Covid-19 Dataset Analysis

# Project overview

The Covid-19 virus crushed the world during the year 2019-2020, killing thousands of people, destroyed companies and the economy of many countries - among other things. It became primarly important to analyze every aspect known of the virus to prevent another outbreak, synthesize a vaccine or understand how to manage the people likely to be contaminated considering their symptoms.

The current dataset focuses on this ending point. Using different analysis conducted on patients, the objective is to determine if they are really infected (false positives) and therefore which unit they need to be conducted to.

Abstract :

« 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. »

# Code and Resources Used

**Python Version:** 3.8  
**Packages:** pandas, numpy, sklearn, matplotlib, seaborn, scipy

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

# Form analysis

The target is the « SARS-Cov-2 exam result » taking « positive » or « negative » state, in a dataset of 5644 lines and 111 columns. The variables analysis shows 74 quantitatives, 37 qualitatives.

Some values are missing and two groups appears separated :

* + ~ 76 % missing values for other virus tests
  + ~ 89 % missing valeurs for blood analysis

# Substance analysis

For the target :

* 10 % positives
* 90% negatives

It’s very unbalanced, and we will need to sample the negatives results during the subset analysis to get relevant information.

Signification of the variables :

* Variables standardized, somethimes asymetrics, concerning the blood samples
* Age quantile : hard to conclude anything because the data have been mathematically shifted or transformed
* Qualitatives variables : are binaries (0, 1) detected/not detected,

NB : Rhinovirus seems to be anormaly high, this hypothesis needs to be checked later.

Relation variables to target :

* Target/Blood, idea of features that may be correlated :
* Leucocyte
* Monocyte
* Platelets

🡪 These rates are different between patient positively and negatively tested for the Covid19. We have to check later if it seems likely correlated.

* Target/Age : Young individuals seems less likely to be tested positives (it doesn’t mean they are not infected). The exact age is unknown.
* Target/Viral : It’s rare to find people with more than one sickness at a time.

As already said, Rhinovirus/Entérovirus positive may implied a negative Covid19 result. This hypothesis requires to be validate because it’s likely that the area from where the data are collected just suffered an outberak simultenously to the Covid19. It may be unrelated.

# Analyse plus détaillée

* Relation between the variables :
  + Blood\_data / Blood\_data : some variables are correlated (+0.9 !)
  + Blood\_data / Age : week correlation
  + Viral / Viral : influeza rapid test gives bad results and needs to be droped
  + Relation sickness / Blood\_data : Blood rates between regular patient and covid19 patient are différent (lymphocyte, hemoglobine et hematocrite)
* NaN analyse : viral 1350 (92%/8%), blood sample 600 (87%/13%), previously : 90% of the dataset

**Student’s test (H0) :**

* Patients infected with covid-19 have higher leucocyte, monocyte et platelets rate than regular individuals
  + H0 = These average rates are EQUALS between people tested positive and negative to covid-19
* Result : Rejected, because patients with a more common decease also get significantly higher rates, not exclusively covid-19 patients

**Conclusions :**

* A large part of the dataset is missing : only 20% is exploitable
* Two main groups interesting : blood and viral analysis
* The blood sample can’t give the certainty of Covid19 cases
* Some missing values need to be replaced, we can’t just drop them all. If we do so, we get 99 lines instead of 5644, so we lose to much information
* Blood\_column : 600 values, viral\_column : 1354