[0.01, 0.5]$	Maximum neighborhood distance in DBSCAN

As data science researchers know, high-dimensional data has fewer degrees of freedom than one might initially assume, which is known as the "Manifold Hypothesis". (FEFFERMAN; MITTER et al., 2016) manifold presents a complete description of the hypothesis and several demonstrations on the subject. A good way to visualize the dimensional reduction performed by UMAP is by comparing Figures 8 and 9. Figure 8 shows elements of the so-called MNIST dataset (LECUN; CORTES, 2010), which is composed of 28x28 pixel images (784 dimensions) of thousands of handwritten digits. In Figure 9, after the UMAP algorithm, we can see that similar points tend to cluster closely, and non-similar digits tend to be more distant. The overall distance is controlled by the parameters´ neighbors and spread in Table 8.

The hypothesis and dimensional reduction provided by UMAP allows us to analyze blood test data with a new perspective: different groups with different manifestations of the disease could be traced using this technique, as these groups will tend to cluster together in the low-dimensionality representation. Moreover, more significant factors could give us some clues about the disease and its progression.

We then propose two “experiments”. In the first, we analyze data from all patients in our dataset with measurements of blood tests (red and white series) and then use the procedure outlined above. In the second one, we filter out our patient data keeping only those with confirmed

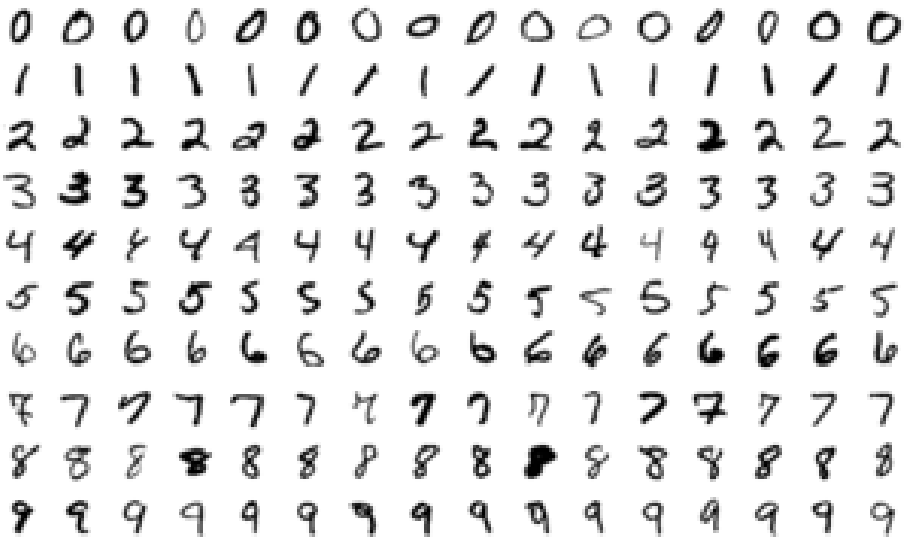

Figure 8 – MNIST data (LECUN; CORTES, 2010) examples. Each example is a 28x28 pixel image.

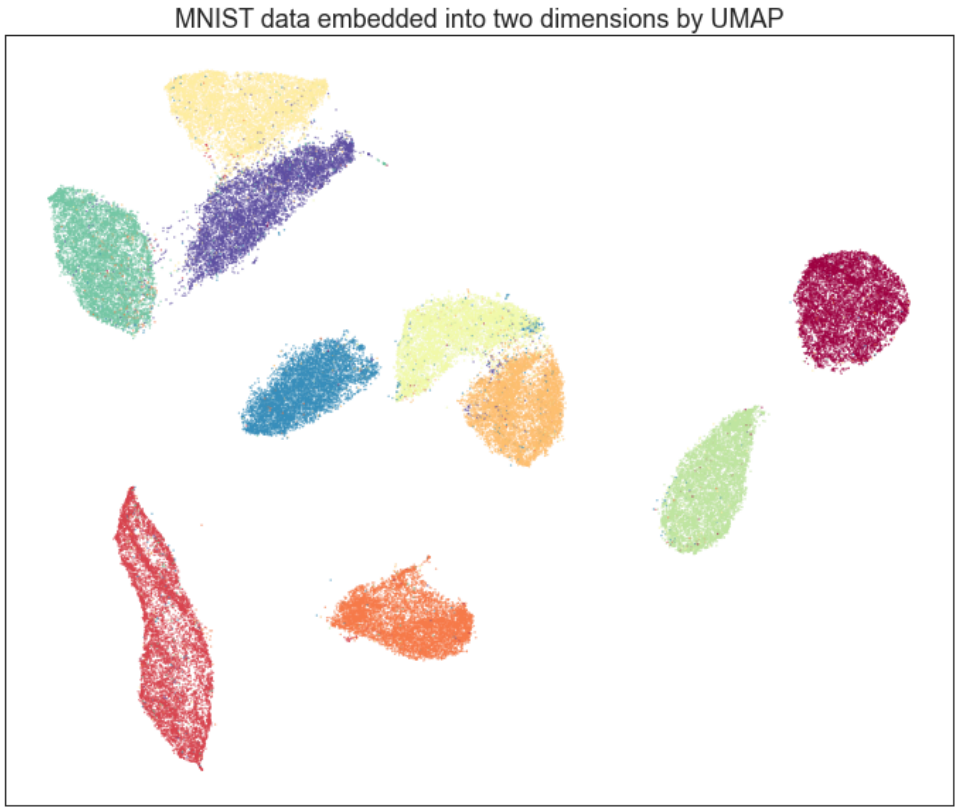

Figure 9 – UMAP dimensionality reduction results on MNIST data. Each one of the colors represents a different number. (the coordinates were omitted).

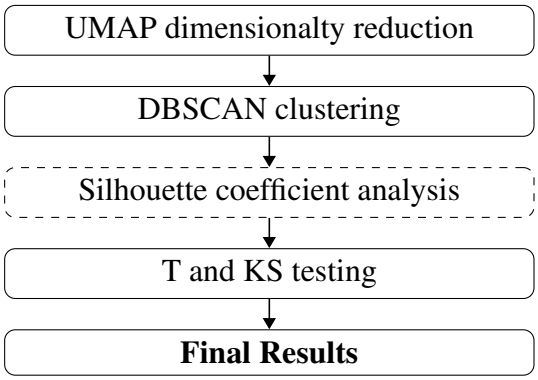

Figure 10 – Steps synthesizing our method for both experiments proposed

COVID-19 and comparing the results using the targets for both situations. It is worth mentioning at this point that none of our analysis aims to be causal. The study was not conceived in this way, and the data are observational. For this purpose, we suggest using Causal Forests (WAGER; ATHEY, 2018), which can deal with observational data and make a satisfactory causal inference whether the number of samples is high enough as the method needs several data splits.

Figure 10 summarizes all the steps we used in both experiments. Silhouette coefficient is used to select the number of clusters for our experiments prior to statistical analysis.

3.4 Computational Results

3.4.1 Data

The data contains anonymous information about 598 patients admitted to the Albert Einstein Hospital during the COVID-19 pandemic. 81 patients tested positive for infection (13%) and 128 patients needed special care treatment (21%, not only related to COVID-19). There are available parameters related to red and white cell counts for each patient, all of them normalized by the mean and standard deviation (z-scores). Table 9 summarizes all the variables used for the study.

Table 9 – Variables used for study.

Fraction	Components
Red Cell	Hematocrit, Hemoglobin, Red Cells, MCHC, MCH, MCV, RDW
White Cell	Platelets, MPV, Lymphocytes, Leukocytes, Basophils, Eosinophils, Monocytes

To further expand on the data, Figure 11 presents white cell distribution for all 598 patients (blue dots negative, orange dots positive infection). No univariate pattern was observed emerging in the data, which leads us to use a multivariate technique.

As mentioned above, two data experiments were performed: The first experiment consists of all 598 patients and tries to understand if there are groups with high prevalence (greater than the average of the dataset) and to point out the main characteristics of these groups. In the second experiment, the focus is primarily on the confirmed COVID-19 diagnostic, aiming to discover any groups with more prominent special care needs than the whole dataset.

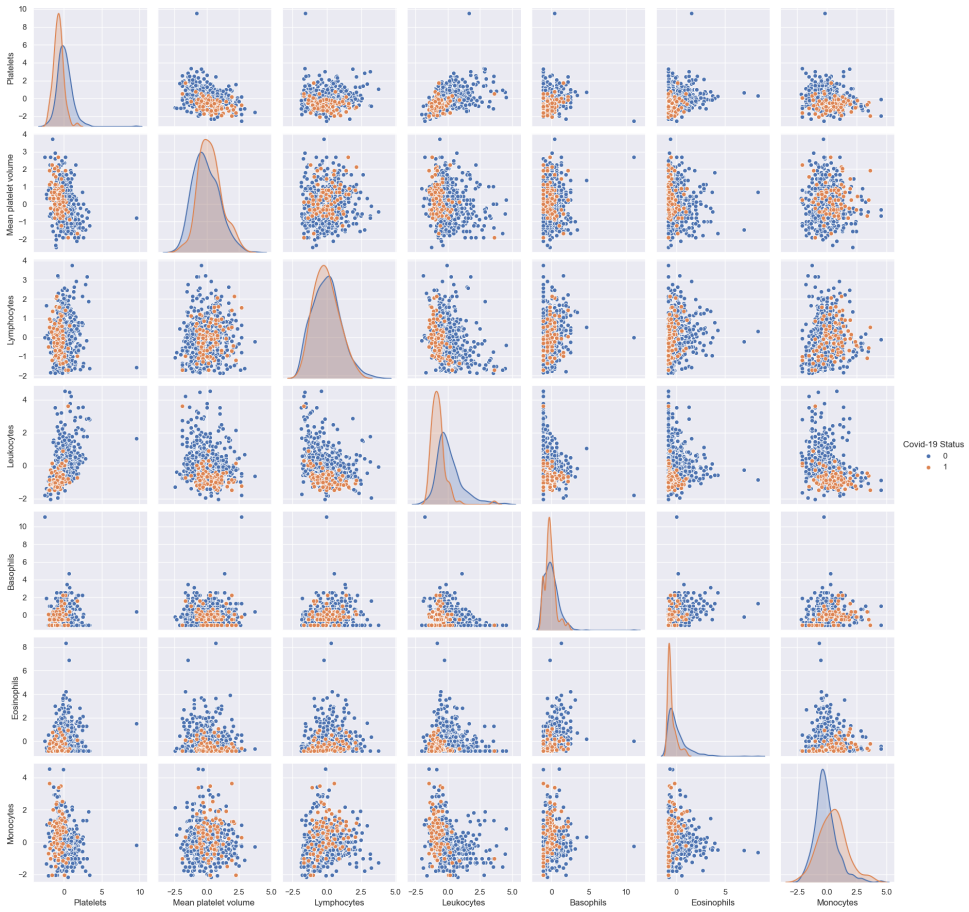

Figure 11 – White cell blood count distributions, normalized for 598 patients.

3.4.2 Experiment I: All patients, focus on the confirmed COVID-19 results

In this first analysis, after performing the aforementioned dimensionality reduction with UMAP and the clustering of the resulting 2-dimensional space variables, we obtained a value of 0.12 for the silhouette coefficient (the clusters obtained are very packed together). Overall, 7 clusters were obtained, with COVID-19 prevalence in the range of $3 - 35\%$. 29 patients did not meet any of the DBSCAN similarity criteria and were not assigned any cluster, thus they were removed from the analysis.

Close inspection of Tables 10 and 11 reveals that most extreme values reside on the first two clusters for white-cell counts. This fact could be interpreted in a two-way manner: Patients could have comorbidities and be more susceptible to being infected by COVID-19, thus having greater white-cell counts, as pointed out by (SOUZA; BUSS et al., 2020). On the other

hand, COVID-19 could be responsible for the values themselves. One observation is about the number of platelets, which is very low, much in line with discoveries shown in (G.D.; J.L., 2021), (GüçLü; KOCAYIğIT et al., 2020), (BATTINELLI, 2020) and (MEI; LUO et al., 2020).

No extreme values were found in red cell samples for high COVID-19 prevalence clusters, but the close observation of the tables regarding the prevalence and the number of people in each cluster may help to "name" each cluster, a procedure that is made when clusters are applied in several contexts. For example, cluster 1 could be named "Non-symptomatic patients", although more data is needed to make such an affirmation.

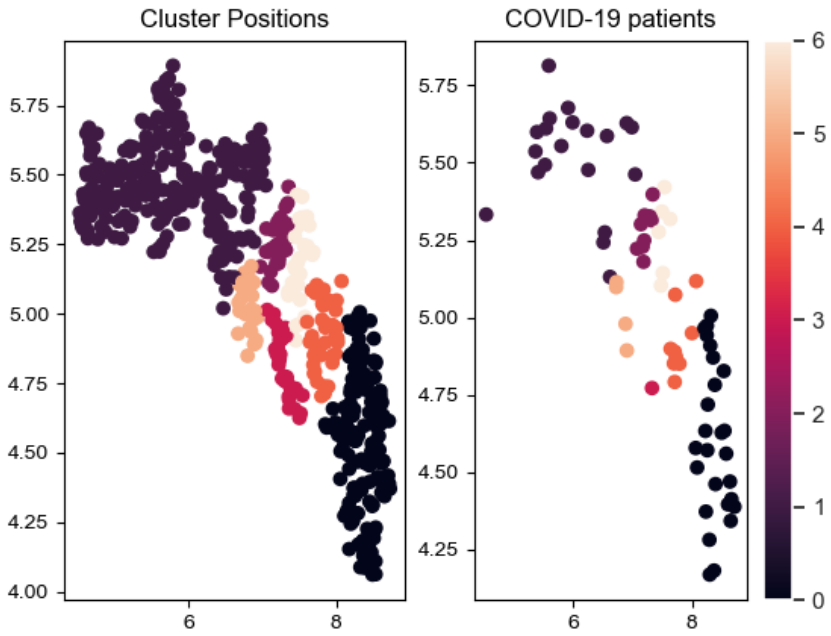

Figure 12 – DBSCAN cluster results for Experiment I. On right, all COVID-19 patients with clusters associated.

Table 10 – Means for variables in clusters found in experiment I (Red components - extreme values in bold).

Cluster	Hematocrit	Hemoglobin	Red Cells	MCHC	MCH	MCV	RDW	**Covid-19** (%)	Patients
2	0.449555	0.360825	0.403754	-0.219273	-0.129423	-0.025629	-0.192997	34.6	26
4	0.331591	0.353596	0.177950	0.187976	0.259933	0.197007	-0.155573	23.1	39
6	**0.890685**	**0.947817**	**0.910157**	**0.404758**	-0.046087	-0.249906	**-0.234152**	19.4	31
0	-0.123704	-0.160615	-0.269449	-0.167212	0.249553	0.363449	0.330257	17.9	145
5	**-0.566416**	**-0.606294**	-0.369784	**-0.313489**	**-0.400994**	**-0.312915**	**0.680545**	16.0	25
1	-0.015429	0.021216	0.056257	0.141979	-0.079685	-0.156664	-0.216660	7.4	269
3	-0.285210	-0.324563	**-0.527503**	-0.212740	**0.398023**	**0.565605**	-0.133359	2.9	34

Table 11 – Means for variables in clusters found in experiment I (White components - extreme values in bold).

Cluster	Platelets	MPV	Lymphocytes	Leukocytes	Basophils	Eosinophils	Monocytes	**Covid-19** (%)	Patients
2	**-0.566694**	0.092664	**0.365603**	**-0.408745**	**0.880585**	-0.018652	**0.227241**	34.6	26
4	-0.327375	-0.262615	-0.127550	-0.291689	-0.231599	**-0.303903**	0.100975	23.1	39
6	0.244400	-0.376571	0.014347	0.068407	0.130960	**0.072528**	-0.105025	19.4	31
0	-0.129383	**0.287677**	-0.154026	-0.031201	0.037454	0.068019	0.019385	17.9	145
5	-0.108903	-0.016250	**-0.803715**	**0.207712**	**-0.419260**	-0.301180	-0.465017	16.0	25
1	0.115441	-0.031031	0.160436	0.065219	-0.026183	0.058192	-0.004671	7.4	269
3	**0.555883**	**-0.477694**	-0.118372	0.242435	-0.133926	0.023392	**-0.575570**	2.9	34

3.4.3 Experiment II: COVID-19 patients, focus on special care

In this analysis, we obtained a value of 0.40 for the silhouette coefficient (the clusters obtained seem very separated, as shown in Figure 13). Overall, two clusters were obtained, with COVID-19 prevalence in the range of $7-61\%$. No patients without clusters were obtained in this analysis.

The number of clusters obtained allows us to go one step further in the analysis. We conducted two-sample one-sided (lower) t- and KS- statistical tests. Tables 12 and 13 show the p-values associated with one of these tests in every parameter. The result is very similar to Experiment I. Red cell components do not display any statistical differences between the two groups, however white cell components show statistical differences. Again, platelets appear as a significant factor, once again indicating a relationship between coagulation factors, COVID-19 and a possible patient prognostic.

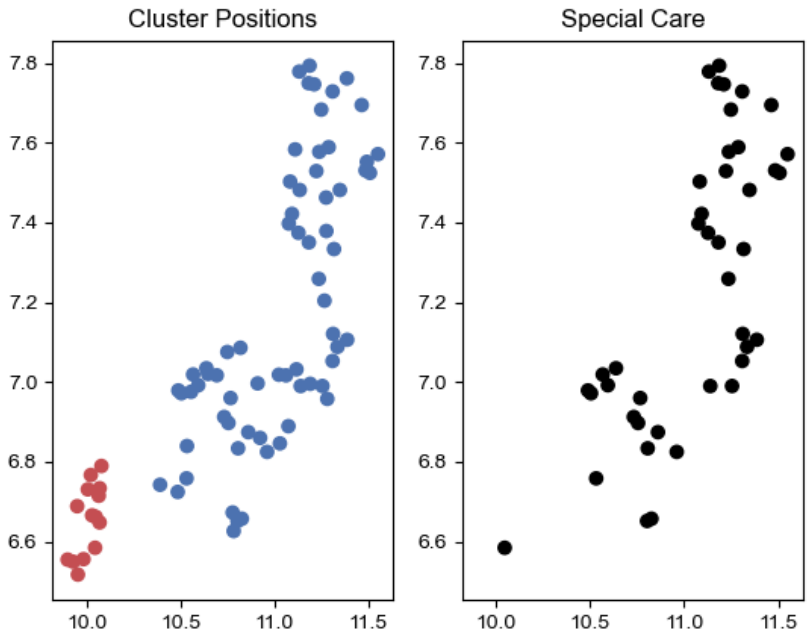

Figure 13 – DBSCAN cluster results for Experiment II. On the right, all special-care patients.

Table 12 – Means for variables and respective t and KS tests for clusters found in experiment II (Red components - no significant p-values in bold).

	Hematocrit	Hemoglobin	Red Cells	MCHC	MCH	MCV	RDW	Special Care (%)	Patients
Mean - Cluster 1	0.192373	0.228284	0.124672	0.187246	0.152039	0.078920	-0.227019	7.0	14
Mean - Cluster 2	0.276826	0.302162	0.261730	0.166864	0.034623	-0.037691	-0.194673	61.0	67
t-test	0.638796	0.619572	0.701361	0.466539	0.285301	0.295333	0.562766	-	-
KS-test	0.440488	0.675420	0.788581	0.458707	0.284728	0.348343	0.863925	-	-

Table 13 – Means for variables and respective t and KS tests for clusters found in experiment II (White components - significant p-values in bold).

	Platelets	MPV	Lymphocytes	Leukocytes	Basophils	Eosinophils	Monocytes	Special Care (%)	Patients
Mean - Cluster 1	-0.445631	0.331228	0.063713	-0.537869	0.016237	-0.305755	0.858424	7.0	14
Mean - Cluster 2	-0.734901	0.263530	-0.049911	-0.741464	-0.205530	-0.516632	0.406545	61.0	67
t-test	**0.061341**	0.399595	0.331979	0.150230	0.156617	**0.056762**	**0.088217**	-	-
KS-test	**0.034455**	0.689187	0.272100	0.564482	0.284728	**0.030776**	0.105875	-	-

Note: On Table 13, we use the KS statistic as measurement of difference for classes. A next step, should we consider to evaluate differences in groups (like white-series or platelets) is to use multivariate tests like *Hotelling T-squared* or *Wilks* Λ.

3.5 Limitations and Possible Extensions

The limitations of this study are in two points. The first one is data: the variables to be analyzed ("wider": more columns) and the number of patients ("taller": more rows) could lead to a substantial improvement in the results achieved so far, allowing us to separate the clusters better.

More variables for each patient also mean that different representations could be obtained. In medical terms, more complex relationships could be extracted. Restricting ourselves only to blood exams, C-reactive protein, AST, ALT, GGT, and LDH could be excellent additions to the analysis. Other data sources could be leveraged: social and economic data could help to trace relationships between infection severity and social strata. Genetic markers could help to understand whether some populations are more susceptible to infections than others. Medical imaging data could help to associate blood parameters with physiological changes in organs and tissues, and so on.

The second point is the non-causality of analysis. None of this study's conclusions are causal for two reasons: The data is observational, and the number of patients and parameters

is not large. This reveals an excellent opportunity for researchers because the procedure applied here could be used to control the experiment data without any modifications. There are some studies in the literature that combine cluster analysis with causal inference but it is still very sparse (OLTEAN; GAGNIER, 2015). Statistically significant samples and more parameters could help to create groups of patients where a treatment (or protective measures) could be tailored for each group. Other diseases could also benefit from the same approach presented here.

Considering the nature of this research, other epidemics (e.g., Dengue fever, Zika Virus, Ebola) could be an excellent investigation opportunity, as the primary source of data used here is inexpensive and could be collected even in developing and emerging countries.

3.6 Final remarks

Using only data science methods, we were able to demonstrate that different prevalence subgroups exist, and that these groups have different medical interpretations and that they make sense. This article opens a window of opportunity for those with access to individual and more granular blood data for patients, paving the way for a more comprehensive analysis with more factors to be analyzed. Moreover, we aim to help to demonstrate that COVID-19 is not only "a simple flu" with only respiratory effects but a more complex disease with several potential implications and outcomes, particularly hematological as described by relevant statistical testing.

Special implications in platelets (which control coagulation), eosinophils and monocytes (related to infection control and adaptive immunity) further disclose that COVID-19 is a multi-systemic, multi-implication disease that must be analyzed from a multi-disciplinary perspective and the clusters found can be the first indication that several approaches must be taken by medical staff, policymakers and governments. On the future, we can use similar techniques with augmented data to address different problems related to COVID-19 such as vaccine distribution, field hospital construction, disease spread analysis and other issues. The technique presented here can be also easily adapted to other diseases as well.

4 Code Library & Results

All code for the articles is available in our github page (https://rb.gy/nvpbl2). The published text of the first article can be found here (https://rb.gy/zzkaax).

The second article is still awaiting acceptance in "Informatics in Medicine Unlocked".

Bibliography

ALASHWAL, H.; HALABY, M. E. et al. The application of unsupervised clustering methods to alzheimer's disease. *Front Comput Neurosci*, v. 13, 2019. Cited on page 40.

ALJAME, M.; AHMAD, I. et al. Ensemble learning model for diagnosing covid-19 from routine blood tests. *Informatics in Medicine Unlocked*, v. 21, 2020. Cited on page 40.

ALSHEREF, F. K.; GOMAA, W. H. Blood diseases detection using classical machine learning algorithms. *International Journal of Advanced Computer Science and Applications*, v. 10, n. 9, 2019. Cited 2 times on pages 20 and 22.

BAKSHY, E.; DWORKIN, L. et al. Ae: A domain-agnostic platform for adaptive experimentation. *NIPS'18: Proceedings of the 31th International Conference on Neural Information Processing Systems*, 2018. Cited on page 32.

BAO, J.; LI, C. et al. Comparative analysis of laboratory indexes of severe and non-severe patients infected with covid-19. *Clinica Chimica Acta*, v. 509, p. 180–194, 2020. Cited on page 39.

BARBOSA, V. A. de F.; GOMES, J. C. et al. Heg.ia: an intelligent system to support diagnosis of covid-19 based on blood tests. *Research on Biomedical Engineering*, 2021. Cited on page 40.

BATISTA, A. F. de M.; MIRAGLIA, J. L. et al. Covid-19 diagnosis prediction in emergency care patients: a machine learning approach. *medRxiv*, Cold Spring Harbor Laboratory Press, 2020. Disponível em: <https://www.medrxiv.org/content/early/2020/04/14/2020.04.04.20052092>. Cited on page 40.

BATTINELLI, E. M. Covid-19 concerns aggregate around platelets. *Blood*, v. 136, p. 1221–1223, 2020. Cited on page 45.

BEELER, C.; DBEIBO, L. et al. Assessing patient risk of central line-associated bacteremia via machine learning. *American Journal Infect Control*, v. 46, n. 9, p. 986–991, 2018. Cited 2 times on pages 21 and 22.

BERTSIMAS, D.; OHAIR, A. K.; PULLEYBLANK, W. R. *The Analytics Edge*. [S.l.]: Dynamic Ideas LLC, Belmont MA, 2015. Cited 2 times on pages 20 and 22.

BEZZAN, V.; ROCCO, C. D. *Predicting special care during the COVID-19 pandemic: A machine learning approach*. 2020. Cited on page 40.

BONANAD, C.; GARCíA-BLAS, S. et al. The effect of age on mortality in patients with covid-19: A meta-analysis with 611.583 subjects. *Journal of the American Medical Directors Association*, v. 21, p. 915–918, 2020. Cited on page 28.

BREIMAN, L. Random forests. *Machine Learning*, v. 45, p. 5–32, 2001. Cited on page 30.

BRIER, G. W. Grabit: Gradient tree-boosted tobit models for default prediction. *Monthly Weather Review*, v. 78, n. 1, 1950. Cited on page 35.

BRINATI, D.; CAMPAGNER, A. et al. Detection of covid-19 infection from routine blood exams with machine learning: A feasibility study. *Journal of Medical Systems*, v. 44, n. 8, 2020. Cited 3 times on pages 13, 21, and 22.

____. Detection of covid-19 infection from routine blood exams with machine learning: A feasibility study. *Journal of Medical Systems*, n. 135, 2020. Cited on page 40.

CAI, J.; LUO, J. et al. Feature selection in machine learning: A new perspective. *Neurocomputing*, v. 300, p. 70–79, 2018. Cited on page 27.

CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. *KDD*, 2016. Cited on page 30.

COLUBRI, A.; SILVER, T. et al. Transforming clinical data into actionable prognosis models: Machine-learning framework and field-deployable app to predict outcome of ebola patients. *PLOS Neglected Tropical Diseases*, v. 10, n. 3, 2016. Cited 2 times on pages 21 and 22.

DUTTA, S.; BANDYOPADHYAY, S. K. Machine learning approach for confirmation of covid-19 cases: positive, negative, death and release. *Iberoamerican Journal of Medicine*, v. 03, p. 172–177, 2020. Cited 2 times on pages 21 and 22.

EFRON, B. Estimating the error rate of a prediction rule: Improvement on cross-validation. *Journal of the American Statistical Association*, v. 78, p. 316–331, 1983. Cited on page 16.

ELAZIZ, M. A.; HOSNY, K. M. et al. New machine learning method for image-based diagnosis of covid-19. *PLOS ONE*, 2020. Cited 2 times on pages 21 and 22.

ESTHER, M.; KRIEGEL, H.-P. et al. A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, p. 226–231, 1996. Cited on page 17.

FAPESP. Covid 19 data sharing brasil. 2020. Cited on page 19.

FATIMA, M.; PASHA, M. Survey of machine learning algorithms for disease diagnostic. *Journal of Intelligent Learning Systems and Applications*, v. 9, n. 1, 2017. Cited 2 times on pages 21 and 22.

FEFFERMAN, C.; MITTER, S. et al. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, v. 29, n. 4, p. 983–1049, 2016. Cited on page 41.

FERRARI, D.; MOTTA, A. et al. Routine blood tests as a potential diagnostic tool for covid-19. *Clinical Chemistry and Laboratory Medicine*, 2020. Cited 4 times on pages 13, 22, 33, and 39.

FEURER, M.; HUTTER, F. *Hyperparameter Optimization. In: Automated Machine Learning, The Springer Series on Challenges in Machine Learning*. [S.l.]: Springer, Cham., 2019. 3-33 p. Cited on page 26.

FREUND, Y.; SCHAPIRE, R. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, v. 55, p. 119–139, 1997. Cited on page 15.

G.D., W.; J.L., M. The impact of covid-19 disease on platelets and coagulation. *Pathobiology*, v. 88, p. 15–27, 2021. Cited on page 45.

GEURTS, P.; ERNST, D.; WEHENKEL, L. Extremely randomized trees. *Machine Learning*, v. 63, p. 3–42, 2006. Cited on page 30.

GUNčAR, G.; KUKAR, M. et al. An application of machine learning to haematological diagnosis. *Scientific Reports*, v. 8, n. 411, 2018. Cited 2 times on pages 20 and 22.

GUYON, I.; ELISSEEFF, A. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 2003. Cited on page 27.

GüçLü, E.; KOCAYIğIT, H. et al. Effect of covid-19 on platelet count and its indices. *Revista da Associação Médica Brasileira*, v. 66, 2020. Cited on page 45.

HALL, D. B. Zero-inflated poisson and binomial regression with random effects: a case study. *Biometrics*, v. 56, n. 4, p. 1030–1039, 2000. Cited on page 36.

HAN, J.; GONDRO, C.; REID, K.; STEIBEL, J. P. Heuristic hyperparameter optimization of deep learning models for genomic prediction. *G3 Genes|Genomes|Genetics*, 02 2021. ISSN 2160-1836. Jkab032. Disponível em: <https://doi.org/10.1093/g3journal/jkab032>. Cited on page 27.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The Elements of Statistical Learning*. [S.l.]: Springer-Verlag, New York, 2009. Cited 2 times on pages 16 and 33.

HEINZE, G.; WALLISCH, C. et al. Variable selection – a review and recommendations for the practicing statistician. *Biometrical Journal*, v. 60, p. 431–449, 2018. Cited on page 27.

HIAE. *Diagnosis of COVID-19 and its clinical spectrum*. 2020. <https://www.kaggle.com/einsteindata4u/covid19>. Accessed: 2021-02-17. Cited on page 39.

JAIN, V.; CHATTERJEE, J. M. *Machine Learning with Health Care Perspective*. [S.l.]: Springer International Publishing, 2020. Cited 2 times on pages 21 and 22.

JHCRC. Johns hopkins coronavirus resource center. 2021. Accessed: 2021-02-17. Cited on page 39.

J.ROUSSEEUW, P. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, v. 20, p. 53–65, 1987. Cited on page 41.

KE, G.; MENG, Q. et al. Lightgbm: A highly efficient gradient boosting decision tree. *Conference on Neural Information Processing Systems*, 2017. Cited on page 30.

KHANDAY, A. M. U. D.; RABANI, S. T. et al. Machine learning based approaches for detecting covid-19 using clinical text data. *International Journal of Information Technology*, v. 12, p. 731–739, 2020. Cited 2 times on pages 21 and 22.

KUKAR, M.; GUNčAR, G. et al. *COVID-19 diagnosis by routine blood tests using machine learning*. 2020. Cited on page 40.

KUMAR, B. A novel latent factor model for recommender system. *Journal of Information Systems and Technology Management*, v. 13, n. 3, 2016. Cited on page 25.

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*, v. 139, 2020. Cited on page 22.

LALWANI, P.; MISHRA, M. K. et al. Customer churn prediction system: a machine learning approach. *Computing*, 2021. Cited on page 27.

LECUN, Y.; CORTES, C. MNIST handwritten digit database. 2010. Disponível em: <http://yann.lecun.com/exdb/mnist/>. Cited 3 times on pages 9, 41, and 42.

LIAO, D.; ZHOU, F. et al. Haematological characteristics and risk factors in the classification and prognosis evaluation of covid-19: a retrospective cohort study. *The Lancet*, v. 7, 2020. Cited on page 39.

LIPPI, G.; PLEBANI, M. Laboratory abnormalities in patients with covid-2019 infection. *Clinical Chemistry and Laboratory Medicine*, De Gruyter, Berlin, Boston, n. 0, 2020. Cited 3 times on pages 10, 22, and 23.

LIU, M.; OXNARD, G. et al. Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free dna. *Annals of Oncology*, v. 31, n. 6, 2020. Cited 2 times on pages 20 and 22.

LORENZO, P. R.; NALEPA, J. et al. Particle swarm optimization for hyper-parameter selection in deep neural networks. *GECCO '17: Proceedings of the Genetic and Evolutionary Computation Conference*, p. 481–488, 2017. Cited on page 27.

LUNDBERG, S. M.; LEE, S.-I. A unified approach to interpreting model predictions. In: GUYON, I.; LUXBURG, U. V.; BENGIO, S.; WALLACH, H.; FERGUS, R.; VISHWANATHAN, S.; GARNETT, R. (Ed.). *Advances in Neural Information Processing Systems 30*. [S.l.]: Curran Associates, Inc., 2017. p. 4765–4774. Cited on page 33.

MAHMOOD, S. S.; LEVY, D. et al. The framingham heart study and the epidemiology of cardiovascular diseases: A historical perspective. *Lancet*, Lancet, Boston, EUA, n. Mar 15 383(9921): 999–1008, 2014. Cited on page 20.

MAZUMDER, R.; HASTIE, T.; TIBSHIRANI, R. Spectral regularization algorithms for learning large incomplete matrices. *Journal of Machine Learning Research*, v. 11, p. 2287–2322, 2010. Cited on page 25.

MCINNES, L.; HEALY, J.; MELVILLE, J. *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. 2020. Cited 3 times on pages 14, 17, and 39.

MCLACHLAN, G. Cluster analysis and related techniques in medical research. *Statistical Methods in Medical Research*, v. 1, 1992. Cited on page 40.

MEI, H.; LUO, L. et al. Thrombocytopenia and thrombosis in hospitalized patients with covid-19. *Journal of Hematology & Oncology*, v. 13, 2020. Cited on page 45.

MLHC. Machine learning for healthcare conference. 2020. Cited on page 21.

MOCKUS, J. On bayesian methods for seeking the extrema. *Optimization Techniques*, p. 400–404, 1974. Cited on page 17.

____. Application of bayesian approach to numerical methods of global and stochastic optimization. *Journal of Global Optimization*, v. 4, p. 347–365, 1994. Cited on page 26.

NGUYEN, M. H. Smote: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, p. 321–357, 2002. Cited on page 26.

____. Impacts of unbalanced test data on the evaluation of classification methods. *International Journal of Advanced Computer Science and Applications*, v. 10, n. 3, 2019. Cited on page 26.

NICULESCU-MIZIL, A.; CARUANA, R. Predicting good probabilities with supervised learning. *Proceedings of the 22 nd International Conference on Machine Learning*, 2005. Cited on page 35.

OLTEAN, H.; GAGNIER, J. J. Use of clustering analysis in randomized controlled trials in orthopaedic surgery. *BMC Medical Research Methodology*, v. 15, 2015. Cited 2 times on pages 40 and 48.

PAUL, R.; SAYED, A. Clustering medical data to predict the likelihood of diseases. *2010 Fifth International Conference on Digital Information Management (ICDIM)*, 2010. Cited on page 40.

PEDREGOSA, F.; VAROQUAUX, G. et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011. Cited on page 26.

PEIFFER-SMADJA, N.; MAATOUG, R. et al. Machine learning for covid-19 needs global collaboration and data-sharing. *Nature Machine Intelligence*, v. 2, p. 293–294, 2020. Cited on page 19.

QOLOMANY, B.; MAABREH, M. et al. Parameters optimization of deep learning models using particle swarm optimization. *13th International Wireless Communications and Mobile Computing Conference (IWCMC)*, 2017. Cited on page 27.

SANCHE, S.; LIN, Y. T. et al. High contagiousness and rapid spread of severe acute respiratory syndrome coronavirus 2. *Emerging Infectious Diseases*, CDC Centers for Disease Control and Prevention, Los Alamos Laboratory EUA, v. 26, n. 7, 2020. Cited on page 19.

SCHUBERT, E.; SANDER, J. et al. Dbscan revisited, revisited: Why and how you should (still) use dbscan. *ACM Transactions on Database Systems*, v. 19, 2017. Cited on page 41.

SIGRIST, F.; HIRNSCHALL, C. Grabit: Gradient tree-boosted tobit models for default prediction. *Journal of Banking & Finance*, v. 102, p. 177–192, 2019. Cited on page 36.

SKERMAN, H. M.; YATES, P. M. et al. Multivariate methods to identify cancer-related symptom clusters. *Research in Nursing & Health*, v. 32, p. 345–360, 2009. Cited on page 40.

SMOLA, A. J.; BARTLETT, P. *Advances in Large-Margin Classifiers*. [S.l.]: MIT Press, Cambridge MA, 2000. Cited on page 35.

SNOEK, J.; LAROCHELLE, H.; ADAMS, R. P. Practical bayesian optimization of machine learning algorithms. *NIPS'12: Proceedings of the 25th International Conference on Neural Information Processing Systems*, v. 2, p. 2951–2959, 2012. Cited on page 26.

SOUZA, W. M. de; BUSS, L. F. et al. Epidemiological and clinical characteristics of the covid-19 epidemic in brazil. *Nature Human Behaviour*, v. 4, p. 856–865, 2020. Cited on page 44.

TIBSHIRANI, R. Regression shrinkage and selection via the lasso. *Journal of Royal Statistical Society*, v. 1, p. 267–188, 1996. Cited on page 27.

TROYANSKAYA, O.; CANTOR, M. et al. Missing value estimation methods for dna microarrays. *Bioinformatics*, v. 17, 2001. Cited on page 25.

WAGER, S.; ATHEY, S. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, v. 113, p. 1228–1242, 2018. Cited on page 43.

WU, C.-C.; YEH, W.-C. et al. Prediction of fatty liver disease using machine learning algorithms. *Computer Methods and Programs in Biomedicine*, v. 170, p. 23–29, 2019. Cited 2 times on pages 21 and 22.

WU, W. B.; WOODROOFE, M. et al. Isotonic regression: Another look at the changepoint problem. *Biometrika*, v. 88, n. 3, p. 793–804, 2001. Cited on page 35.

YUAN, X.; HUANG, W.; YE, B. et al. Changes of hematological and immunological parameters in covid-19 patients. *International Journal of Hematology*, 2020. Cited 2 times on pages 23 and 33.