Using Machine Learning to support health system planning during the COVID-19 pandemic: a case study using data from São José dos Campos (Brazil)

Leila Abuabara
Universidade Federal de São Paulo – UNIFESP,
Instituto Tecnológico Aeronáutico – ITA,
São José dos Campos (SP), Brasil
leila.abuabara@unifesp.br

Maria Gabriela Valeriano
Universidade Federal de São Paulo – UNIFESP,
Instituto Tecnológico Aeronáutico – ITA,
São José dos Campos (SP), Brasil
maria.valeriano@ga.ita.br

Carlos Roberto Veiga Kiffer
Infectious Diseases Discipline,
Universidade Federal de São Paulo – UNIFESP,
Escola Paulista de Medicina – EPM,
São Paulo (SP), Brasil
carlos.rv.kiffer@gmail.com

Horácio Hideki Yanasse
Universidade Federal de São Paulo – UNIFESP,
São José dos Campos (SP), Brasil
horacio.yanasse@unifesp.br

and

Ana Carolina Lorena
Instituto Tecnológico Aeronáutico – ITA,
São José dos Campos (SP), Brasil
aclorena@ita.br

Abstract

Many efforts were made by the scientific community during the COVID-19 pandemic to understand the disease and better manage health systems' resources. Believing that city and population characteristics influence how the disease spreads and develops, we used Machine Learning techniques to provide insights to support decision-making in the city of São José dos Campos (SP), Brazil. Using a dataset with information from people who undergo the COVID-19 test in this city, we generated and evaluated predictive models related to severity, need for hospitalization and period of hospitalization. Additionally, we used SHAP (SHapley Additive exPlanations) values for models' interpretation of the most decisive attributes influencing the predictions. We can conclude that patient age linked to symptoms such as low saturation and respiratory distress and comorbidities such as cardiovascular disease and diabetes are the most important factors to consider when one wants to predict severity and need for hospitalization in this city. We also stress the need of a greater attention to the proper collection of this information from citizens who undergo the COVID-19 diagnosis test.

Keywords: Predictive models, COVID-19, Brazil, severity, hospitalization.

Resumo

Muitos esforços foram despendidos pela comunidade científica durante a pandemia de COVID-19 a fim de compreender a doença e gerenciar melhor os recursos disponíveis nos sistemas de saúde. Acreditando que as características da cidade e da população influenciam na forma de propagação e no desenvolvimento da doença, usamos ferramentas de Aprendizado de Máquina para fornecer *insights* para o suporte à tomada de decisão na cidade de São José dos Campos (SP), Brasil. Usando uma base de dados com informações provenientes das pessoas que fizeram o teste de COVID-19 nesta cidade, foram gerados e avaliados modelos preditivos relacionados a gravidade, necessidade de hospitalização e quantidade de dias de internação. Além disso, valores SHAP (explicações aditivas SHapley) foram usados para a interpretação dos atributos mais decisivos que influenciavam as predições dos modelos. É possível concluir que a idade do paciente atrelada a sintomas como baixa saturação e desconforto respiratório e comorbidades como doença cardiovascular e diabetes são os fatores mais importantes a se considerar quando se deseja prever gravidade e necessidade de internação nesta cidade. Destacamos também a necessidade de uma maior atenção à coleta adequada de informações dos cidadãos que se submetem ao teste para diagnóstico de COVID-19.

Palavras-chaves: Modelos preditivos, COVID-19, Brasil, gravidade, internação.

1 Introduction

The COVID-19 pandemic has led scientists from different areas to use their knowledge to answer a wide range of questions. More than a year after the start of the pandemic, a certain volume of data has been accumulated which can be useful in different studies to support the decision process of managers and public policies makers. This data can be used to answer questions related to medical care planning for the population and for the direction of resources to fight the disease more effectively.

Whenever a citizen takes a COVID-19 test in Brazil, a set of information about the symptoms developed and comorbidities he/she has is registered into governmental systems. The health department of the municipality of São José dos Campos gathers such data from the city citizens in a daily basis and analyze them to support decision making regarding hospital, health infrastructure and professionals needed, among others. The objective of this work is to analyze such data using more sophisticated techniques and to extract useful information for supporting the management of medical-hospital assistance resources. Specifically, we build predictive models using Machine Learning (ML) techniques to address the following *research questions*:

RQ 1: Which patients, among those who *tested positive*, will probably develop a **serious health condition**?
The objective here is to predict whether a citizen with a positive diagnosis for COVID-19 will develop a serious condition, requiring greater medical attention and the reserve of hospital resources. Our proxy for a serious condition is an hospitalization stay longer than 10 days or death. The predictive model has two possible outcomes for a given case: serious or non-serious.

RQ 2: Which patients, among those who *tested positive* for COVID-19, will need to **be hospitalized**?
The intention here is to predict whether a citizen positively diagnosed for COVID-19 will require hospitalization in either Intensive Care Unit (ICU) and non-ICU beds. This question is similar to RQ 1, but it also includes short and medium term hospitalization stays and disregards citizens' deaths. The possible outcomes are: requires hospitalization or not.

RQ 3: Among patients who were *hospitalized*, **how long** will they stay hospitalized?
The objective here is to predict the period of hospitalization of a citizen. Three possible outcomes are considered: short (up to 5 days), medium (6 to 10 days) and long (more than 11 days) terms.

The previous research questions are useful for estimating resources needed in hospitals, health centers and health units to fight the COVID-19 pandemic. RQ 1 can support professionals on focusing attention to some particular cases with higher chances of developing serious conditions, RQ 2 and RQ 3 can support bed demand planning in hospitals and health centers. Indeed, data already collected regularly by the city's health department is used for building the required predictive models.

Next section presents the main related work on how cities are fighting the pandemic using a data analytics approach. Section 3 describes our data and how it was organized and processed in order to give the views we need to answer the research questions RQ 1 to RQ 3. In Section 4 we present the main results achieved and discuss them. In Section 5 we conclude our work pointing out contributions, limitations and future studies.

2 Literature Review

This section covers two main topics. First, the location and main characteristics of the city of São José dos Campos are presented. Next we present some related work on initiatives for fighting COVID-19 at the city level. While some of them also use ML models, others use standard statistics techniques to analyze the data.

2.1 The city in the focus

São José dos Campos (for simplification purposes, here referred as 'SJC') is a city located at the São Paulo State in Brazil, about 81 km east from São Paulo city. The population is estimated at 729,731 people (1), being the fifth most populous city in the State. Figure 1 shows the location of SJC from the political map of Brazil. The city is located at the metropolitan region of the Paraíba river valley, which extends between the cities of São Paulo and Rio de Janeiro. It has an industrial economy with focus on aeronautics and a high human development index (HDI) according to Brazilian standards. Many universities and research centers are also located in this city, making it an engineering hub in Brazil.

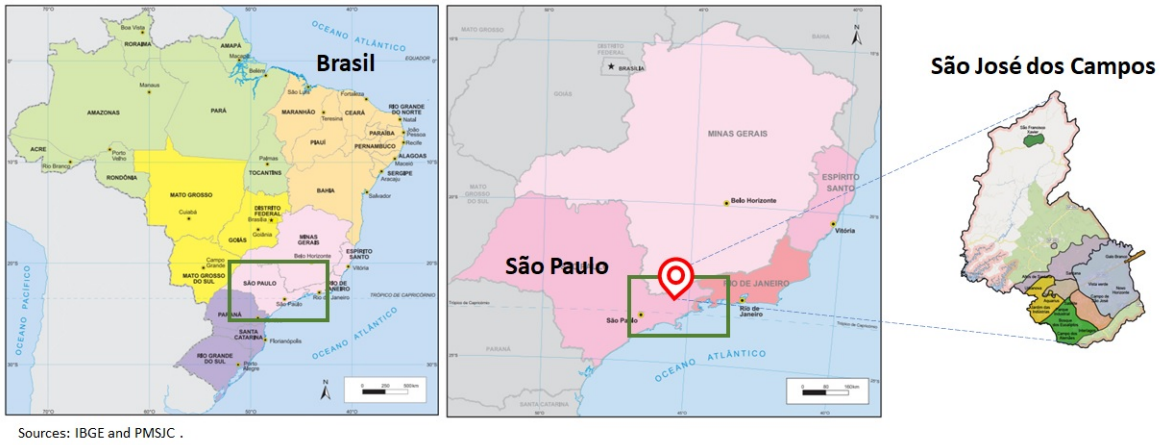

Figure 1: São José dos Campos city location (elaborated by the authors).

Important roads permeate SJC: one that links to the coast, another that links to the south of Minas Gerais State and a federal highway that links São Paulo to Rio de Janeiro, two important capitals in Brazil, making SJC to be labelled as a 'transition city'. This geographic and strategic position can also mean a population very susceptible to infection by COVID-19 according to recent studies (2) which also points the GRU airport (74 km from SJC) as one of the main entry points of SARS-CoV-2 into the country.

COVID-19 variants, which are mutant forms of the virus, have been similarly and fast wide-spreading through these same routes. The P1 lineage (called 'gamma') emerged in November 2020 in Northern Brazil and has quickly spread to other regions, such as the Southeast (3). Gamma is the predominant virus strain in Brazil and likely in the SJC region during the period covered in our research. Nonetheless, since the variant type is usually not identified during COVID-19 tests, the predominant strand in the city is hard to assess and many variants are possibly present in our data.

2.2 Related Works

Since the beginning of the pandemic, many studies in various fields of knowledge have developed decision support systems and data analysis to fight the outbreak. Each city in the world was impacted in different ways during this period. And studies at a city level were performed by local researchers to understand specific needs and support local decisions.

There are studies focusing on health systems such as testing the virus seroprevalence in medical staff inside a hospital in Lyon (France) (4); checking barriers to healthcare utilization by international students during the pandemic period in Ankara (Turkey) (5), and analysing the treatment offered to residents in Salford (UK) (6). Other researches are in fields such as economy by evaluating government (full and partial) lockdown effectiveness in Vienna (Austria) (7), also associating economic impacts to regions with distinct HDI (Human Development Indexes) in Santiago (Chile) (8), and analyzing ongoing lockdown decisions through dynamic models calibrated with real data from the New York City (USA) (9). These studies are very recent, and their main contributions are listed and summarized in Table 9 of the **Appendix** section.

Many of the previous studies analyzed the three month period known as pandemic *first wave*, the time course between March and May 2020 (4; 7; 6), as the demand for answers was urgent and used to sustain public policy actions. Our work differentiates from the previous initiatives by taking data routinely collected every time a citizen is tested for COVID-19 in the city of São José dos Campos and building some predictive models using ML techniques which can support decision making regarding the need for medical care and hospital beds.

3 Dataset and Methodology

In this session we detail two important components of our work: first we describe our dataset, second we detail the steps of the methodology adopted to build the ML predictive models.

3.1 Understanding the data source

In this subsection we describe how the raw dataset is composed before any pre-processing activity. First, we list the types of COVID-19 testing that are available for citizens in public and private facilities such as pharmacies, clinical laboratories, emergency care units and hospitals in SJC:

- **Antibody test** (also referred to as **Serology**): used for detecting any previous COVID-19 infection. It is performed using a blood sample of the person after infection. The result is the *immune response* to the infection (10).

- **PCR test**: the examination material (sample) is collected by rubbing the nasopharyngeal cavity using a special cotton swab. The test detects the SARS-CoV-2 virus in upper and lower respiratory specimens, meaning an *active infection*.

Second, it should be clear that the set of tests we are using in this work do not come from any *mass testing program* but from a **voluntary and protocol testing** of citizens. From a purposive sampling of citizens who have undergone the COVID-19 test, we tried to understand the main reasons why the test was taken. Figure 2 summarizes the results. Basically, we have two main groups of citizens, one composed by **symptomatic** people who took the test due to the presence of any suspicious symptom such as fever, cough, loss of taste or smell, and so on, to check if there is an active infection in order to seek for medical care and self-isolation. The second group is generally composed by **asymptomatic** people who took the test to meet certain protocols required by schools (e.g., to return to traditional face-to-face classes after an infection period); by companies (e.g., to certify that there is no active infection before visiting a supplier or performing a group work; or to return to face-to-face work after an infection period); by hospitals (e.g., as a pre-surgical procedure); by nursing homes (e.g., to leave an elderly relative for a day care); or even because the person had close contact with someone diagnosed with COVID-19. Additionally, for personal reasons before **visiting relatives** and friends, especially the elderly. One of the clinical laboratories we visited reported us that prior to celebrations such as Mother's Day, Christmas and other holidays, the demand for these tests has intensified a lot.

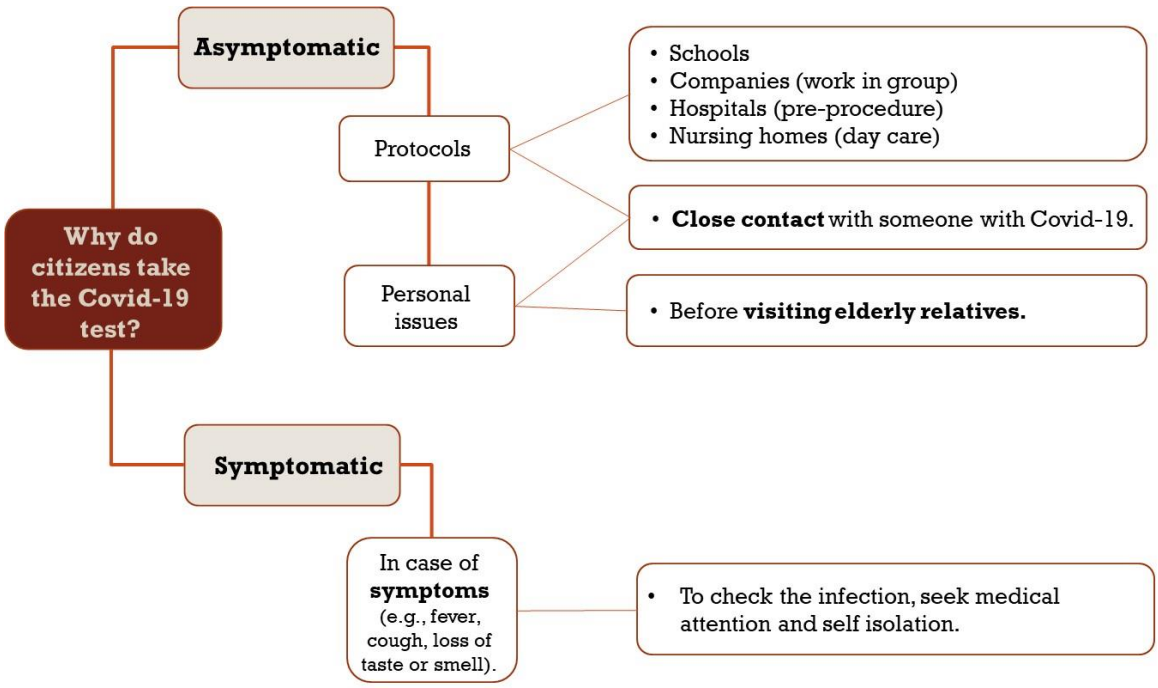

Figure 2: Main reasons for citizens to take a COVID-19 test in SJC (elaborated by the authors).

A **quiz** is mandatory for each test performed, although not all fields are obligatorily answered. The information collected from these quizzes composes most of our dataset. The quiz is filled in by the pharmacy, laboratory, or hospital employee and is based on the patient self-declarations. Therefore, it is not unusual

to have missing information, typos, and mistaken information. Taking, for instance, the postal code it is common to find codes of health unities where the tests were taken instead of the real address of the citizens taking the test. Address and borough information are open fields, so that there are more than six thousand boroughs registered in the raw dataset, while the city has about 412 boroughs/subdivisions only. Other attributes registered are date of birth, gender, symptoms (cough, fever, sore throat, dyspnea, oxygen saturation, respiratory distress, diarrhea, vomit, and others) and comorbidities (chronic cardiovascular disease, diabetes mellitus, chronic respiratory diseases, asthma, puerperal, immunodeciency immunodepression, chronic kidney disease, chronic hematologic disease, Down's syndrome, chronic liver disease, chronic neurological disease, other chronic lung disease, obesity, other risks, high risk pregnancy, and chromosomal diseases).

A private non-profit Research and Planning Institute (IPPLAN - Instituto de Pesquisa e Planejamento) associated to the City Hall of SJC gathers such data, along with other type of government management data in a daily basis. This data includes information from COVID-19 test quizzes along with data from health systems, including the need for hospitalization and reported deaths. With such data, IPPLAN already provides the Health's Secretary some analysis about the profile of contaminated citizens with the aid of dashboards. In this paper we go one step further and also build some predictive models using such data. The raw dataset used here comprises tests performed from March 1^{st}, 2020 to May 14, 2021, totaling 255,815 tested cases. Next we describe how this raw data is pre-processed prior to building our predictive models.

3.2 Methodology step-by-step

Figure 3 illustrates the methodological framework used in this research. We follow by detailing each of the steps in the correspondent subsections. Steps 1 and 2 are initial and comprise the preparation of raw data and computing some descriptive statistics. The following steps 3 to 6 may involve some specific activities depending on the research question addressed.

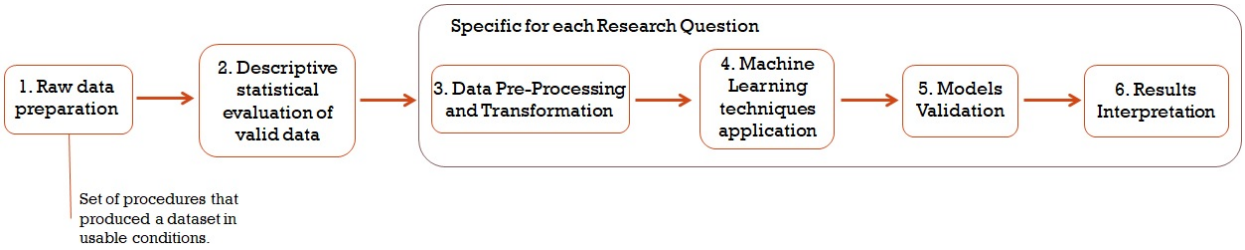

Figure 3: Methodology (elaborated by the authors).

3.2.1 Raw Data Preparation

The raw dataset consists of five basic categories of information: (i) *personal information* (e.g. gender, birth date, address, borough, zip code and on); (ii) *COVID-19 test result* (positive or negative); (iii) *disease evolution information* (e.g. date of first symptoms, hospitalization date, hospitalization place and on); (iv) *symptoms* (e.g. fever, cough, sore throat, vomit and on); and (v) *comorbidities* (e.g. chronic cardiovascular disease, imumnodeciency immunodepression and on).

Step 1 refers to a set of procedures that were performed to clean, adjust, transform, and bring our dataset into an usable condition. Some of the procedures were recommended by IPPLAN professionals. The activities carried out in this phase are listed in Table 10 from the **Appendix** section.

3.2.2 Data Statistical Description

Once the raw data is initially pre-processed, some statistical description procedures summarize important information from our dataset such as the number of women and men tested, positive and negative cases, the age of the people who were tested, which symptoms and comorbidities are the most frequent and the death toll by age group.

3.2.3 Data Pre-Processing and Transformation

In this subsection we describe the pre-processing and transformation activities needed to build the input datasets of the predictive models. All data pre-processing and transformation activities were performed using the R tool.

The first activity required for all research questions was to filter positive cases for the COVID-19 disease only. We also removed any cases in which hospitalization occurred 15 days after the beginning of the first symptoms to ensure the correlation between the events of COVID-19 positive diagnosis and hospitalization. In the case of an event of death, only cases where death occurred up to 30 days after the date of the first symptoms are considered to ensure the correlation between the events of COVID-19 positive diagnosis and death. Cases for which the citizens were already hospitalized when taking the test were disregarded too. Any comorbidities and symptoms reported for less than 1% of the population included in the datasets were also removed, as they are underrepresented in our population. Some attributes were only used to support the computation of other input features and are not employed as input to the predictive models. For instance, date of birth for computing **age**, and hospitalization dates for computing **hospitalization days** that were used for each survey question, respectively.

Next we present some specific pre-processing required for dealing with each of the research questions. For **RQ 1**, the following additional actions were performed:

1. Removing cases with date of hospitalization but no exit date, as we cannot access the hospitalization stay period for them.
2. Removing cases with place of hospitalization but no dates of hospitalization.
3. Adding an attribute referring to the number of days the patient has been hospitalized. Any cases whose value is greater than 100 days of hospitalization or negative were removed, being considered outliers and typos, respectively.
4. Adding a '**severity**' column. Cases representing citizens who died or were hospitalized for 10 days or more were labeled as having a serious condition. Otherwise, they are labeled as a non-serious COVID-19 case.

For **RQ 2**, we performed the following specific additional actions:

1. Removing cases corresponding to citizens that *died* but were not hospitalized.
2. Removing cases with place of hospitalization but no dates of hospitalization.
3. Adding a column to indicate whether the patient was hospitalized or not.

For **RQ 3**, we performed the following specific additional actions:

1. Removing cases without any hospitalization information: no hospitalization date nor date of entry into the ICU.
2. Removing cases which have a date of hospitalization (either ICU and non-ICU beds), but no date of exit nor any information on case evolution.
3. Adding an attribute hosp_days referring to the hospitalization period. The following categories are considered: **short term** stay for a period up to 5 days, **medium term** stay for a period between 6 and 10 days, and **long term** stay for a period of more than 11 days. These values were adopted based on the median hospitalization time (11) beyond to producing three reasonably balanced classes.
4. Removing cases with zero days of hospitalization.

A summary of the characteristics of the datasets built for answering each research question is presented in Table 1. Symptoms and comorbidities are binary attributes, assuming the value 1 when the corresponding symptom/comorbidity is reported and 0 otherwise. Gender is also codified as a binary value. Age is a real-valued attribute. The attributes used in each research question are listed in Table 11 of the **Appendix** section.

Table 1: Summary of the characteristics of the datasets built for each RQ.

Dataset	♯ Observ.	♯ Attrib.	Majority class (%)	Other classes (%)
RQ 1	18,995	18	non-serious (91.0%)	serious (9.0%)
RQ 2	20,011	18	no-hospitalization (79.0%)	hospitalization (21.0%)
RQ 3	2,828	22	short-term hospitalization (37.6%)	medium-term hospit. (28.7%) long-term hospit. (33.7%)

3.2.4 Machine Learning Prediction Models

Each of the datasets obtained previously represent classification problems, in which one wants to predict a qualitative label for a new observation. There are several ML classification techniques in the literature (12). Here we have chosen some representatives which commonly show a highlighted predictive performance for structured data like ours. They are:

- **Support Vector Machine (SVM)**: it is a predictive method based on the statistical learning theory and successfully applied to several pattern recognition tasks (13). SVM builds an hyperplane in a high dimensional feature space which separates the different classes with a maximum margin;
- **Gradient Boosting Classifiers (GB)**: it is an ensemble technique which combines multiple weak learning models into a more robust predictor. At each round of the technique, any observations wrongly classified in previous rounds receive a larger weight when building the new classification model. At the end, all classification models built are joined by a weighted majority voting strategy in order to give the final predictions (14).
- **RandomForest (RF)**: this is also an ensemble technique combining multiple decision tree models. The trees are build from bootstrapping samples of the data and a set of randomly chosen features. Next, their predictions are joined by a simple majority voting scheme (15).

We limited our study to the three previous algorithms, although other classification techniques can be explored in the future. For instance, Decision Tree models can generate decision rules supporting the health professionals, but their predictive performance was much lower than those of the previous algorithms in our problems and was disregarded.

3.2.5 Model Validation

All techniques had their performance evaluated by a 10-fold cross-validation strategy. Within it, the datasets are first divided into ten folds of approximately the same size. Nine folds are used for training a predictive model, while the remaining fold is left-out for testing, simulating the presentation of new data to the model. This is done ten times, alternating the fold left out for testing.

Since the datasets built for RQ 1 and RQ 2 are highly imbalanced, we have also randomly undersampled the majority class in the training sets to have the same number of observations as the minority class. This is required because when faced to imbalanced datasets the ML techniques tend to favor the majority class in detriment of the minority class (16), which is frequently the class of most interest. Since the choice of the observations to keep from the majority class is made at random, we repeat this procedure ten times. Therefore, for research questions RQ 1 and RQ 2 a total of 100 tests are performed and their results are averaged. Furthermore, the repetitive *random sampling-training-testing* process contributed to minimize any bias. Since the dataset from RQ 3 is quite balanced, the random undersampling step was not needed and the results are averaged across the 10 cross-validation test sets.

For all datasets and ML algorithms, a grid-search of the main hyperparameter values was performed considering training data only, with an inner 3-fold cross-validation. This is an important step as the best hyperparameter values per dataset are previously unknown and are decisive on classification performance. Thus, grid-search was used to find the combinations of hyperparameters values of the ML models, from a specified set of values, resulting in the most 'accurate' predictions (17; 18).

For all prediction models, the average AUC (Area Under the ROC Curve) in the test sets is reported, along with the per class accuracy (for binary classification problems, they correspond to sensibility and specificity measures). AUC varies between 0 and 1 and values closer to 1 are better, while values around 0.5 correspond to random predictions. Sensibility and specificity vary between 0 and 1 and the higher the value, the better the performance.

In addition to the predictive results obtained, we used SHAP (acronym for SHapley Additive exPlanations) values to interpret the decisions made by the prediction models. SHAP assigns each feature an importance

value for a particular prediction (19), allowing to prospect which input attributes are more decisive for obtaining the given outcomes. For building such plots, age must be previously normalized.

4 Results & Discussion

In this section, we present and discuss our results. The dataset covered test information for the period from March 1^{st}, 2020 (earliest date of first symptoms) to May 14, 2021. Before the raw data preparation, there were information from 255,815 tests recorded and 44 attributes describing each of the observations. After preparing the raw data, 43,579 observations and 36 attributes were kept, corresponding to 17% of the original number of tests recorded.

Regarding gender, there are 19,484 (44.7%) male and 24,095 (55.3%) female individuals in the dataset. Among the test results, 22,535 (51.7%) were negative and 21,044 (48.3%) were positive. This can be related to the fact that many tests are done for protocol reasons and not because people actually have any suspicion on the disease.

Table 2 presents the frequency of symptoms reported in our filtered dataset. Cough, fever, sore throat and dyspnea were the most frequent symptoms, quite similar to common colds, which may have led people to get tested. "Other symptoms" are also quite frequent. This indicates the form to be filled by the tested individuals is incomplete. For instance, although frequently associated to the COVID-19 disease, the loss of smell and taste is absent from the dataset.

Table 2: Occurrence of reported symptoms (for the filtered dataset).

Symptoms	**Number of patients**	**Percentage**
cough	33,937	77.9%
fever	25,675	58.9%
sore throat	23,294	53.5%
dyspnea	16,878	38.7%
oxygen saturation	4,608	10.6%
respiratory distress	4,092	9.2%
diarrhea	949	2.2%
vomit	585	1.3%
other symptoms	30,840	70.8%

Table 3 presents the comorbidities in the tested population (only for the filtered dataset). The main comorbidities reported are related to chronic cardiovascular disease, diabetes and chronic respiratory diseases. Some comorbidities have a very low frequency in the population tested, but again the "other risks" group is quite populated. They represent commorbidities not discriminated in the form's options.

Table 3: Occurrence of comorbidities (for the filtered dataset).

Comorbidities	**Number of patients**	**Percentage**
chronic cardiovascular disease	7,919	18.2%
diabetes mellitus	5,252	12.1%
chronic respiratory diseases	2,481	5.7%
immunodeficiency immunodepression	666	1.5%
chronic kidney disease	403	0.9%
chronic hematologic disease	71	0.2%
Down's syndrome	17	0.0%
chronic liver disease	53	0.1%
asthma	363	0.8%
puerperal	13	0.0%
chronic neurological disease	378	0.9%
other chronic lung disease	341	0.8%
obesity	416	1.0%
other risks	2,692	6.2%
high risk pregnancy	367	0.8%
chromosomal diseases	260	0.6%

Figure 4 shows the histogram of ages (in years) of the filtered dataset. People of all ages were tested,

including children. However, most of them are working-age adults with a medium age around 40 years.

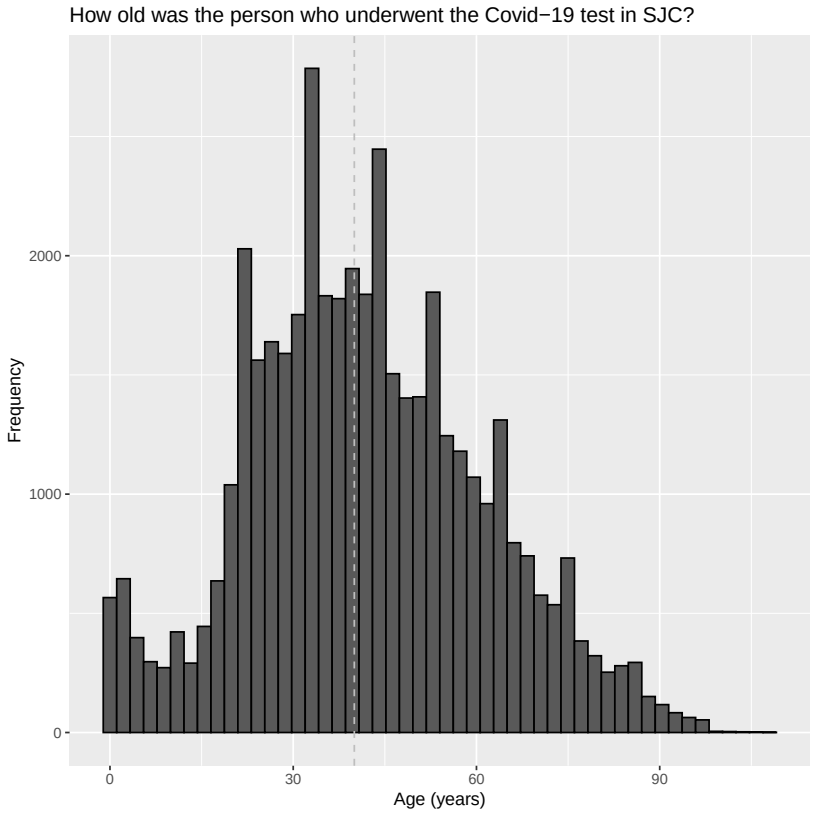

Figure 4: Histogram of ages of tested individuals in the SJC dataset.

In our data excerpt, 1,677 citizens died from COVID-19, being 967 (58%) male and 710 (42%) female individuals. The mortality rate of COVID-19 being higher for male individuals compared to females is confirmed by other studies (20). The work (21) also points as most important factors to the death toll as those related to immune system response and the role of sex hormones (20). The age of people who died of COVID-19 in our sample dataset is represented in the histogram from Figure 5. We highlight people over 50 years as those who were the most affected by the death toll, which is also in accordance to the literature on the disease, that affects more severely elderly people (22).

4.1 Results for Research Question 1

As presented in the previous section, the dataset related to RQ 1 has **18,995** lines (cases) and 18 columns (attributes). Among the cases, **17,381** (91%) were **non-serious** and **1,710** (9%) were **serious** cases. Table 4 presents the average (± standard deviation) predictive performance of the cross-validated models built using such data. Best results per metric are highlighted in boldface. We can observe that all measures present high values and a good predictive performance can be observed when answering RQ 1, with very similar results for all ML techniques. In general, specificity was higher than sensibility and the models are more accurate in identifying non-severe cases. GB had a slight superior sensibility and AUC compared to the other techniques. And SVM showed the highest specificity.

Table 4: Predictive results for RQ 1.

Technique	**Specificity**	**Sensibility**	**AUC**
RF	0.917 ± 0.008	**0.909 ± 0.018**	**0.954 ± 0.007**
GB	0.917 ± 0.008	0.906 ± 0.014	**0.954 ± 0.007**
SVM	**0.919 ± 0.007**	0.905 ± 0.020	0.944 ± 0.010

Figure 6 shows a summary plot of SHAP values for RQ 1. This plot was generated using all data from RQ 1 and the *RF* algorithm. Each point is an observation from the dataset. The x-axis distance shows how

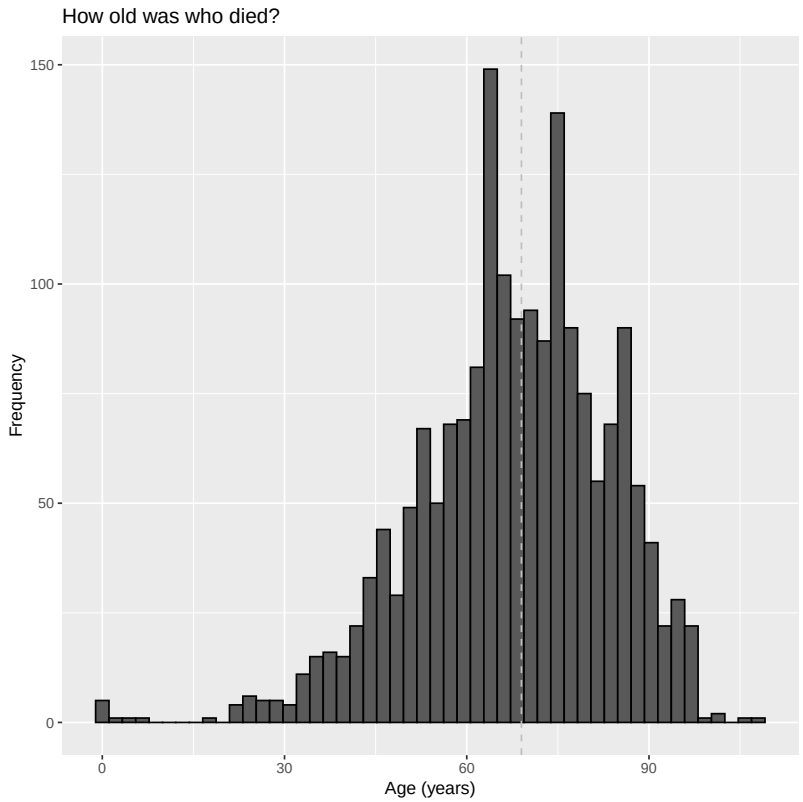

Figure 5: Histogram of the ages of the dead by COVID-19 in SJC in our dataset.

much that variable influenced so that the model considered the instance serious (if in the right quadrant) or non-serious (if in the left quadrant). Colours represent high (in pink) or low (in blue) values of the variables. In case of binary variables, pink dots represent **presence** of a symptom or a comorbidity and blue dots represent **absence**. In the case of age, pink dots correspond to higher ages, blue dots to lower ages, and purple dots are intermediate ages. The variables are ordered according to their importance in the model performance. Therefore, in this model, oxygen saturation is the most important variable, with the 1 value attributed to patients with low saturation and 0 otherwise. Except for age, all other variables are binary, making interpretation easier. For example, whenever the patient has low oxygen saturation or respiratory distress, this pushes the output of the model to as serious. The same happens for comorbidities such as chronic cardiovascular disease, diabetes and obesity. The opposite happens to sore throat, whose presence pushes the outcome of the model to as a non-serious case. The male sex (blue color) is also more indicative of a serious case than the female counterpart (in pink). Elderly people also tend to have predictions towards the serious class. Nonetheless, one must observe that, for each instance the model output is always the result of the sum of the influence of all variables and none of them alone leads to a higher predictive accuracy than their combination. The presence of the "other risks" variable among the top-ranked evidences there are other commorbidities impacting the predictions towards the severity class, although they are not formally discriminated in our data excerpt.

4.2 Results for Research Question 2

Dataset from RQ 2 has **20,011** lines (cases) and 18 columns (attributes). Among the observations, **15,965** (79%) did **not need** to be hospitalized and **4,196** (21%) **needed** to be hospitalized. We report in Table 5 the average and standard deviation of the AUC, sensibility and specificity of the predictive models built.

Table 5: Predictive results for RQ 2.

Technique	**Specificity**	**Sensibility**	**AUC**
RF	0.992 ± 0.001	0.876 ± 0.013	0.969 ± 0.004
GB	0.985 ± 0.004	**0.883 ± 0.011**	**0.970 ± 0.004**
SVM	**0.995 ± 0.001**	0.872 ± 0.014	0.944 ± 0.012

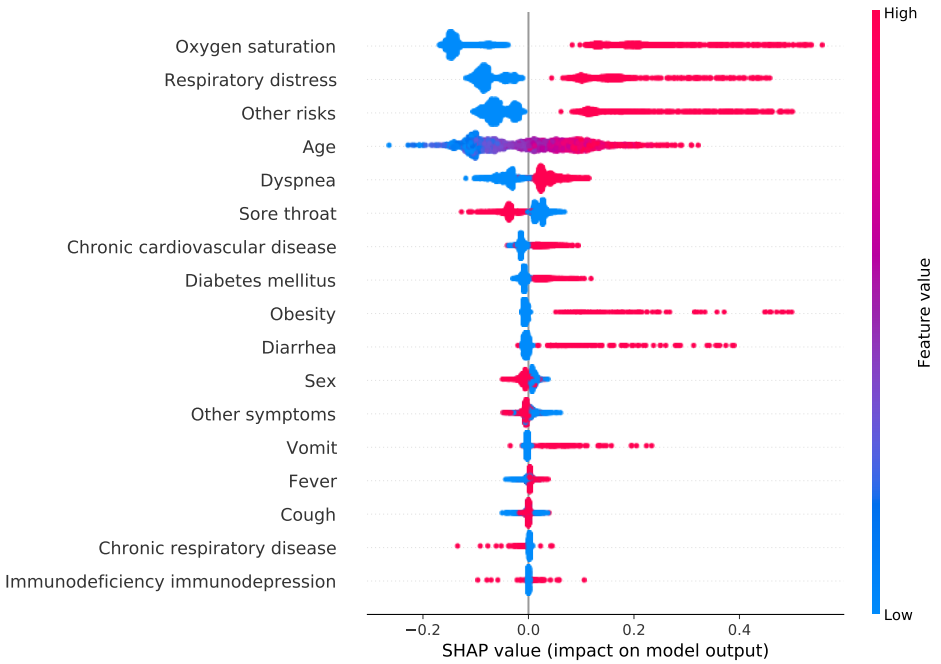

Figure 6: SHAP values associated to RQ 1.

Again all models have a high predictive ability, for both hospitalization and non-hospitalization classes. Here specificity and sensibility values are always above 90%. SVM again showed the highest specificity values, although all models performed similarly. RF models had highlighted AUC and sensibility performance. The predictive results achieved for all models indicate they can be accurately employed to support a better decision making regarding the allocation of hospital resources.

Figure 7 shows SHAP values related to **RQ 2**. This plot was also generated using all data assembled for RQ 2 and the *RF* algorithm. Similar considerations drawn for RQ 1 can be made here. The attributes that most influenced the need for hospitalization refer to symptoms: low oxygen saturation and respiratory distress. Other risks, related to comorbidities, also play an important role in determining the risk of hospitalization and evidence the lack of important specific comorbidity data. Next dyspnea pushes the predictions towards the need for hospitalization, followed by age with a predominance of elderly people. And sore throat is again not a determinant symptom for hospitalization. Sex does not influence the predictive results so clearly as in RQ 1. Nonetheless, again one must point that the model output is always the result of the combined influence of all variables.

4.3 Results for Research Question 3

The dataset associated to RQ 3 has **2,828** lines (cases) and 22 columns (attributes). Among the cases, **1,062** (37.6%) stayed hospitalized for a **short term** (up to 5 days), **812** (28.7%) stayed hospitalized for a **medium term** (6 to 10 days) and **954** (33.7%) stayed hospitalized for a **long term** (more than 11 days). Table 6 shows some statistics related to the number of stay days in our dataset. And Figure 8 presents the histogram of hospitalization days. Most of the hospitalizations are of short term and very few hospitalizations exceed 40 days and can be regarded as outliers. The dashed line in Figure 8 represents the median hospitalization days, which is inferior to a week (6 days, according to Table 6).

Table 6: Number of hospitalization days (ICU and non-ICU).

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.00	3.00	6.00	9.53	12.00	98.00

Unlike the previous models, here the predictive results achieved by all ML models were low and not satisfactory, with AUC values slightly above the 0.5 baseline. Table 7 presents the average and standard deviation of the AUC and accuracy per class results. There may be some obstacles for a ML model to ascertain the period for which a patient will be hospitalized. The results are particularly deceptive for the medium-term stay class. We also tried to develop regression models to predict the actual number of hospitalization days, but the predictive results attained remained low.

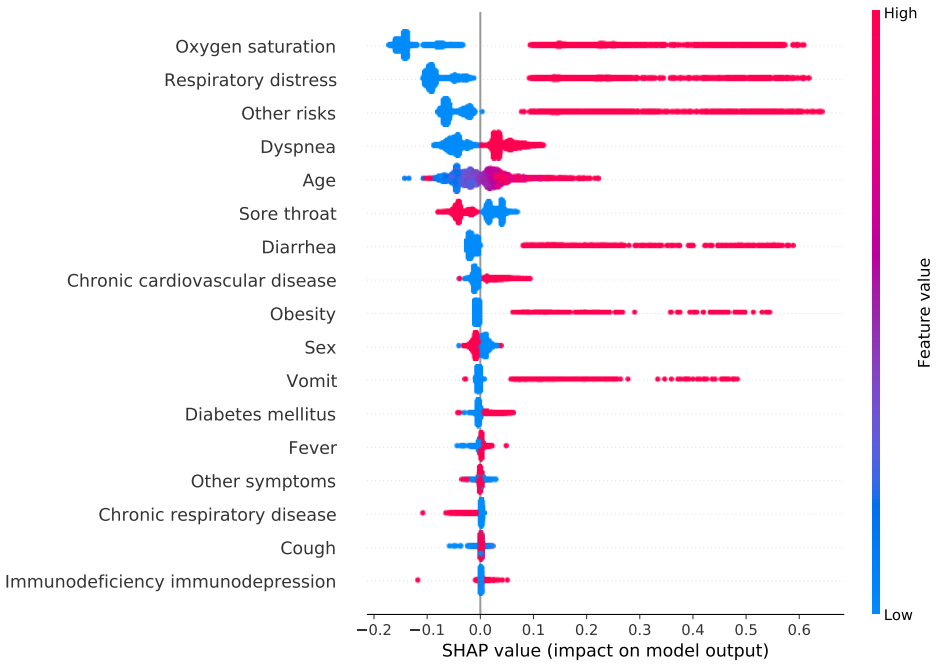

Figure 7: SHAP values associated to RQ 2 using the RF model.

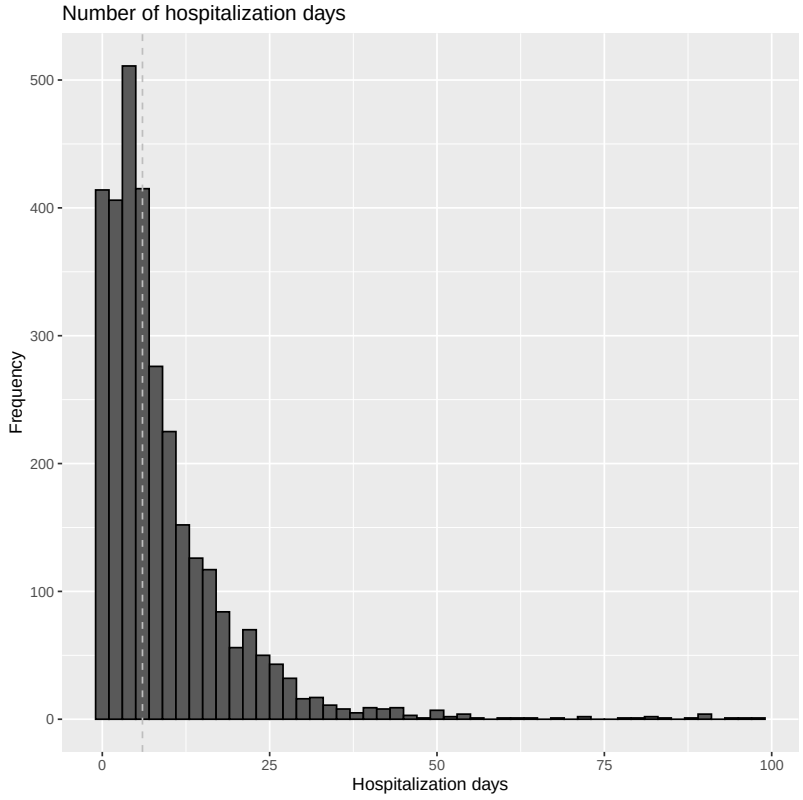

Figure 8: Histogram of hospitalization days.

We can think of two reasons for the difficulty of the prediction of hospitalization stay based on initial symptoms and comorbidities of the patients. The first may be a reason of administrative origin in which the period of hospitalization is influenced not only by the patient's conditions, but by the number of available beds and the waiting list of patients with even more serious conditions. The second reason can be linked to the recovery time within the treatment offered to the patient. In this case, each person has a particular response which may be not linked to the attributes under consideration. These assumptions were certified

by a medical expert (one of the authors). Therefore, other factors than simply age, sex, initial symptoms and comorbidities are required for tracking whether the patient will require a short, medium or long term hospitalization. This may include blood exams and other monitoring results during hospitalization.

Table 7: Predictive results for RQ 3.

Technique	Accuracy Short	Accuracy Medium	Accuracy Long	AUC
RF	**0.572 ± 0.065**	0.066 ± 0.051	**0.530 ± 0.042**	**0.573 ± 0.032**
GB	0.494 ± 0.070	**0.179 ± 0.094**	0.469 ± 0.098	0.559 ± 0.045
SVM	0.541 ± 0.073	0.070 ± 0.059	0.519 ± 0.041	0.554 ± 0.024

Finally, Table 8 presents a confusion matrix for one particular run of the RF model. We can notice that class medium term is confused as either a long term or a short term hospital stay. In fact, the boundary between short-medium and medium-long term stays is close and may be ambiguous. Very few predictions are obtained for the medium term stay class, so that long and short term stays tend to be more confused between each other too. Therefore, a third source of difficulty of the classification problem associated to RQ 3 is that the boundaries between the considered classes are close or even overlap given the input attributes considered, which are not discriminative enough for this problem.

Table 8: Confusion Matrix of the RF model for one of the cross-validation rounds of RQ 3.

		Actual		
		Long	**Medium**	**Short**
Predicted	**Long**	506	366	396
	Medium	76	54	62
	Short	372	392	607

5 Conclusions & Future Work

During the COVID-19 pandemic, which is still of world concern, many efforts have been made by the scientific community to better understand the situation and provide insights for a better decision-making. The focus of our research was particularly on health resources management. By experimenting different classification ML techniques for predicting the propensity of severe cases and of hospitalization using data regularly collected when COVID-19 tests are carried out in São José dos Campos - São Paulo, in Brazil, we were able to provide some valuable tools for supporting decision making policies by health professionals and managers. The predictive results are very good for both the severity and hospitalization prediction problems. But one cannot predict the hospital stay period using the available data, a problem which may require other attributes as input. Experimentally, the high accuracy verified in RQ 1 and RQ 2 was measured by the AUC (Area Under the ROC Curve) measure using the available data, with a high sensibility and specificity. In contrast, the models built for RQ 3 present a low AUC accuracy and RQ 3 cannot be answered properly using the available data.

In addition, for the severity and hospitalization prediction problems we also used SHAP values to identify the attributes that were influencing the results of the Random Forest predictor the most, resulting in inherently interpretable results (23). Symptoms as low oxygen saturation and respiratory distress have significant influence in determining the severity of the cases and the need for hospitalization. For both problems, other risks related to comorbidities are relevant variables, among other health conditions evaluated. Elderly and male individuals also push the predictions towards severity and need for hospitalization. All observations are in accordance to the literature on the COVID-19 disease. But, interestingly, we have proven that some common information routinely collected when a SARS-COV-2 diagnosis test is undertaken can be used to build accurate predictive models to determine if a new case will require hospitalization or develop a severe condition. This type of information is available along all Brazilian territory, so that other cities can perform similar analysis using data about their citizens.

The ML models proposed in this research focused on estimating the resources needed in hospitals to accommodate COVID-19 patients. In case the models show high false positive rates (low specificity), the need for resources will be overestimated and the managers will likely allocate resources unnecessarily, pressing investments that could be postponed. While in cases of many false negatives (low sensibility), there could be not enough resources allocated, causing overload and chaos in the health system, which had been frequently reported at onset of the pandemic. Therefore, trade-offs of high sensibility and specificity performances should be sought.

But it is also important to point out that frequently the population do not fill properly forms related to their health conditions. Our work has evidenced how such information can support a better management of hospital resources, which is of general interest of the population and public authorities and can encourage a more conscious filling of medical records and forms. The forms may also need some adjustments to better reflect the main symptoms and comorbidities of the population, as the variable other risks greatly influenced the predictions in our analysis.

As future work, we plan to test our models in practice in partnership with the city's health secretariat. We also want to evaluate the benefits of including more input attributes to the models, such as the boroughs of the reported cases. In addition to being cut by one of the main Brazilian highways (BR-116 that connects São Paulo to Rio de Janeiro), SJC city has a network of express avenues that connect different blocks of boroughs. This means that the population is very spread out and we can assume that certain boroughs are more affected by COVID-19 cases than others. Another approach to be tested is the evaluation of data from more recent periods in which vaccination campaigns have been intensively carried on. Finally, it will be worth to investigate how the predictive models behave face to cases of new disease variants, such as delta.

Acknowledgment

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001. M. G. Valeriano holds a CAPES scholarship from the project "Data Science fighting outbreaks, epidemics and pandemics in hospitals" (grant 88887.507037/2020-00). We also thank IPPLAN and the health secretariat of SJC for providing the raw database for our project.

References

[1] IBGE, "INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA," *Cidades*. [Online]. Available: https://cidades.ibge.gov.br/brasil/sp/sao-jose-dos-campos/panorama

[2] M. A. L. Nicolelis, R. L. G. Raimundo, P. S. Peixoto, and C. S. Andreazzi, "The impact of super spreader cities, highways, and intensive care availability in the early stages of the COVID-19 epidemic in Brazil," *Scientific Reports*, vol. 11, no. 13001, pp. 1–12, 2021. [Online]. Available: https://doi.org/10.1038/s41598-021-92263-3

[3] M. F. Ziegler. (2021, August) Variante Gama (P.1) é mais agressiva, mas pode ser contida com vacina e lockdown, comprova estudo. [Online]. Available: https://agencia.fapesp.br/variante-gama-p1-e-mais-agressiva-mas-pode-ser-contida-com-vacina-e-lockdown-comprova-estudo/36531/

[4] E. Vivier, C. Pariset, S. Rio, S. Armand, F. Doroszewski, D. Richard, M. Chardon, G. Romero, P. Metral, M. Pecquet, and A. Didelot, "Specific exposure of ICU staff to SARS-CoV-2 seropositivity: a wide seroprevalence study in a French city-center hospital," *Annals of Intensive Care*, vol. 11, no. 1, December 2021. [Online]. Available: https://doi.org/10.1186/s13613-021-00868-8

[5] A. N. Masai, B. Güçiz-Dogan, P. N. Ouma, I. N. Nyadera, and V. K. Ruto, "Healthcare services utilization among international students in Ankara, Turkey: a cross-sectional study," *BMC Health Services Research*, vol. 21, no. 311, 2021. [Online]. Available: https://doi.org/10.1186/s12913-021-06301-x

[6] J. W. Tankel, D. Ratcliffe, M. Smith, A. Mullarkey, J. Pover, Z. Marsden, P. Bennett, and D. Green, "Consequences of the emergency response to COVID-19: a whole health care system review in a single city in the United Kingdom," *BMC Emergency Medicine*, vol. 21, no. 55, 2021. [Online]. Available: https://doi.org/10.1186/s12873-021-00450-2

[7] M.-K. Breyer, R. Breyer-Kohansal, S. Hartl, M. Kundi, L. Weseslindtner, K. Stiasny, E. Puchhammer-Stöckl, A. Schrott, M. Födinger, M. F. Michael Binder, E. F. M. Wouters, and O. C. Burghuber, "Low SARS-CoV-2 seroprevalence in the Austrian capital after an early governmental lockdown," *Scientific Reports*, vol. 11, no. 10158, December 2021. [Online]. Available: https://doi.org/10.1038/s41598-021-89711-5

[8] N. Gozzi, M. Tizzoni, M. Chinazzi, L. Ferres, A. Vespignani, and N. Perra, "Estimating the effect of social inequalities on the mitigation of COVID-19 across communities in Santiago de Chile," *Nature Communications*, vol. 12, no. 2429, pp. 1–9, 2021. [Online]. Available: https://doi.org/10.1038/s41467-021-22601-6

[9] V. V. L. Albani, R. M. Velho, and J. P. Zubelli, "Estimating, monitoring, and forecasting COVID-19 epidemics: a spatiotemporal approach applied to NYC data," *Scientific Reports*, vol. 110, no. 9089, 2021. [Online]. Available: https://doi.org/10.1038/s41598-021-88281-w

[10] CDC, "Serology testing for COVID-19 at CDC," *Centers of Disease Control and Prevention*, November 2020. [Online]. Available: https://www.cdc.gov/coronavirus/2019-ncov/lab/serology-testing.html

[11] Y. Wu, B. Hou, J. Liu, Y. Chen, and P. Zhong, "Risk factors associated with long-term hospitalization in patients with COVID-19: A single-centered, retrospective study," *Frontiers in Medicine*, vol. 7, no. 315, pp. 1–10, 2020. [Online]. Available: https://doi.org/10.3389/fmed.2020.00315

[12] F. Osisanwo, J. Akinsola, O. Awodele, J. Hinmikaiye, O. Olakanmi, and J. Akinjobi, "Supervised machine learning algorithms: classification and comparison," *International Journal of Computer Trends and Technology (IJCTT)*, vol. 48, no. 3, pp. 128–138, 2017. [Online]. Available: https://doi.org/10.14445/22312803/ijctt-v48p126

[13] A. C. Lorena and A. C. de Carvalho, "Uma introdução às Support Vector Machines," *Revista de Informática Teórica e Aplicada*, vol. XIV, no. 2, pp. 43–67, 2007. [Online]. Available: https://www.seer.ufrgs.br/rita/article/viewFile/5690/3543

[14] D. Nelson. (2019, aug) Gradient boosting classifiers in python with scikit-learn. [Online]. Available: https://stackabuse.com/gradient-boosting-classifiers-in-python-with-scikit-learn

[15] T. Yiu. (2019, June) Understanding Random Forest. [Online]. Available: https://towardsdatascience.com/understanding-random-forest-58381e0602d2

[16] A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk, and F. Herrera, *Learning from imbalanced data sets*. Springer, 2018.

[17] R. Joseph. (2018, december) Grid search for model tuning. [Online]. Available: https://towardsdatascience.com/grid-search-for-model-tuning-3319b259367e

[18] A. B. Jiménez, J. L. Lázaro, and J. R. Dorronsoro, "Finding optimal model parameters by Discrete Grid Search," *Innovations in Hybrid Intelligent Systems. Advances in Soft Computing*, vol. 44, pp. 120–127, 2007. [Online]. Available: https://doi.org/10.1007/978-3-540-74972-1_17

[19] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA, 2017. [Online]. Available: https://arxiv.org/abs/1705.07874

[20] A. Capuano, F. Rossi, and G. Paolisso, "COVID-19 kills more men than women: An overview of possible reasons," *Frontiers in Cardiovascular Medicine*, vol. 7, no. 131, pp. 1–7, July 2020. [Online]. Available: https://doi.org/10.3389/fcvm.2020.00131

[21] J.-M. Jin, P. Bai, W. He, F. Wu, X.-F. Liu, D.-M. Han, S. Liu, and J.-K. Yang, "Gender differences in patients with COVID-19: Focus on severity and mortality," *Frontiers in Public Health*, vol. 8, no. 152, pp. 1–6, April 2020. [Online]. Available: https://doi.org/10.3389/fpubh.2020.00152

[22] D. Cortis, "On determining the age distribution of COVID-19 pandemic," *Frontiers in public health*, vol. 8, no. 202, May 2020. [Online]. Available: https://doi.org/10.3389/fpubh.2020.00202

[23] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, May 2019. [Online]. Available: https://doi.org/10.1038/s42256-019-0048-x

Appendix

Table 9 organizes a summary of the studies herein referenced, which analyzed COVID data at a city level. Table 10 lists the pre-processing initially applied to clean up the raw data provided by IPPLAN. And Table 11 lists the attributes used as input in RQ1, RQ2 and RQ3.

Table 9: Summary of COVID-19 studies at a city level.

Ref.	City	Objective	Study Design	Some conclusions
(7)	Vienna (Austria)	To verify SARS-CoV-2 seroprevalence in Vienna in order to evaluate first government lockdown (from March to May 2020) in general population aged between 6-85 years.	Survey questionnaire and SARS-CoV-2 antibody testing of over 12,000 participants (including household members). And statistical analysis of positive cases.	Lockdown is effective. There is a significant underreporting of infection due to asymptomatic cases. Disturbance of smell and taste are the most prominent symptoms of Covid infection. There is a high probability of transmission within household contacts. Children from 6 to 9 years old have no lower seroprevalence than adults.
(4)	Lyon (France)	To verify SARS-CoV-2 seroprevalence in the staff of a hospital in Lyon, several weeks after the first epidemic wave (March and April of 2020).	Survey questionnaire and SARS-CoV-2 antibody testing offered to all persons who had worked in a certain hospital from March 1 to April 30, 2020. And statistical analysis of positive cases.	Seroprevalence is higher among staff than in the general population. The seropositivity rates are particularly high for staff in contact with COVID-19 patients, especially those in the emergency department and in the COVID-19 unit but were much lower among ICU staff.
(8)	Santiago (Chile)	To model the spatial and temporal spread of SARS-CoV-2 in Santiago by estimating the impact of nonpharmaceutical interventions (NPIs) on Covid spreading and estimate the effect of social inequalities.	Anonymized mobile phone data from 1.4 million users, 22% of the population in the area. And a stochastic mechanistic epidemic model integrating mobility, physical contacts, and census data in three phases: (i) before any restrictions; (ii) during partial lockdown, and (iii) during full lockdown.	Partial lockdown reduced mobility by 48%. Full lockdown reduced mobility by 65%. A greater decrease in mobility is generally associated with a higher HDI.

Continued on next page

Table 9 – *Continued from previous page*

Ref.	City	Objective	Study Design	Some conclusions
(9)	New York (USA)	To simulate and monitor COVID-19 epidemic evolution considering different levels of disease severity, its impact on age ranges, and the distribution of the population in different boroughs.	To provide a SEIR (Susceptible-Exposed-Infected-Recovered) type dynamic model calibrated with real data updated from time to time from the NYC Health website for COVID-19.	The model had good results in predicting: (i) the average of new daily infections; (ii) hospitalizations and deaths by boroughs; and (iii) changes in the pattern of disease transmission as measures are taken.
(5)	Ankara (Turkey)	To investigate healthcare access and behavior in a group of international students (higher education) and identify potential barriers that affect healthcare utilization.	Survey questionnaire of 535 international students from 83 countries in Sep-Oct/2020. Statistical analysis was used to evaluate the relationships between variables related to access to health services.	The main barriers affecting healthcare access by international students are related to: (i) lack of awareness of healthcare support systems; (ii) perceived stigma associated with mental health services; and, (iii) language barriers. However, three quarters of the respondents reported that if they suspected to have infections of public health concern, such as COVID-19, they would call an ambulance.
(6)	Salford (United Kingdom)	To analyze the care and treatment offered to Salford residents whose death was registered in Salford due to COVID-19 in the "first wave" of the pandemic (from March 20 to May 18, 2020). This was a period of excess of deaths in the city. And to evaluate some way to mitigate the excess of deaths in similar situations in the future.	Review of the pathway of care of patients whose death was registered in Salford between Mar 1 and May 18 (first wave): primary care, secondary care, and 111 and 999 calls. An expert panel judged avoidability of death according to a scale from *definitely avoidable* to *definitely not avoidable*.	There were 522 deaths, with a mean age of 79 ($\pm$ 9) years. 64% of the cases had cardiovascular comorbidity. Regarding avoidability of death: 80% had Score 6 (definitely not avoidable); 18% had score consistent with some degree of avoidability; 15% had score of 5 or 4 (slight or possible avoidability); 3% had score of 3 or 2 (likelihood of being avoidable); None scored 1 (definitely avoidable); 2% had no final score allocated as these were sudden deaths in a patient's home with no further information available.

Table 10: Pre-processing activities applied to raw data.

	Item	**Pre-processing Activity**
1	Blank cells	Filled in with: 'NA' (not applicable).
2	All attributes referring to dates	Adjusted to DATE format (instead of TEXT format), allowing operations with dates.
3	Attribute '**date_first_symptoms**' (refers to the date of the first symptoms)	Excluded lines prior to 2020-mar-01 (probably typo, as the disease started to spread in Brazil in march 2020).
4	Attribute '**evolution**' (refers to the disease evolution)	Excluded lines in which this attribute has been filled in as 'ignored' (according to IPPLAN recommendation).
5	Attribute '**date_notification**' (refers to the notification date)	Excluded attribute (according to IPPLAN, this data is not reliable, as the results of the tests can take a long time to be recorded).
6	Attribute '**HASH**' (refers to anonimized personal data)	Excluded attribute. Adopted '**id_notification**' (refers to the test identification) as an identifier.
7	Attribute '**classification**' (refers to healthcare professional or not)	Excluded attribute (note: in most cases, it was not filled out.)
8	Attributes: '**place_admission**', '**active_transmission**', '**monitored_transmition**'	Excluded attributes (they were not evaluated.)
9	Attributes: '**address**', '**borough**', '**zip code**'	Excluded attributes (note: due to typos and lack of standardization, there are more than 6,000 different records, but the city has just 412 boroughs/subdivisions.)
10	Attribute '**id_notification**' (refers to the identification of the notification)	In case of identical ID, just the most recent was kept. There must be only one exam per ID.
11	Attributes related to symptoms ('**fever**' | '**cough**' | '**sore throat**' | '**dyspnea**' | '**respiratory discomfort**' | '**saturation**' | '**dyarrea**' | '**vomit**' | '**outher symptoms**')	Maintained lines with at least 3 filled cells. A heuristic was applied to identify these cases (note: cells can be 'blank' meaning 'no', lack of information or typo.)
12	Attributes related to comorbidities ('**chronic cardiovascular disease**' | '**imumnodeficiency immunodepression**' | '**chronic renal disease**' | '**diabetes mellitus**' | '**chronic hematological disease**' | '**down syndrome**' | '**chronic hepatic disease**' | '**asthma**' | '**puerperal**' | '**chronic neurological disease**' | '**other chronic lung disease** ' | '**obesity**' | '**outher risks**' | '**chronic respiratory diseases**' | '**high risk pregnancy**' | '**cromossomic diseases**')	
13	Attribute '**result_exam**' (refers to the test result)	Excluded lines where the exam result has status of awaiting result: [**Aguardando resultado**].
14	Add a column with the attribute '**age**'	Removed lines whose age is **greater than 110 years** (probably a typo in one of the dates).
15	Attribute '**sex**' (refers to the patient's gender)	Became numerical: Female=1; Male=0.
16	Attribute '**result_exam**'	Became numerical: Positive=1; Negative=0.
17	Attributes related to symptoms and comorbidities	Became numerical: Filled =1 (yes); Blank=0 (no).

Table 11: Attributes used to RQ1, RQ2 and RQ3.

	Personnal	**Symptons**	**Comorbities**	**New**
RQ1	Gender Age	Fever Cough Sore throat Dyspnea Respiratory distress Oxygen saturation Diarrhea Vomit Other symptoms	Chronic cardiovascular disease Immunodeficiency immunodepression Diabetes mellitus Obesity Other risks Chronic respiratory disease	Severity
RQ2	Gender Age	Fever Cough Sore throat Dyspnea Respiratory distress Oxygen saturation Diarrhea Vomit Other symptoms	Chronic cardiovascular disease Immunodeficiency immunodepression Diabetes mellitus Obesity Other risks Chronic respiratory disease	Hospitalization
RQ3	Gender Age	Fever Cough Sore throat Dyspnea Respiratory distress Oxygen saturation Diarrhea Vomit Other symptoms	Chronic cardiovascular disease Immunodeficiency immunodepression Chronic kidney disease Diabetes mellitus Asthma Chronic neurological disease Other chronic pneumopathy Obesity Other risks Chronic respiratory disease	Hospitalization Days