Dataset: <https://www.kaggle.com/einsteindata4u/covid19>

### Background

The World Health Organization (WHO) characterized the COVID-19, caused by the SARS-CoV-2, as a pandemic on March 11, while the exponential increase in the number of cases was risking to overwhelm health systems around the world with a demand for ICU beds far above the existing capacity, with regions of Italy being prominent examples.

Brazil recorded the first case of SARS-CoV-2 on February 26, and the virus transmission evolved from imported cases only, to local and finally community transmission very rapidly, with the federal government declaring nationwide community transmission on March 20.

**Until March 27, the state of São Paulo had recorded 1,223 confirmed cases of COVID-19, with 68 related deaths, while the county of São Paulo, with a population of approximately 12 million people and where Hospital Israelita Albert Einstein is located, had 477 confirmed cases and 30 associated death, as of March 23.** Both the state and the county of São Paulo decided to establish quarantine and social distancing measures, that will be enforced at least until early April, in an effort to slow the virus spread.

One of the motivations for this challenge is the fact that in the context of an overwhelmed health system with the possible limitation to perform tests for the detection of SARS-CoV-2, testing every case would be impractical and tests results could be delayed even if only a target subpopulation would be tested.

### Dataset

**This dataset contains anonymized data from patients seen at the Hospital Israelita Albert Einstein, at São Paulo, Brazil, and who had samples collected to perform the SARS-CoV-2 RT-PCR and additional laboratory tests during a visit to the hospital.**

**All data were anonymized following the best international practices and recommendations. All clinical data were standardized to have a mean of zero and a unit standard deviation.**

### Task Details

TASK 1

• Predict confirmed COVID-19 cases among suspected cases.

Based on the results of laboratory tests commonly collected for a suspected COVID-19 case during a visit to the emergency room, would it be possible to predict the test result for SARS-Cov-2 (positive/negative)?

TASK 2

• Predict admission to general ward, semi-intensive unit or intensive care unit among confirmed COVID-19 cases.

Based on the results of laboratory tests commonly collected among confirmed COVID-19 cases during a visit to the emergency room, would it be possible to predict which patients will need to be admitted to a general ward, semi-intensive unit or intensive care unit?

### Questions and More Info

Additional questions and clarifications can be obtained at data4u@einstein.br

### Answers to most voted questions

### Missing data

Decision making by health care professionals is a complex process, when physicians see a patient for the first time with an acute complaint (e.g., recent onset of fever and respiratory symptoms) they will take a medical history, perform a physical examination, and will base their decisions on this information. To order or not laboratory tests, and which ones to order, is among these decisions, and there is no standard set of tests that are ordered to every individual or to a specific condition. This will depend on the complaints, the findings on the physical examination, personal medical history (e.g., current and prior diagnosed diseases, medications under use, prior surgeries, vaccination), lifestyle habits (e.g., smoking, alcohol use, exercising), family medical history, and prior exposures (e.g., traveling, occupation).

The dataset reflects the complexity of decision making during routine clinical care, as opposed to what happens on a more controlled research setting, and data sparsity is, therefore, expected.

### Variables in addition to laboratory results

We understand that clinical and exposure data, in addition to the laboratory results, are invaluable information to be added to the models, but at this moment they are not available.

### Additional laboratory variables

A main objective of this challenge is to develop a generalizable model that could be useful during routine clinical care, and although which laboratory exams are ordered can vary for different individuals, even with the same condition, we aimed at including laboratory tests more commonly order during a visit to the emergency room. So, if you found some additional laboratory test that was not included, it is because it was not considered as commonly order in this situation.

# Our message to all participants

Hospital Israelita Albert Einstein would like to thank you for all the effort and time dedicated to this challenge, the community interest and the number of contributions have surpassed our expectations, and we are extremely satisfied with the results.

These have been challenging times, and we believe that promoting information sharing and collaboration will be crucial to gain insights, as fast as possible, that could help to implement measures to diminish the burden of COVID-19.

The multitude of solutions presented focusing on different aspects of the problem could represent a valuable resource in the evaluation of different strategies to implement predictive models for COVID-19. Besides the data visualization methods employed could make it easier for multidisciplinary teams to collaborate around COVID-19 real-world data.

Although this was not a competition, we would like to highlight some solutions, based on the community and our review of results.

Lucas Moda (<https://www.kaggle.com/lukmoda/covid-19-optimizing-recall-with-smote>) utilized interesting data visualization methods for the interpretability of models. Fellipe Gomes (<https://www.kaggle.com/gomes555/task2-covid-19-admission-ac-94-sens-0-92-auc-0-96>) used concise descriptions of the data and model results. We saw interesting ideas for visualizing and understanding the data, like the dendrogram used by CaesarLupum (<https://www.kaggle.com/caesarlupum/brazil-against-the-advance-of-covid-19)>. Ossamu (<https://www.kaggle.com/ossamum/eda-and-feat-import-recall-0-95-roc-auc-0-61>) also sought to evaluate several data resampling techniques, to verify how it can improve the performance of predictive models, which was also done by Kaike Reis (<https://www.kaggle.com/kaikewreis/a-second-end-to-end-solution-for-covid-19>) . Jairo Freitas & Christian Espinoza (<https://www.kaggle.com/jairofreitas/covid-19-influence-of-exams-in-recall-precision>) sought to understand the distribution of exams regarding the outcomes of task 2, to support the decisions to be made in the construction of predictive models.

We thank you all for the feedback on available data, helping to show its potential, and taking the challenge of dealing with real data feed. Your efforts let the feeling that it is possible to build good predictive models in real life healthcare settings.