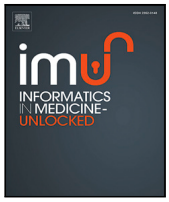
Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Informatics in Medicine Unlocked 32 (2022) 101023

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: www.elsevier.com/locate/imu



Risk assessment in COVID-19 patients: A multiclass classification approach

Roberto Bárcenas *, Ruth Fuentes-García

Departamento de Matemáticas, Facultad de Ciencias, Universidad Nacional Autónoma de Mexico, Mexico

ARTICLE INFO

Keywords:
COVID-19
Machine Learning
Risk assessment
Multiclass classification
Feature importance

ABSTRACT

Understanding SARS-CoV-2 infection that causes COVID-19 disease among the population was fundamental to determine the risk factors associated with severe cases or even death. Amidst the study of the pandemic, Artificial Intelligence (AI) and Machine Learning (ML) have been successfully applied in many areas such as biomedicine. Using a dataset from the Mexican Ministry of Health, we performed a multiclass classification scheme for the detection of risks in COVID-19 patients and implemented three Machine Learning algorithms achieving the following accuracy measures: Random Forest (89.86%), GBM (89.37%) XGBoost (89.97%). The key findings are the identification of relevant components associated with different severities of COVID-19 disease. Among these factors, we found sex, age, days elapsed from the beginning of symptoms, symptoms such as dyspnea and polypnea; and other comorbidities such as diabetes and hypertension. This setting allows us to establish predicting algorithms to model the risk that an individual or a specific group of people face after contracting COVID-19 and the factors associated with developing complications or receiving appropriate treatment.

1. Introduction

The SARS-CoV-2 virus is a type of coronavirus (from the same family as the one causing the SARS-severe acute respiratory syndrome), first detected in humans in December 2019 in the Chinese town of Wuhan. It causes coronavirus disease since 2019, named as COVID-19. It is considered a global health emergency by the World Health Organization (WHO) and on 11 March 2020, the same organization reported the COVID-19 outbreak as a pandemic [1]. The environment during the onset and spread of coronavirus 2 (SARS-CoV-2) is surely accompanied by uncertainty. Initially, on its main epidemiological, clinical, and virological characteristics and, in particular, on its ability to spread in humans and its virulence.

Symptoms of COVID-19 include coughing, sneezing, fever, and shortness of breath and contagion occurs by inhaling tiny droplets that are emitted from the affected person to the healthy person through expectorations such as coughing and sneezing, in an interpersonal contact one or two meters away [2]. According to the centers for disease control and prevention (CDC) in the U.S. [3], there are some keys to identifying the disease: shortness of breath, if not an early symptom, certainly a serious one; high fever and dry cough, and chills and body aches. Also, fatigue and pain in the joints and muscles, headache, sore throat, and congestion. As well as conjunctivitis and loss of smell and taste, sudden confusion (inability to wake up) and in some cases digestive problems (diarrhea). Currently, it is unknown whether recent memory loss, difficulty paying attention, and slowness to process information, among others, are due to lack of oxygenation during the critical stage of infection or whether the direct action of the virus on neurons is due [4].

Recent findings indicate that COVID-19 is not only a respiratory disease, but a thrombo-inflammatory microvascular syndrome, which also affects the lungs more frequently and severely, causing neurological, kidney, cardiac and liver damage [5]. Based on the evidence found so far [6], there are traces that older people and those with chronic diseases, such as high blood pressure, heart disease, obesity, or diabetes, develop severe cases of the disease more often than others. In a study from the China Centers for Disease Control and Prevention within a large number of COVID-19 positive patients, it is argued that age, cardiovascular disease, diabetes, chronic respiratory diseases, comorbidities as obesity, hypertension, and cancer were associated with a high risk of death. Also in many cases, variables such as sex and smoking were related to increased risks of severe cases [7].

Regarding the COVID-19 epidemic in Mexico, and in particular the interpretation of official statistics, health authorities have pointed out that patients with this new disease who have other comorbidities could have a significantly higher risk of dying [8]. An important finding, observed in a study conducted by the National Institute of Respiratory Diseases in Mexico, shows that the lungs were blocked by clots causing death. People die for other reasons because COVID-19 originates multisystemic damage, leading to organ failure and eventually death [9].

* Correspondence to: Departamento de Matemáticas, Facultad de Ciencias, UNAM Circuito Exterior s/n, 04510, Mexico City, Mexico
E-mail address: rbacenas@ciencias.unam.mx (R. Bárcenas).

https://doi.org/10.1016/j.imu.2022.101023
Received 26 April 2022; Received in revised form 9 July 2022; Accepted 14 July 2022
Available online 19 July 2022

From a statistical perspective, ML leads to pattern detection through the use of several algorithms. Here, predictive models deal with learning from data to explicitly find knowledge from the available information. Applications of Artificial Intelligence (AI) and Machine Learning (ML) have made remarkable progress in biomedicine: image recognition and genetics, to learn the overall properties of genes such as DNA configurations and sequences. Until recently there were no tools for their analysis, but many studies have used ML algorithms for the first time to explain genetic, biological, clinical, and social processes. This opens up the possibility of this technology moving to other areas; however, these advances have been best done in single-task applications where inaccurate results and occasional errors can be handled. Medical practice is singular and represents a large-scale challenge. Dealing with sensitive information using relevant clinical data leads us to re-formulate how ML could help solve these problems [10].

There are approaches that electronically monitor cough recordings, converting them into spectrograms. Different AI algorithms are able to detect changes in the cough pattern, facilitating early diagnosis of the infection. Cough is not only an early and characteristic symptom of COVID-19, but also of other respiratory diseases such as tuberculosis or influenza. Therefore, there are proposals to train learning algorithms to detect changes in cough patterns [11]. In fact, voice spectral analysis is an objective and non-invasive evaluation method used for research purposes. It uses acoustic recordings obtained directly from the patient in order to diagnose the disease from cough signals [12]. The classification itself could be able to distinguish between COVID-19 and different respiratory diseases.

This paper aims to introduce learning methods for detecting and classifying risk scenarios and factors in COVID-19 patients. The primary goal is to explore and recognize features that influence the distinction between developing a serious illness and mild cases, and to identify patients susceptible to developing a critical condition or even at the risk of death. Then, given a set of labels that describe three risk scenarios depending on the evolution of the disease, a multiclass classification setting is proposed. In this way, we can alert the population with the means and tools to understand the risk factors, which can be used by the authorities to schedule lock-down agendas and expedite vaccination in each country.

1.1. Related work

The new contribution of ML and Data Mining techniques in medical fields can provide an alternative way to create better applications for the future. The systematic review [13] is a comprehensive state-of-the-art study for COVID-19 learning algorithms based on data mining and ML. From this substantial amount of information (1305 articles), they considered the reliability and acceptability of the datasets and features extracted from the technologies implemented in the literature. An interesting conclusion reveals that it is important to merge all gained knowledge and expand it massively to identify solutions to the major problems of this pandemic and introduce novel approaches, which mean valuable time savings.

In this context, original contributions arise to provide information that helps to deal with the pandemic: the most popular is directed towards the early detection, treatment, and control of COVID-19. Predictive models could improve COVID-19 diagnostic prediction in the early stages of infection. An analysis of the patient's characteristics, case history, comorbidities, symptoms, diagnosis, and results appears in Ahamad et al. (2020) [14]. They train supervised machine learning algorithms to accurately examine the characteristics of COVID-19 disease diagnoses. Some include variables such as age, sex, fever, travel history, clinical details, including the severity of the cough and the incidence of lung infection. In [15] a meta-analysis is performed to study data and build computational models to predict whether a patient has COVID-19, based on their clinical information. That is, clinical information from patients is re-analyzed to state the diagnosis of COVID-19, rather than relying only on their symptoms regardless of the association between different clinical variables. Li et al. (2020) [15] report that their classification algorithms could acceptably discriminate positive COVID-19 patients against influenza patients.

As for methods closer to our approach, there are proposals based on Machine Learning (ML) for risk prediction in patients with COVID-19. We believe that it is important to have tools to identify people with higher mortality risk. This infection often leads to nosocomial spread, affecting health workers and general health care services. Hospitalization loads can saturate health systems due to a prime need for oxygen, prolonged ventilation, and even extracorporeal membrane oxygenation, particularly in patients with acute respiratory distress syndrome. In Jiang et al. (2020) [16] historical records are used to predict whether a patient will develop a severe case of COVID-19. Based on data from two hospitals in Wenzhou, Zhejiang, China, they considered an algorithmic predictive model that combines previous records and clinical patient information, specifically a slightly elevated liver enzyme, myalgias (body aches), elevated hemoglobin (red blood cells), and proneness to Acute Respiratory Difficulty Syndrome. This is one of the early attempts of such initiatives, whose accuracy ranges from 70% to 80% predicting serious cases.

In a retrospective study, Assaf et al. (2020) [17] reveal the need for an efficient triage to tackle the COVID-19 pandemic. To assess the capabilities of ML models, they train three different algorithms: Neural Network, Random Forest, and CART to predict patient deterioration. Among the 6995 patients evaluated, they achieve outstanding performance by predicting their critical COVID-19 risk based on baseline clinical parameters and other covariates. Burdick et al. (2020) [18] conduct as real a clinical trial as possible to evaluate the performance of an algorithm designed in some U.S. health systems. To make predictions, they consider admissions to the Intensive Care Unit (ICU), invasive ventilation cases, and deaths. The advantage is the ability to test the behavior of risk prediction models under uncertainty, and the convenience of replicating analysis in prospective clinical settings. This is important in an accelerated outbreak of these features and when critical care resources and hospital beds are limited, forcing doctors to make tough decisions. This could guide future research and underlines the relevance of implementing proposals such as the one raised in our work.

The rest of this article is organized as follows. In Section 2, we analyze the materials to characterize risk scenarios, based on available feature information. In Section 3 as a second stage, three statistical learning algorithms are implemented in a multiclass classification context to train and determine their performance in the task of assigning a risk outcome to a positive SARS-CoV-2 patient. In Section 4, we present the results that validate the relevance of our approach, including an analysis of feature importance obtained from the classification algorithms. For Section 5, we discuss the proceeds of the study, and finally, we examine some conclusions and the areas for progress in Section 6.

2. Materials

The COVID-19 emergency has caused a deep crisis in many areas of our lives. Given the nature of the emergency, governments and society have had to balance priorities and needs in all sectors (social, economic, educational, health) at the same time. The community's response to contingency has manifested itself in different ways, addressing with different approaches the problems generated by the impact of the virus. Probably the major challenges we face are: (i) in health terms, to protect and alert people of the risks of contracting the disease and to detect it with no extensive testing; (ii) in managing terms to end the emergency, to regenerate the economy and all stopped processes. Here, an upcoming genuine option is the vaccination protocol. Initially in Mexico states sought to contain the immediate impact of COVID-19 and have planned their government-coordinated vaccination process. However, in this new reality, there is still much to study and learn. The crucial discussion is to take advantage of the expertise generated during this crossing and from this knowledge available, quantify the risk that anyone may have when exposing COVID-19.

2.1. Dataset description

In this paper, we considered a data subset on COVID-19 infections, provided by the Epidemiological Surveillance System for Respiratory Diseases under the Directorate-General for Epidemiology of the Ministry of Health in Mexico. These are patient observations recorded from January 17, 2020, to June 28, 2020, including confirmed cases of COVID-19, but are limited to the target analysis and dissemination of official information available online in a system created in collaboration between the Institute of Biotechnology and IIMAS, both from UNAM, Mexico.[1] The data collected through these initiatives could be substantial to improve recommendations on case definitions and COVID-19 surveillance. Additionally they could be useful to understand the underlying epidemiological characteristics of the disease, as well as for understanding and determining its spread, clinical spectrum, severity, and community effects. This provides prompt information for developing guidance on implementing measures to contain it, such as case isolation and contact monitoring. The programs for this paper were generated using R software for statistical computation and graphics [19].

Essentially, the dataset includes information on (a) sociodemographic variables, (b) symptoms, (c) comorbidities, and (d) health system records. A relevant point is that at the time of generation of the report, we have the profile of the patient for positive cases of COVID-19 and its respective outcomes. In total, there is a record of 583,678 patients, but we only considered a total of 220,657 confirmed positive cases of SARS-CoV-2. The exclusion criteria in this study were as follows: cases that do not meet the definition or that were not well certified in the database. Or, cases that do not include a complete set of covariates. As a reference, we considered the variable labeled as *Progression* to study the evolution of the disease in each patient. It comprises six cases: Non-severe, Recovered, Severe, Decease, Treatment, and Monitoring. In the first part of the analysis we considered sex and age which have been useful to distinguish more severe cases. Additionally the days since the initial symptoms is observed to understand the dynamics of the disease. Here, we have basic information and intuition about the risk factors that lead patients to experience only moderate clinical manifestations or a serious illness, even reaching a tragic outcome. At best, the patient is sent home, under treatment and monitoring [20].

2.2. Statistical data analysis

According to global experience, COVID-19 has a similar incidence in men and women, but the aspect that seems to be different is vulnerability and mortality, Emerging evidence suggests higher mortality in men. One explanation could be related to sex-based immune response, where not only men's immune systems are weaker, but they are also affected by social factors such as lower hygiene, increased exposure to the disease due to work activities and increased tobacco use. Other research highlights the differences between genders in more characteristics, from hormones and cell receptors to genetics linked to the X chromosome [21]. In our case, the distribution by sex is 45.25% women versus 54.75% men. While a difference in the disease's prevalence is not so clear, it is more evident in the most severe cases. In the observed period, mortality was higher in men (65.97%) than in women (34.03%). This represents approximately 1.9 deaths in men for each death registered in women. This trend was similar to that of severe cases, which are 62.2% for men and 37.8% for women. Therefore, if we grouped together deaths and serious cases, the ratio would be 64.43% men and 35.57% women, respectively. Also frequencies per evolution preserve the same order in each sex group (see Fig. 1a).

Many studies emphasize that age is another risk factor for COVID-19, but that it is not the only one to consider [4,22]. According to the centers for disease control and prevention (CDC) in the US, 8 out of 10 COVID-19 deaths occur at the adult age of 65 and older. Compared to younger adult patients, older adults are more likely to need hospitalization if they contract COVID-19. This may be because the immune system deteriorates as people age, regardless of gender. However, what is a mild decline for women is an abrupt drop for men. Various sources in the biological area show that the response of men's immune cells (T cells) between ages 30 and 50 is equivalent to that of a 90-year-old woman. These not only detect infected cells and eliminate them along with help manage the response of antibodies. In male patients, scientists have found that T-cell response was weaker. Its damage can affect the ability of the heart muscles to function, leading to heart failure [23]. According to the evidence we found, they are not the only disproportionately impaired immune characteristics in men, who also have more associated pathologies, from hypertension to diabetes. Therefore, aging is strongly associated with an increased risk of death in both genders, but, in particular, for ages over age 50, male patients have a significantly higher risk of mortality, making older people a vulnerable group (Fig. 1b).

The interval between onset of symptoms and diagnosis is effective for medical care and for limiting disease transmission through case epidemiological study and contact traceability. COVID-19 disease shows an extremely variable course, from asymptomatic to abruptly lethal at brief time intervals. Sometimes, it affects young and seemingly healthy people, for whom the severity of the disease is not induced by age or any comorbidity. Besides genetic predisposition, we should consider other probable reasons for a severe course: the amount of viral exposure, the route by which the virus enters the body, the virulence of the pathogen and possible (partial) immunity from previous viral diseases. Inhaling numerous viruses deeply immediately causes many inoculators viruses in the lung system. This can have a much faster effect than receiving a small inoculating virus, which would lead to slow and even asymptomatic disease [24].

To perform this part of the analysis, we considered the interval given by the number of days between the symptom start date and the system registration date. Knowledge of the diagnostic interval can contribute to improving the processes that must be carried out to issue the confirmatory alerts from an analytical phase, and even in the result's issuance. If these intervals are more precise, delays will be avoided, for example, in the recognition of cases and contacts, and in establishing control measures. In Fig. 1c, we found that when comparing severe cases and deaths, the median number of days from the onset of symptoms to the registry was the same on almost all progression labels. However, this period should be considered depending on the patient's age and immune system status. In our observations, the difference lies in the behavior of the days from the onset of symptoms of the patient's progression between the distinct groups. Also, in Fig. 1c we observed that the distribution is more uniform in non-severe and recovered cases, showing some changes for deaths and severe cases. In fact, in the latter group, the cut is more abrupt. Here, we might think that the transition to the severity of disease occurs rapidly, then it is maintained slowly in intensive care. For the death cases, something similar happens, it either occurs at the onset of symptoms, or later. In the latter probably after medical care, either by emergency treatments or intubation, days lengthen to the lethal outcome. Treatment and monitoring dynamics have more variability, given the nature of clinical follow-up, but behave with some sync.

The previous analysis allowed us to notice a natural grouping between pairs of progression labels. The frequencies and distribution of variables behave homogeneously between non-severe and recovered cases and between cases of treatment and monitoring. In severe cases and deaths, although there are some differences, their dynamic is similar. Thinking of a cautious strategy for risk characterization, in the sense of preventing severe cases and deaths, we will consider a

[1] See http://covid-19.iimas.unam.mx/.

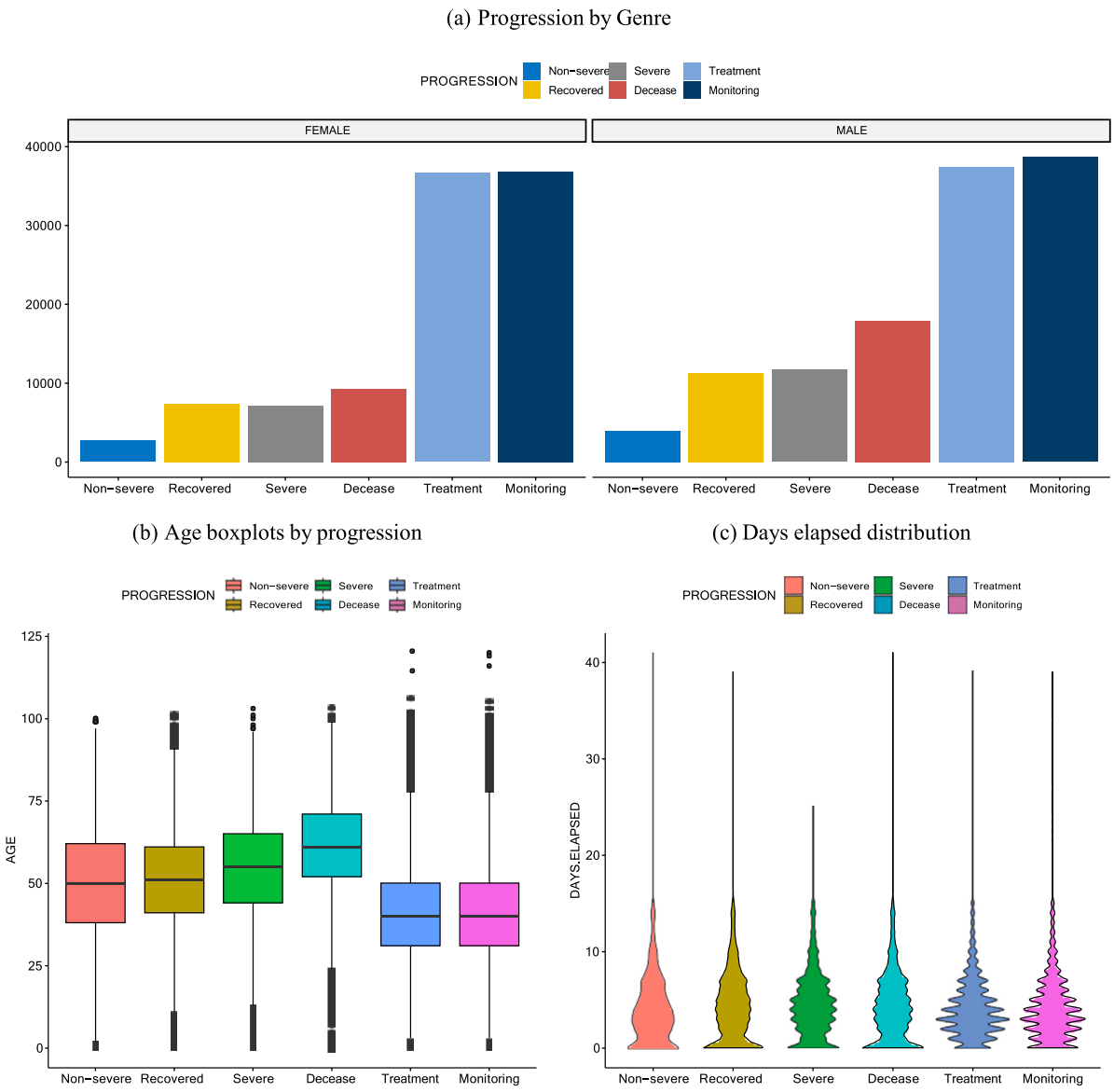


Fig. 1. Progression grouped by Sex, Age and Days elapsed.

grouping of classes in three risk scenarios. We suggest that case labels monitoring and treatment are most often given in a low-risk group. Then, the non-severe and recovered correspond to a moderate-risk scenario because while they developed the disease, it did not go to a degree of needing more attention. Finally, cases of interest, given by the severe manifestation and outcome of death, will be treated as the high-risk scenario. This rearrangement will allow us to clearly describe the relationships between symptoms and comorbidities since the patterns correspond to a specific dynamic, efficiently identifying risk. It simplifies our problem to a multi-classification task of three classes, which decreases variance and makes it more efficient to manage the performance measures of the algorithms.

One facet, widely studied, which could increase the risk of severe COVID-19 is having certain underlying chronic conditions. Knowing the factors that involve greater exposure to a serious scenario allows us to decide what provisions we should take in daily life [25]. In Mexico, several chronic diseases prevail that have influenced the new coronavirus as comorbidity factors. Official estimates from the 2018 National Health and Nutrition Survey show that obesity and overweight in Mexico affect 75.2% of adults over the age of 20; diabetes is present in 14.4% of the elderly population and rises to 30% in over-50s [26]. According to our analysis, we have identified three main comorbidities in people infected with the SARS-CoV-2 virus. In 16.4% of confirmed cases hypertension occurs, 20.08% are obese, and 19.6% report diabetes. Obesity is the first risk factor for causing other serious disorders such as diabetes and hypertension, which have also been associated with severe cases of COVID-19. As for the incidence of these, the specific percentages by risk group showed that in cases of high-risk, 61.97%, 69.26%, and 45.3% suffer from diabetes, hypertension, and obesity, respectively. Among cases of moderate and low-risks, the most prevalent pre-existing diseases were also hypertension, diabetes, and obesity. In fact, some patients had several at once, making them more likely to suffer serious conditions from COVID-19 (see Fig. 2a).

A mosaic plot [27] is a graphical display of cross-table data in which a rectangle of size proportional to the count represents each cell count. Mosaics are suitable to identify high or small counts, pointing to dependencies between variables. It sets the positions and sides of the rectangles to facilitate comparisons between counts in the cells. Here, a log-linear model by Iterative Proportional Fitting is carried for the multidimensional contingency tables. Extended mosaic displays visualize standardized residuals of the log-linear model for the table by color and outline of the mosaic tiles. Recall, the color shows the Pearson residuals sign, black for the positive, and red for the negative residuals.

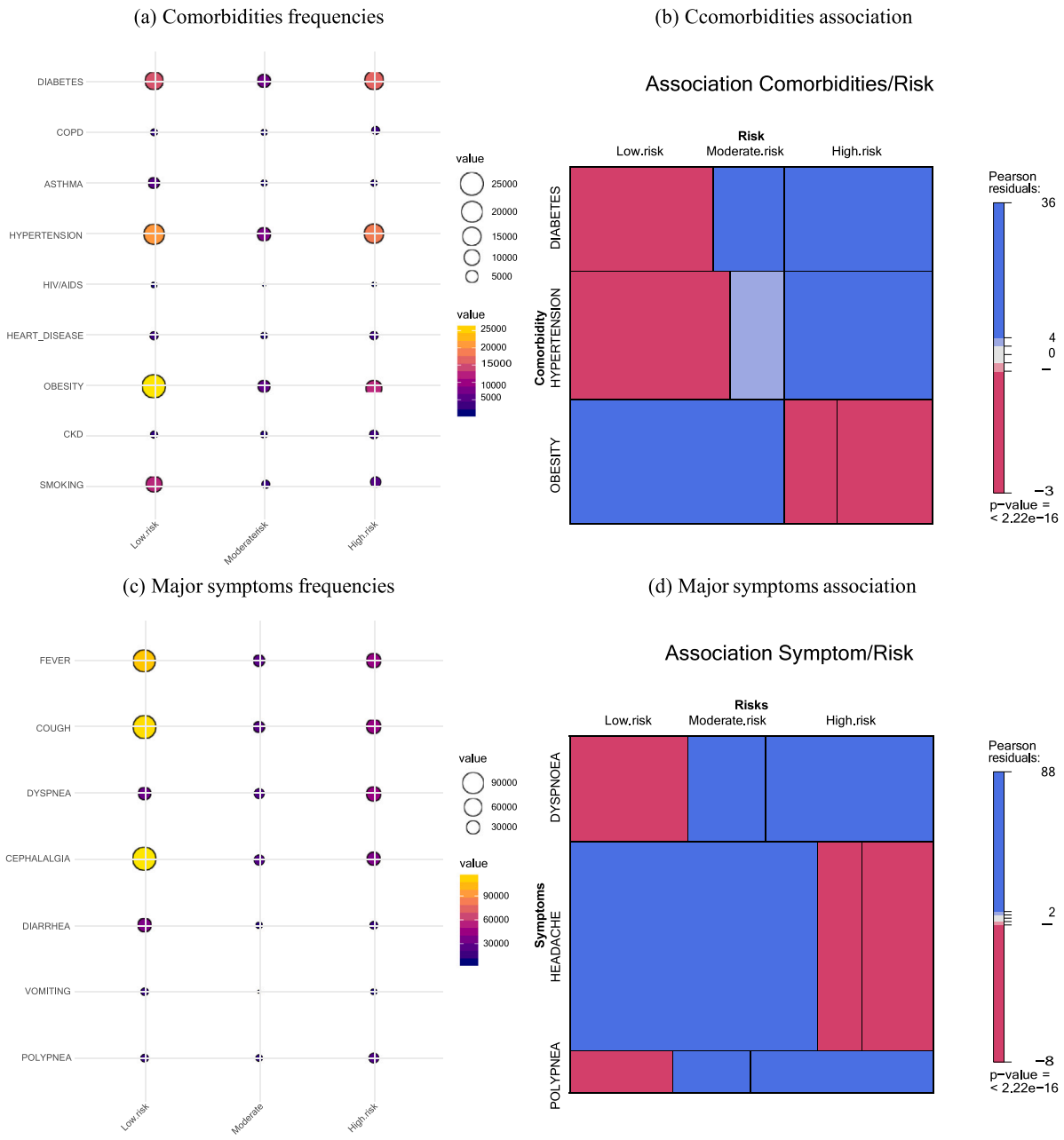


Fig. 2. Risk level by comorbidities and major symptoms. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

It sets the saturation of a residual according to its size: high saturation for large, and low saturation for small residuals. In Fig. 2b, the plot illustrates big-size rectangles with large positive residuals (greater than 4) for diabetes/high-risk, diabetes/moderate-risk, and obesity/low-risk, where all three are depicted in deep black. There is a large negative residual (less than 4) for hypertension/low-risk, colored in deep red. There are also medium-sized positive residuals between 2 and 4 for hypertension/moderate-risk, which is less saturated.

In this data, the most common major symptoms for COVID-19 infection are dyspnea, cephalalgia, dry cough, polypnea, vomiting, diarrhea, and fever. Some patients also have a secondary symptoms: odynophagia, myalgia (discomfort in the body through joint and muscle pain), arthralgia, conjunctivitis, anosmia, dysgeusia, and chills. These symptoms may appear 2 to 14 days after exposure to the virus. Because these symptoms also occur in common colds, it is mandatory to know others that can occur, such as nasal congestion (rhinorrhea), diarrhea and rashes, or color changes in fingers or toes. Based on our information, the major symptoms among infected patients were: fever, headache, cough, and dyspnea (78.51%, 76.14%, 74.28%, and 38.54%), which have been consistent in several studies with the highest reported percentages [6,10]. We can point to dyspnea as the comorbidity that can distinguish more patients in a risk scenario because its prevalence percentage is 86.29% of cases. For moderate and low-risk groups it represents 69.87% and 18.6%, respectively (see Fig. 2c). When using the mosaic plot (Fig. 2d), the prior relationships are distinctly observed. There is a positive association between the dyspnea and the high-risk group. Instead, in the low-risk scenario, the relationship is reversed for the same covariate. The headache association is noticeable for the low-risk group. Note that 74.79% of non-serious or recovered patients develop this symptom.

An unusual and worth exploring issue is attention in medical units. The information describing the clinical arrangement by sector and unit

of patients with COVID-19 in Mexico is still limited, however, there are variables available to inspect their relationship with outcomes. Recall, a confirmed case has a positive test to SARS-CoV-2 and a diagnosis endorsed by the National Network of Public Health Laboratories recognized by the Institute of Epidemiological Diagnostics and Reference; or, by the definition criteria, a person of any age who in the last seven days has presented at least two of the following signs and symptoms: cough, fever, or headache. Also, someone accompanied by at least one of the following signs or symptoms: dyspnea (gravity sign), arthralgias, myalgias, odynophagia, rhinorrhea, conjunctivitis, chest pain. Anosmia and dysgeusia should override symptoms; not necessarily rhinorrhea, which is an influenza indicator. According to the standardized guideline by the National Committee for Epidemiological Surveillance, given a confirmed case and based on the clinical diagnosis of admission, a patient is considered an outpatient or hospitalized. It is said to be outpatient if the patient returned home or is referred to as hospitalized if they are admitted to the health unit [28]. In our study, we have a proper variable that tells us the type of patient according to whether he received outpatient medical care, that is, we know if a patient received medical and diagnostic care but did not have to spend the night in the health unit, or if the patient required hospitalization. This information together with the progression variable, described as the patient's evolution to register the database, and allowed us to outline relationships focused on the primary factors of clinical care.

From Fig. 3a, we noted outpatients account for 69.04% of total cases and can be treated as suspects, but with mild symptoms, so they do not require hospitalization. Thus, it associates them with low-risk scenarios if they only require treatment or are monitored. There were a few high-risk cases that were not hospitalized (1.8% from outpatient), in these cases, we do not know the reason, however, it could be related to availability at the time of registration or the patient did not want to remain in the unit. Suspected cases with respiratory distress symptomatology are severe patients, and they meet the valid definition of severe acute respiratory infection for all medical units in the country. Here, we confirmed that cases of moderate-risk, and most high-risk cases, 99.96%, and 93.11%, respectively, are hospitalized (Fig. 3a).

In this way, ailing patients have to enter the different units according to their condition or when there is a reasonable medical evaluation given their clinic and epidemiological history. Therefore, there should be a basis for diagnosis and prognosis for the procedures to be followed, holding the biological factors, and the patient needs [28]. We considered the covariate Income Unit, which characterizes the service to which the patient arrived within the medical sector that provided the care. These have been labeled as Infectology, Internal Medicine, Pneumology, Intensive Care Unit (ICU), Neonatal Intensive Care Unit (ICU), Adult Emergencies, Emergency Surgery, Pediatric Emergencies, Pediatric Intensive Therapy Units (UTIP). The criteria for admission to an Intensive Care Unit (ICU) usually consider multiple factors (e.g. oxygen saturation less than 90%). From Fig. 3b, we could establish that the distribution of low-risk patients is concentrated in External Consultation and a lower proportion in Emergency Observation, which did not require hospitalization. In contrast, cases of moderate-risk and high-risk are mostly found in Internal Medicine and Adult Emergencies.

The Mexican health system comprises two sectors, namely the public and the private sectors. Within the public sector, we find the social security institutions: Mexican Institute of Social Security (IMSS), Institute of Security and Social Services of State Workers (ISSSTE), PEMEX, Defense's Secretary (SEDENA), Navy's Secretary (SEMAR); and institutions and programs that care for the population without social security, Ministry of Health (SSA), IMSS-Opportunities Program (IMSS-O), Popular Health Insurance (SPS). According to data from INEGI's 2015 Intercensal Survey, 82% of the Mexican population was affiliated to health services, and 17.3% were not insured. The insured Mexicans are distributed as follows: 49.9% are registered to the Ministry of Health (SSA), 39.2% to IMSS, 7.7% to ISSSTE, and 3.3% to the private sector. PEMEX, SEDENA, and SEMAR together provide medical services in clinics and hospitals of PEMEX and the Armed Forces for their employees, which represent 1.2% of the population with social security. The Health Sector variable of the Medical Unit designates the type of institution of the National Health System that provided the care and identifies whether it is a public or private health institution [29]. From the data, we estimate that patient distribution is especially concentrated in the Ministry of Health (SSA with 53.3%) and the IMSS (32.6%), and the rest to ISSSTE and the private and state sectors. Low and moderate-risk cases in the SSA and IMSS sectors are in proportion to the number of patients affiliated with these health service units. That is not the situation with the most severe cases (Fig. 3c).

3. Methods

In this section, the aim is to classify patients into a category using a set of features to characterize and evaluate the COVID-19 patient's risk level. Following the preliminary analysis of different subsets of data variables, observing their nature, distribution, and using that we have achieved a plausible description of three risk scenarios, we considered a multiclass classification problem. That is a pattern recognition task with more than two classes where the labels correspond to risk levels: low, moderate, and high. Also, as we have seen, they have a natural interpretation since they emerged from the disease's eventual outcome in each patient.

Classification problems, and in particular, multiclass applications do not have a single solution strategy, i.e. there is not a unique methodology or immediate procedure for dealing with all aspects of the problem. The reason is that the classification depends on the data, the choice varies depending on the objectives [30–32]. An essential and decisive issue in these tasks is to choose the right features for the training phase, which corresponds to the most demanding stage of the work [33]. Notice we have a significant size in the dataset to carry out these processes. Thus, an appropriate selection of algorithms and relevant variables will enable the implementation of our learning algorithms, avoiding expensive computational complexity. All calculations in this paper were conducted using R statistical language.

3.1. Multiclass classification

Recently, the number of ML techniques has grown. Neural Networks have become popular, and there are many other highly acceptable techniques, based on decision trees [34]. For example, Bagging allows a decrease in the variance of predictions through the combination of the results of several classification trees, each trained with different subsets of observations taken from the same population. Random Forests provide an improvement over Bagging through a slight change that decorrelates trees[2] [35,36]. Boosting or Boosting Machine is based on fitting multiple weak classifiers (simple models that predict only slightly better than expected by chance) [37]. This combination is performed sequentially in such a way that each new incorporated model attempts to correct the errors of the previous ones, improving from iteration to iteration. Gradient Boosting Machine (GBM) [38,39] is a generalization of the Boosting Machine model that allows us to apply the Gradient Descent method to optimize any loss function during model fitting. XGBoost (Extreme Gradient Boosting) [40] is an optimized distributed Gradient Boosting alternative designed to be highly efficient, flexible, and portable. It implements ML algorithms under the Gradient Boosting framework and provides a parallel tree boosting that solves many problems in a fast and accurate way.

In the following part of our research, we implemented ML algorithms to predict the variable risk level response. First, we trained an initial model via Random Forest, including all covariates, only

[2] The main difference between Bagging and Random Forests (RF) is the choice of the size of the subset of predictors.

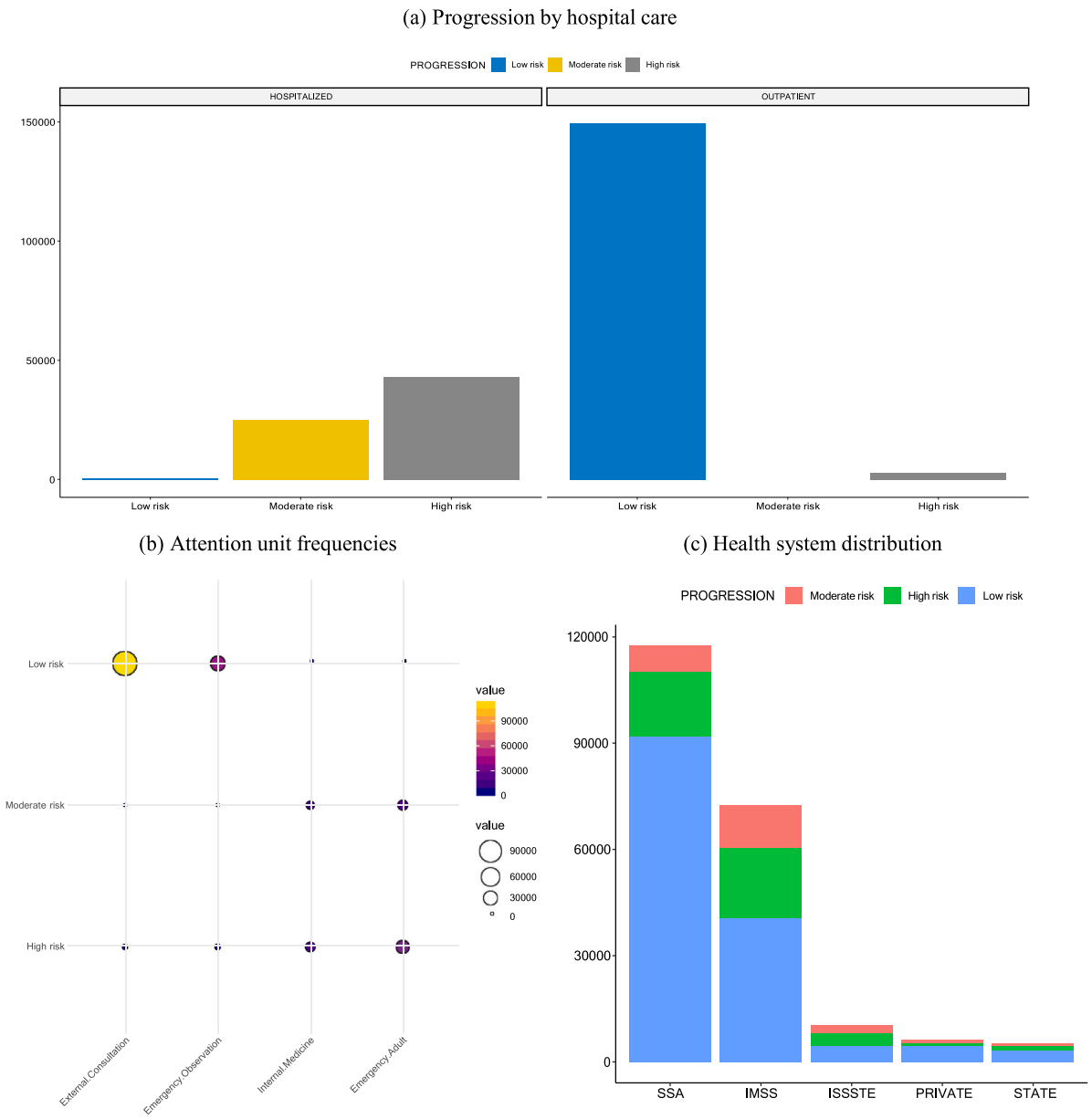


Fig. 3. Risk level by Health System features.

specifying the maximum number of trees and the complexity parameters. We used Ranger, a fast implementation of recursive partitioning from R statistical software, particularly suited for high-dimensional data. There, classification forests are implemented as in the original Breiman's Random Forest proposal [35] and the Gini index (total variance between all classes as an impurity measure) is applied as the default split rule for decision trees. Next, we used a GBM model but considering a subset of variables chosen through the Variable Selection Using Random Forests (VSURF) criteria based on a three-step selection procedure [41]. The first step is to remove non-informative variables from the training dataset. The second step considers all variables that support the interpretation of the model. The last step, the prediction step, refines the search and removes redundant variables from the set of variables selected in the interpretation stage.

After applying a GBM model, we considered a practical variant when speed and accuracy are required: the XGBoost algorithm, which implements parallel processing and is faster compared to GBM. It considers different parameters and values to be specified, therefore, XGBoost requires parameter tuning to maximize its advantages over other algorithms. In turn, the standard GBM implementation has not been regularized as XGBoost, so the latter also helps reduce overfitting. In fact, XGBoost is also known as a *regularized boosting* technique. Another plus point of XGBoost is that it can work with encoded categorical data, which is not a common property of classification algorithms. Accordingly, for the small set of influential variables in our application, we use a transformation of the features into dummy-type variables. In other words, we adopt a one-hot encoding that creates new (binary) columns, indicating the presence of each feasible value from the original data.

It is worth noting that all these methods incorporate many parameters and hyper-parameters necessary to control the algorithm's behavior. As for supervised classification procedures, it is common to divide our data into a training set, which will be used to learn the about the data and generate a prediction model; and a test set, with which we validate the generated model. However, because parameter tuning is required, data is divided into a validation set (1% from total), and the rest are split into training and test sets, whose percentages are 80% and 20%, respectively. To improve the predictive capacity of each model, we searched the optimal parameter values in the validation

set through cross-validation with five partitions. In addition, the multiclass classification approach applied is the one-vs-all strategy, which comprises fitting one (final) classifier per class. For each classifier, we tested the class against all the other classes. Besides its computational efficiency, one advantage of this approach is its interpretability. Since each class is represented by one and only one classifier, it is possible to gain knowledge about the class by inspecting its corresponding classifier. This is the most used strategy and is a fair default choice in programming routines.

4. Results

As in any predictive study, not only fitting the model is important, also assessing its ability to predict new observations. Once the algorithms were trained, we evaluated their predictive ability using the test set. Evaluation of a classification algorithm performance is measured through the confusion matrix, which contains information about the (true) actual and the predicted class. The rows are the Predicted class, and the columns are the Actual class. In the confusion matrix, TN is the number of negative outcomes correctly classified (True Negatives), TP is the number of positive outcomes correctly classified (True positives), FP is the number of negative outcomes incorrectly classified as positive (False Positives); and FN is the number of positive outcomes incorrectly classified as negative (False Negatives).

Table 1 shows the results for the confusion matrix obtained for each algorithm, allowing the comparison of predictions among different algorithms using the classification of positive cases (or reference class) against negative ones (classes opposed to the reference), based on the approach we have adopted (one vs. all). For a multi-class classification problem, the definition of True Positive is the same as in the confusion matrix of two classes. However, here we calculated the true positives for each class in the confusion matrix. Therefore, True Positives is the number of predictions in which data were correctly classified. Also, true negatives are defined as in the confusion matrix of two classes. Now for a given class, True Negatives are calculated by taking the sum of the values in each row and column, except for the current ones. The same is done for the calculation of false positives and false negatives.

Simple inspection shows an acceptable classification for the low-risks in all three outcomes. There is only a slight increase in the number of false negatives (15 classified as high risk) in the low-risk classification using Random Forest. Predictions for moderate-risk and high-risk classes show greater differences. In the GBM model, there was a decrease in the moderate-risk classification. The same occurs in greater failures in the XGBoost performance for this label. However, the XGBoost model presents the largest gain when classifying high-risk patients (8111), as the best classification of the three algorithms for this class is obtained.

The most widely used measure of classification performance, which represents the number of correctly predicted examples over all instances, is accuracy, given by

$$Acc = \frac{TP + TN}{TP + FN + FP + TN}.$$

Here, results gave an overall accuracy of 89.86% for Random Forest, 89.37% for GBM, and 89.97% for XGBoost (Table 2). Therefore, we can say that they performed well with a global effectiveness close to 90%. Essentially the classification of the low-risk group has no drawbacks. We found no differences between the measures used. Patients in this class can be adequately identified; under the three algorithms, the correct classifications are almost perfect.

There are two comparative measures in the performance of positive cases and negative cases:

- **Sensitivity**, also called True Positive Rate (TPR) or Recall. For an specific class, represents the proportion of TP over observed positive:

$$sensitivity = \frac{TP}{TP + FN}.$$

- On the other hand, **specificity** describes the proportion of the negative samples that were correctly classified:

$$specificity = \frac{TN}{FP + TN}.$$

In the moderate-risk class, we have a decrease in sensitivity ranging from 42.13% in Random Forest to 33.63% in GBM, with the lowest level of 25.43%. in XGBoost. We highlight the gradual improvement of the high-risk group classification, advancing from an 83.03% sensitivity in Random Forest to an 85.2% in GBM, and reaching the highest level of 88.38% in XGBoost.

In addition, other measures are considered in the tables, one of them on predictions and two more useful in unbalanced data settings:

- **Precision** reflects the performance of the prediction as the proportion of predicted positive samples that were correctly classified to the total number of positive predicted samples:

$$precision = \frac{TP}{TP + FP}.$$

- **Balanced accuracy** combines the sensitivity and specificity measures:

$$BA = \frac{sensitivity + specificity}{2}.$$

- F_1-score, whose formulation is:

$$F_1 = \frac{2 \cdot TP}{2 \cdot TP + FP + FN},$$

represents the harmonic mean of precision and recall or sensitivity.

It would be worth discussing the precision metric, which declines to 68.42% for the high-risk class. However, note that incorrect classifications of low risk and moderate risk groups are relevant in this case since they are part of the decisions and opportunities that we face in this problem. Other measures also involve a trade-off effect when weighing performance in reference classes and opposite classes within each approach but in general, they were stable (see Table 2).

4.1. Feature importance

This section describes how variable importance is calculated for the three approaches we have implemented. We could study the influence of the features by quantifying the change in a specific measure produced by each variable in the set of trees or base predictors composing the model. In a few words, the importance of a feature corresponds to the increase in the prediction error after modifying the value of that attribute. That is when the link between the attribute and the model output is interrupted. The importance measure automatically considers all interactions with diverse attributes. Permutations of the variable also nullify the interaction effects with diverse features, both main and relational effects of attributes are affected by such permutations.

In a Random Forest, the importance of the features can be assessed from a measure based on increased purity of nodes, representing the decrease in the impurity measure (Gini index) caused by the permutation of a variable in the trees. For each tree, the prediction accuracy on the out-of-bag portion of the data is registered. We did the same after permuting values for each predictor variable. The differences between the two accuracies are averaged over all trees and normalized by the standard error. Then variables are sorted according to their mean variable importance (VI), in decreasing order. This order is kept all along with the procedure. In Fig. 4a only the 12 most important variables for the RF implementation are shown. In this visualization, we can distinguish three groups of features. For the first group, variables related to the health system: admission unit, preliminary diagnosis, and type of patient, are the most important. They are followed by age and symptom variables with less influence: dyspnea and polypnea. Then,

Table 1
Confusion matrices.

Random Forest				GBM				XGBoost			
Predicted	*True*			*Predicted*	*True*			*Predicted*	*True*		
	Low Risk	Moderate Risk	High Risk		Low Risk	Moderate Risk	High Risk		Low Risk	Moderate Risk	High Risk
Low Risk	29927	0	**536**	Low Risk	29937	3	**546**	Low Risk	29935	0	**542**
Moderate Risk	2	2112	**1021**	Moderate Risk	1	1686	**812**	Moderate Risk	3	1275	**524**
High Risk	15	2901	7620	High Risk	6	3324	7819	High Risk	6	3738	8111

Table 2
Performance measures.

Random Forest				GBM				XGBoost			
Accuracy	89.86%			Accuracy	89.37%			Accuracy	89.97%		
	Low Risk	Moderate Risk	High Risk		Low Risk	Moderate Risk	High Risk		Low Risk	Moderate Risk	High Risk
Sensitivity	99.94	**42.13**	**83.03**		99.98	**33.63**	**85.2**		99.97	**25.43**	**88.38**
Specificity	96.22	97.38	91.66		96.13	97.92	90.47		96.18	98.65	89.29
Precision	98.24	67.37	72.32		98.2	67.47	70.13		98.22	70.75	68.42
F_1-score	99.08	51.84	77.31		99.08	44.89	76.94		99.09	37.42	77.13
BA	98.08	69.75	87.35		98.05	65.78	87.84		98.08	62.04	88.84

there are others receiving more limited influence as a state, days since the beginning of symptoms, and job occupation. Finally, comorbidities such as diabetes and hypertension appear.

The starting point for the VSURF analysis is the same as in the RF. Variables are determined by the mean variable importance in decreasing order. As a result, the VSURF method selected 14 variables at the interpretation step and 4 variables at the prediction step. Our selected model was given for the interpretation step. The most important variables are Patient, Admission Unit, Preliminary Diagnosis, Age, Dyspnea, Health System, Arthralgia, State, Diabetes, Myalgia, Polypnea, Cough, Cyanosis, Abdominal Pain. In Fig. 4b the first step plots the standard deviation of variable importance ordered according to their mean variable importance in decreasing order. The top right subgraph represents the "thresholding step", "interpretation step" and "prediction step". For interpretation, it grows embedded random forest models, starting with the random forest build with only the most important variable and ending with all variables selected in the first step. However, now the variables are added to the model in a step-wise manner. Fig. 4b also shows the mean error rate of embedded random forests models. It calculates the mean jump value using variables that have been left out by the second step.

Notice the GBM method stops splitting a node when we find a negative loss in the partition. In that sense, it is more like a greedy algorithm. Thus, the feature importance in GBM is quantified through the total reduction of quadratic error by calculating the relative influence of each variable: once that variable was selected to split on during the tree building process. As a result, it measures how much the squared error (over all weak classifiers) is increased (decreased). The basic idea is to consider a variable important if it has a positive effect on the prediction accuracy (classification). GBM randomly permutes each predictor variable at a time and computes the associated reduction in predictive performance. This is the idea of the variable importance Breiman uses for RF [35], instead GBM runs the entire training dataset (not the out-of-bag observations). The method uses the same approach as a single tree but sums the importance over each boosting iteration. With a similar plot, Fig. 4c scales the importance scores to be between 0 and 100. This option is used to make the image more readable. From the subset of features obtained from VSURF, the order of the variables is very similar to that of the RF algorithm. In this way, there is some consistency in the perception of variables that are relevant for classification purposes only. Any other interpretation requires another type of analysis. There is no causal effect, nor have correlations been thoroughly explored.

On the contrary, XGBoost splits to a specified maximum depth and then prunes the tree backwards eliminating partitions beyond which there is no positive gain. Here gain refers to the improvement in accuracy brought to the branches by the features. The idea is that before adding a new split on a specific feature to the branch there were some wrongly classified elements; after including the split on this feature, there are two new more accurate branches. The gain scores are given as relative scores to the most relevant variable. The most important feature (as reference) will have a score of 1 and the gain scores of the other variables will be scaled, precisely, over the gain score of reference. Recall that for the XGBoost algorithm, we have performed a one-hot encoding transformation. This modification has been useful since results appear in a more specific way, allowing identification of relevant variables. We have the dichotomic variables corresponding to the Level Hospitalized Patient Type, Age, Preliminary Diagnosis under the label Influenza Disease, and no evidence of dyspnea appears. We find three other variables related to the health system (External Consultation, SSA, and IMSS health institutions), the feature Mexico City from the variable State; and two more indicator variables on the presence of polypnea symptoms and the absence of recurrent disease, diabetes, respectively (see Fig. 4d).

5. Discussion

Indeed, one aspect that we can highlight from the various studies on the COVID-19 pandemic is that it has driven unprecedented technological developments. Many of these relate to Artificial Intelligence and Machine Learning, and their interaction with diverse areas of knowledge, whether medical, social or economic; including the sciences of massive data and computational analysis [10,13,42]. Amidst this scenario, the health sector will need to incorporate these resources into its analysis and diagnosis support systems, not only of infectious diseases but of any other nature. Therefore, with these contributions, we aim to improve the care of each patient or community and prepare society for any unforeseen eventualities in the future. Such advances could help health centers reduce operating costs of various kinds, where diagnostic and response time plays a key role in addressing situations in a context of uncertainty.

The characteristics associated with SARS-CoV-2 infection may be overlooked in the presence of other well-known symptoms and comorbidities associated with serious disorders. Although SARS-CoV-2 infection can cause many symptoms, it remains to be decisively established whether the disease has susceptibility for specific comorbidities and whether there are predominant symptoms [6]. For Mexico, some implications of our statistical analysis are that at first sight groups of patients whose particularity encompasses the evolution of the disease

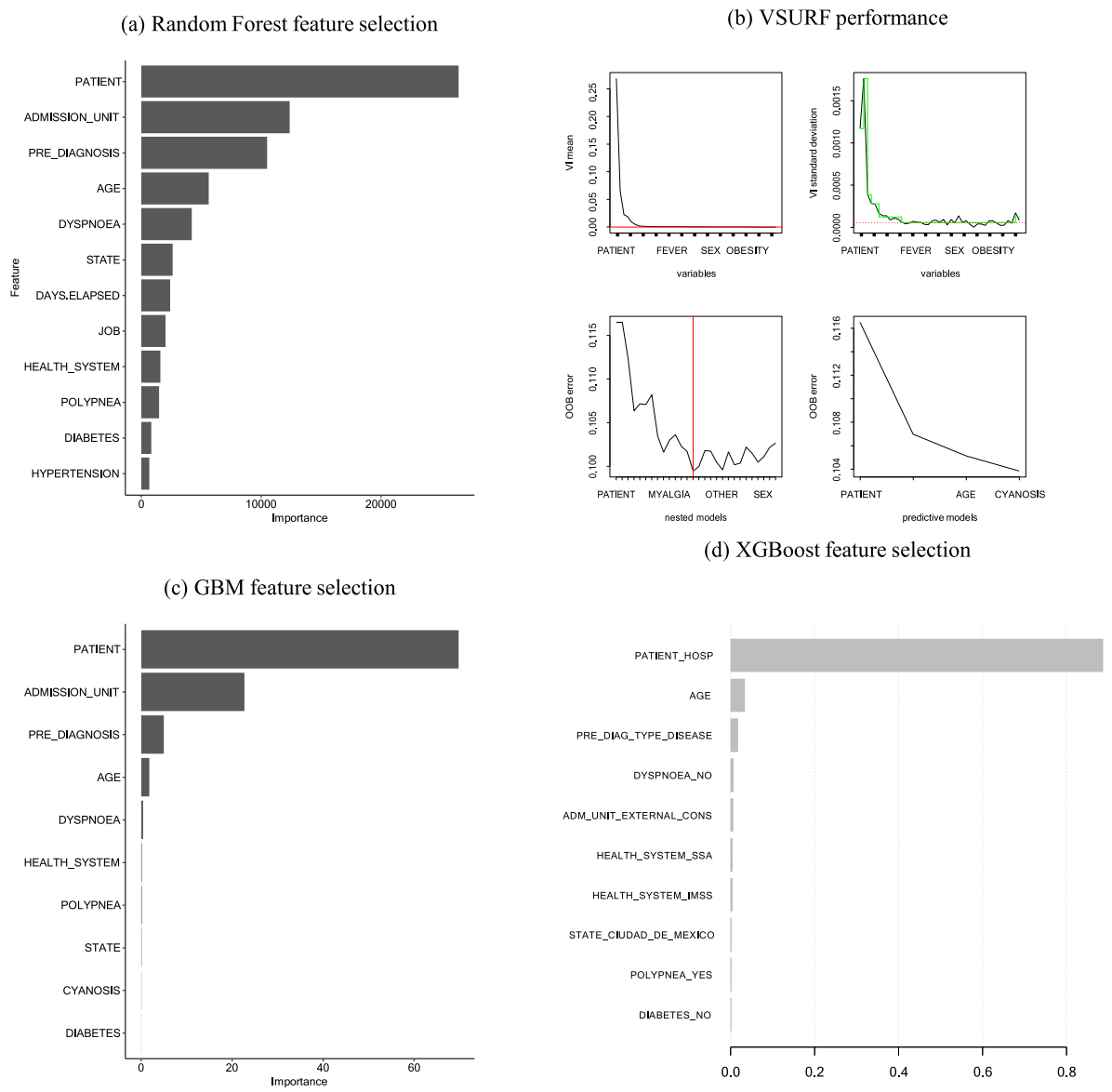


Fig. 4. Training algorithms: Feature importance.

can be discriminated through their association with age, sex, and days from the beginning of symptoms. Given the progression of the positive patient, we proposed three clusters, leading to three risk scenarios: low (treatment and monitoring), moderate (recovered and non-severe), and high (severe and deaths). A constructive aspect is that results shown in Section 2 are consistent with those reported worldwide. Particularly that COVID-19 is associated with the following (major) signs and symptoms: Cough, fever or headache (headache), dyspnea, polypnea, vomiting or diarrhea; accompanied by at least one of these symptoms (secondary): shortness of breath, joint pain (arthralgias), muscle pain (myalgia), sore throat/burning (odynophagia), nasal runoff (rhinorrhea) difficulty breathing, redness of eyes (conjunctivitis), chest pain, loss of smell (anosmia) and taste (dysgeusia) and chills. If we focus on major symptoms, high frequency of dyspnea and polypnea in high-risk patients is regularly reported. However, headache is a prevalent symptom in low-risk patients. The observation and timely record of all this information, together with the lack of oxygenation, are crucial to prevent critical stages of infection with this virus.

According to the evidence found so far, there are indications that the prevalence of previous pathologies are influential factors in severe and complicated forms of COVID-19. Thus, older people and those with chronic diseases, the most predominant being high blood pressure, heart disease, diabetes, and obesity, report severe cases of the disease more often than others [43]. Unfortunately, in Mexico, the incidence of overweight and obesity among the Mexican population has reached alarming levels. Mexico has one of the highest rates of obesity in the OECD, with 73% of the Mexican population overweight. Additionally, 34% of obese people suffer from morbid obesity. Since 2000, diabetes has been the leading cause of death among women and the second among men. The Ministry of Health states that one in four Mexican adults has high blood pressure, i.e. 25.5% of the population. Approximately 40% ignore the condition, further affecting their health. Furthermore, from the 60% who are aware of the diagnosis, only half of them are under control. The combination of these three comorbidities has a latent effect on both the probability to get COVID-19 as well as to have a more severe outcome [26]. Thus, the control of these chronic conditions remains an enormous challenge to avoid high mortality in the Mexican population. Based on the 2020 Mortality Report by the Statistical Information Institution INEGI, COVID1-9 was the second most common cause of death in that year [44]. In brief, patients positive for SARS-CoV-2, who are men at age 60 and older, people living with diseases such as hypertension or diabetes, having symptoms

such as fever, cough, dyspnea, and cephalalgia, are considered higher exposure groups to dangerous *outcome* complications. In their case, immediate medical attention is suggested.

Furthermore, the severity criteria related to COVID-19 disease are analyzed from a demographic perspective and using clinical records. Advanced age is severity criteria (>60 years); and comorbidities, with diabetes and high blood pressure being the most common, followed by diabetes and obesity. These scales cannot replace clinical judgment but may be useful for preliminary diagnosis. Confirmed cases with major symptoms and other severity criteria, type of patient, medical admission unit, and disease progression, are also examined. Finally, the number of deaths is 1.64 times higher in IMSS. However it is worth nothing that IMSS has 27.57% of high risk patients, while SSA has only 15.41%. Patient risk increases significantly within the IMSS health system, but we must consider the hospital conversion processes carried out by the authorities. These implemented exclusive centers to treat COVID-19 patients, hospitals that were mostly sponsored by the IMSS. Finally we note that the division of the health system in Mexico has led to a highly centralized care for COVID-19 patients, while coverage has not been sufficient and in many places, demand has outsized hospital capacity.

In comparison to other studies in Mexico, in [45] an analysis of data from the same source on a different timeline explored risk factors for lethality in COVID-19, including diabetes, obesity, chronic obstructive pulmonary disease, advanced age, hypertension, immunosuppression, and chronic kidney disease (CKD). Their results showed that diabetes, obesity, and comorbidity burden change risk profiles in patients with COVID-19 Mexico and significantly improve the prediction of mortality related to COVID-19. Specifically, early onset diabetes confers a higher risk of intensive care unit admission and intubation. A common thread with our analysis is the predictive pattern of increased lethality of COVID-19 across gender, age older than 65 years, diabetes, hypertension and obesity. For age under 40 years, knowledge of CKD, hypertension, and immunosuppression enables discrimination of fatal COVID-19 cases from non-lethal ones.

Using data on smaller patient samples, in [46], logistic regression and Cox survival analysis models were fitted to estimate the association between hospitalization and mortality mediated by other covariates. Here, the presence of either diabetes, hypertension, or obesity was statistically significant when compared with patients not having those chronic disorders. As in our case, patients suffering from a combination of diabetes and hypertension had a higher risk of suffering a severe disease, a higher probability of being hospitalized, and a higher probability of death. In addition men were approximately 1.54 times more likely to be hospitalized than women and for the 50–74 age groups they were clearly more likely to be hospitalized than people aged 25–49.

Although they reported the lack of further information on light (asymptomatic) or moderate COVID-19 cases they argued a higher death risk for a profile similar as the one we described, namely: male; being in an older age group; having chronic renal disease, COPD, or a combination of comorbidities; hospitalization developing pneumonia; intubation, ICU admission or being treated in a public health institution.

In a prospective cohort study among patients with confirmed COVID-19 disease cared for in a hospital care center in Mexico City, Olivas et al. (2022) [47] found that in-hospital mortality was 30.1%, and 49.2% in ICU beds. Furthermore, it was detected that the risk of in-hospital death was significantly higher in males than in females, and in obese than in non-obese patients, as well as in diabetics than in non-diabetics.

Comparing their results to ours, the similarities lie in indicating greater severity in the face of factors such as comorbidities, especially diabetes and obesity (the degree of the latter being directly proportional to mortality), male sex, increased inflammatory markers and laboratory findings related to organ failure. All were associated with an increased risk of in-hospital death.

Thus, in summary, factors associated with increased risk of death in COVID-19 cases were male age higher than 65 years with diabetes, hypertension, obesity, CKD, COPD, and immunosuppression.

About the information system and the decision-making plan, we recognize that sociodemographic information, clinical history, and symptomatology description have provided medical staff with the indispensable tools, first, to determine the hospitalization of the patient. Then, to drive them to the appropriate admission unit so that the patient can receive the correct medical attention and care. Hence these variables are reliable to discriminate the risks to which most patients are exposed. After our exploratory statistical analysis of different subsets of features, considering their biomedical nature and distribution, we proposed a multiclass classification problem. It is interesting and adapts to the risk assessment from the information record of each patient. Linked to the previous point, our classification schemes benefit from the characterization of the risk levels we have decided on. Using three scenarios has certainly reduced the variability of classifiers and lead to a better probabilistic output.

As seen in Tables 1 and 2 the results in classifiers essentially lie in the predictions of moderate-risk and high-risk labels. Let us distinguish between false positives and false negatives under the following premise. Note that while there is no specific order between classes, in terms of interpretation, severity tells us that there is an imbalance between predicting the label of a low-risk patient as moderate or high-risk, and a high-risk labeled as low risk. In this type of application, we find structural and ethical dilemmas, where criteria should be used to evaluate certain decisions. This could have implications for diagnosis, treatment, and financial costs, but most importantly, human resources clearly involve a patient's life. Then, building a cautious framework on predictions, an ideal strategy could be based on the target of reducing as much as possible the negative false classifications of high risks (in bold in the confusion tables). Since the classifications of low-risk patients are homogeneous in all three algorithms and perform accurately, perhaps telling a patient that they have a low or moderate risk while being of high risk, in reality, causes the most serious error. Therefore, under certain conditions, not very restrictive and assuming some costs, it is still affordable to take the opposite direction. That is classifying patients to a higher risk step, as long as the actions are only to have greater care, contact their hospital care, as well as strengthen their treatment and monitoring. All these are surely nontrivial and time-enhancing. However, it is fundamental that as a comprehensive society we coordinate to discuss and maintain consensus to recover from this difficulty in the immediate future.

Our results demonstrate specific patterns and characteristics change risk profiles in patients with COVID-19 Mexico and significantly improve the prediction of risk scenarios related to severe COVID-19 and its lethality.

6. Conclusions

The monitoring and study of this pandemic naturally involves the quantification of concerning aspects whose impact on society goes from the number of people infected, hospitalizations, deaths to even economics, such as job losses, increased poverty, and social needs coverage such as access to medical services and vaccination programs, among many others. Many disciplines, as AI and ML techniques through data analysis, diagnostic automation, and pattern recognition applications, have been involved in this work. This is where our proposal comes in, which uses a multiclass classification approach to evaluate different risk incidences in COVID-19 patients based on a set of well-defined features.

We also intend this approach to serve as a guide, to find methods that exploit the capacity of learning algorithms to find new insights from the data. Therefore, when pieces of information with a large number of observations and an output variable of interest are available, the opportunity to train algorithms that direct attention to predicting scenarios based on subsets of relevant characteristics should be

considered. Another interesting aspect is that we can determine in probabilistic terms who may have serious COVID-19 outcomes once the diagnosis is given. Within this type of research, these algorithms will be able to spread and apply to other types of diseases, effectively contributing to improving public health services.

The fundamental contributions of this work are the exploitation of a valuable database, from which we address the problem of systematically identifying patterns to detect and evaluate risky and non-risk factors and scenarios in COVID-19 patients. Here, we showed that multiclass classifiers are favorable under different conditions, we observed satisfactory overall performance for the three we implemented, with the following Accuracy (test) metrics: Random Forest (89.86%), GBM (89.37%) XGBoost (89.97%). In addition, we discovered statistically significant features for classification. Variables such as age and those that come from the medical triage (patient type, health system, and admission unit) lead our algorithms. The early detection of symptoms such as dyspnea and polypnea is essential. However, it is also necessary to know if there is a record of comorbidities such as diabetes and hypertension. Elapsed days from the beginning of symptoms are crucial to determining attention during action courses.

As a last remark, this research represents one of the primary efforts to assess risks for patients with COVID-19 under multiclass supervised learning rules. The purpose is to assist people and health authorities in the mitigation of the COVID-19 pandemic. Knowing which positive patients are riskier to get to critical disease stages, it is possible to execute action plans aimed at focusing on the immediate challenges of this emergency.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Funding

This research was supported by DGAPA postdoctoral scholarship and PAPIIT project IN118720, UNAM.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://covid-19.iimas.unam.mx/home.

References

[1] World Health Organization, WHO Director-General's. Remarks at the media briefing on 2019-nCoV on 11 2020. 2020, https://t.co/WAxG44ufyi.

[2] Sohrabi C, Alsafi Z, O'Neill N, Khan M, Kerwan A, Al-Jabir A, et al. World health organization declares global emergency: A review of the 2019 novel coronavirus (COVID-19). Int J Surg 2020;76:71–6. http://dx.doi.org/10.1016/j.ijsu.2020.02.034.

[3] CDC. Symptoms of coronavirus (COVID-19) [Internet], centers for disease control and prevention. 2020, [cited 2020 sep 11]. Available from: https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/symptoms.html.

[4] Larsen JR, Martin MR, Martin JD, Kuhn P, J.B. Hicks. Modeling the onset of symptoms of COVID-19. Front Publ Health 2020;8(473). http://dx.doi.org/10.3389/fpubh.2020.00473.

[5] Morris SB, Schwartz NG, Patel P, et al. Case series of multisystem inflammatory syndrome in adults associated with SARS-CoV-2 Infection – United Kingdom and United States. MMWR Morb Mortal Wkly Rep March-69 (2020), 2020, p. 1450–6. http://dx.doi.org/10.15585/mmwr.mm6940e1.

[6] Jordan RE, Adab P, Cheng KK. Covid-19 risk factors for severe disease and death. BMJ 2020;368(1198). http://dx.doi.org/10.1136/bmj.m1198.

[7] Wu Z, McGoogan JM. Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China: Summary of a report of 72 314 cases from the Chinese center for disease control and prevention. JAMA 2020;323(13):1239–42.

[8] Pan American Health Organization / World Health Organization. Epidemiological update: Coronavirus disease (COVID-19). [26 2020], Washington, D.C. 2020, PAHO/WHO.

[9] Juárez-Hernández F, García-Benítez MP, Hurtado-Duarte AM, et al. Hallazgos tomográficos en afectación pulmonar por COVID-19 experiencia inicial en el instituto nacional de enfermedades respiratorias ismael cosío villegas, méxico. Neumol Cir Torax 2020;79:71–7. http://dx.doi.org/10.35366/94630.

[10] Verdonk C, Verdonk F, Dreyfus G. How machine learning could be used in clinical practice during an epidemic. Critical Care 2020;24(1):265. http://dx.doi.org/10.1186/s13054-020-02962-y.

[11] Melek M. Diagnosis of COVID-19 and non-COVID-19 patients by classifying only a single cough sound. Neural Comput Appl 2021;30:1–12. http://dx.doi.org/10.1007/s00521-021-06346-3, Epub ahead of print. PMID: 34345119; PMCID: PMC8323961.

[12] Rahman T, Ibtehaz N, Khandakar A, Hossain MSA, Mekki YMS, Ezeddin M, et al. QUCoughScope: An intelligent application to detect COVID-19 patients using cough and breath sounds. Diagnostics 2022;12(920). http://dx.doi.org/10.3390/diagnostics12040920.

[13] Albahri AS, Hamid RA, Alwan Jk, et al. Role of biological data mining and machine learning techniques in detecting and diagnosing the novel coronavirus (COVID-19): A systematic review. J Med Syst 2020;44(122). http://dx.doi.org/10.1007/s10916-020-01582-x.

[14] Ahamad MM, Aktar S, Md Rashed-Al-Mahfuz M, Uddin S, Lio P, Xu H, et al. A machine learning model to identify early stage symptoms of SARS-CoV-2 infected patients. Expert Syst Appl 2020;160:160–85, 10.1016%2Fj.eswa.2020.113661.

[15] Li W, Ma J, Shende N, et al. Using machine learning of clinical data to diagnose COVID-19 a systematic review and meta-analysis. BMC Med Inform Decis 2020;20(247). http://dx.doi.org/10.1186/s12911-020-01266-z.

[16] Jiang X, Coffe M, Bari A, Wang J, Jiang X, Huang J, et al. Towards an artificial intelligence framework for data-driven prediction of coronavirus clinical severity. Comput Mater Contin (CMC) 2020;63(1):537–51. http://dx.doi.org/10.32604/cmc.2020.010691.

[17] Assaf D, Gutman Y, Neuman Y, et al. Utilization of machine-learning models to accurately predict the risk for critical COVID-19. Intern Emerg Med 2020;15:1435–43. http://dx.doi.org/10.1007/s11739-020-02475-0.

[18] Burdick H, Lam C, Mataraso S, Siefkas A, Braden G, Dellinger RP, et al. Prediction of respiratory decompensation in Covid-19 patients using machine learning: The READY trial. Comput Biol Med 2020;124. http://dx.doi.org/10.1016/j.compbiomed.2020.103949.

[19] RCore Team. R: A language and environment for statistical computing. 2013, http://www.R-project.org/.

[20] Secretaría de Salud (SSA). Criterios para las poblaciones en situación de vulnerabilidad que tienen mayor riesgo de desarrollar una complicación o morir por Covid-19 en la reapertura de actividades económicas en los centros de trabajo. From: https://coronavirus.gob.mx/wp-content/uploads/2020/08/Criterios_Vulnerabilidad_12Ago2020.pdf.

[21] Adams RB. Gender equality in work and COVID-19 deaths. Covid Econ 2020;11(16):23–60. http://dx.doi.org/10.2139/ssrn.3601651.

[22] Bhopal SS, Raj Bhopal R. Sex differential in COVID-19 mortality varies markedly by age. The Lancet 2020;396:532–3. http://dx.doi.org/10.1016/S0140-6736(20)31748-7.

[23] CDC. Older adults (COVID-19) [internet], centers for disease control and prevention. 2020, [cited 2020 Apr 26]. Available from: https://www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/older-adults.html.

[24] Lieberman NAP, Peddu V, Xie H, Shrestha L, Huang ML, et al. In vivo antiviral host transcriptional response to SARS-CoV-2 by viral load sex, and age. PLOS Biol 2020;18(9):e3000849. http://dx.doi.org/10.1371/journal.pbio.3000849.

[25] CDC. People with certain medical conditions (COVID-19) [internet], centers for disease control and prevention. 2020, [cited 2020 Jul 22]. Available from: https://www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/people-with-medical-conditions.html.

[26] Shamah-Levy T, Vielma-Orozco E, Heredia-Hernández O, Romero-Martínez M, Mojica-Cuevas J, Cuevas-Nasu L, et al. Encuesta nacional de salud y nutrición 2018-19 resultados nacionales cuernavaca. México: Instituto Nacional de Salud Pública; 2020, https://ensanut.insp.mx/encuestas/ensanut2018/doctos/informes/ensanut_2018_informe_final.pdf.

[27] Meyer D, Zeileis A, Hornik K. The strucplot framework: Visualizing multi-way contingency tables with vcd. J Stat Softw 2006;17(3):1–48.

[28] Secretaría de Salud. Subsecretaría de prevención y promoción de la salud, dirección general de epidemiología. Lineamiento estandarizado para la vigilancia epidemiológica y por laboratorio de la enfermedad respiratoria viral; 2020, https://coronavirus.gob.mx/wp-content/uploads/2020/06/Lineamiento_VE_Lab_enfermedad_respiratoria_viral-_20052020.pdf.

[29] World Health Organization, Regional Office for Europe, European Observatory on Health Systems and Policies, González Block MA, Reyes Morales H, Cahuana Hurtado L, Balandrán A, Méndez E, Allin S. Mexico: health system review. world health organization. Reg Off Eur Health Syst Transition 2020;22(1). https://apps.who.int/iris/handle/10665/334334.

[30] Devroye L, Györfi L, Lugosi G. A probabilistic theory of pattern recognition. Springer-Verlag New York; 1996.

[31] Abe S. Support vector machines for pattern classification. Springer-Verlag London; 2005.

[32] Bishop CM. Pattern recognition and machine learning. Springer Science and Business Media; 2006.

[33] Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res 2003;3:1157–82.

[34] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. second ed.. Springer; 2008.

[35] Breiman L. Bagging predictors. Mach Learn 1996;24:123–40.

[36] Breiman L. Random forests. J Mach Learn Archive 2001;45:5–32.

[37] Schapire R, Freund Y. Boosting: foundations and algorithms. The MIT Press; 2012.

[38] Friedman JH. Greedy function approximation: A gradient boosting machine. Ann Statist 2001;29(5):1189–232.

[39] Natekin A, Knoll A. Gradient boosting machines, a tutorial. Front Neurobiot 2013;7:21. http://dx.doi.org/10.3389/fnbot.2013.00021.

[40] Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International conference on knowledge discovery and data mining (KDD '16). Association for computing machinery. 2016, p. 785–94. http://dx.doi.org/10.1145/2939672.2939785.

[41] Genuer R, Poggi JM, Tuleau-Malot C. VSURF: an R package for variable selection using random forests. R Journal 2015;7(2):19–33.

[42] Ozturk T, Talo M, Yildirim EA, Baloglu UB, Yildirim O, Acharya RU. Automated detection of COVID-19 cases using deep neural networks with X-ray images. Comput Biol Med 2020;121. http://dx.doi.org/10.1016/j.compbiomed.2020.103792.

[43] Li X, Xu S, Yu M, et al. Risk factors for severity and mortality in adult COVID-19 inpatients in Wuhan. J Allergy Clin Immunol 2020;46(1):110–8. http://dx.doi.org/10.1016/j.jaci.2020.04.006.

[44] INEGI. México [Internet], Características de las defunciones registradas en México. 2020, [cited 2021 jan 27]. Available from: https://www.inegi.org.mx/contenidos/saladeprensa/boletines/2021/EstSociodemo/DefuncionesRegistradas2020_Pnles.pdf.

[45] Bello-Chavolla OY, Bahena-López JP, Antonio-Villa NE, Vargas-Vázquez A, González-Díaz A, Márquez-Salinas A, et al. Predicting mortality due to SARS-CoV-2 a mechanistic score relating obesity and diabetes to COVID-19 outcomes in Mexico. J Clin Endocrinol Metab 2020;105(8). http://dx.doi.org/10.1210/clinem/dgaa346, dgaa346.

[46] Carrillo-Vega MF, Salinas-Escudero G, García-Peña C, Gutiérrez-Robledo LM, Parra-Rodríguez L. Early estimation of the risk factors for hospitalization and mortality by COVID-19 in Mexico. PLoS One 2020;15(9):e0238905. http://dx.doi.org/10.1371/journal.pone.0238905.

[47] Olivas-Martínez A, Cárdenas-Fragoso JL, Jiménez JV, Lozano-Cruz OA, Ortiz-Brizuela E, et al. In-hospital mortality from severe covid-19 in a tertiary care center in mexico city; causes of death, risk factors and the impact of hospital saturation. PLoS One 2022;17(5):e0269053. http://dx.doi.org/10.1371/journal.pone.0269053.