Original dataset:

* Diagnosis of COVID-19 and its clinical spectrum
* AI and Data Science supporting clinical decisions (from 28th Mar to 3st Apr)
* Original dataset contains anonymized data from patients in Brazil
* Samples collected to perform the SARS-CoV-2 RT-PCR and additional laboratory tests during a visit to the hospital.
* (5644 patients, 111 test/sample categories):

Removing some categories with a lot of missing data:

* Used .isna() to find out percentage of missing data from each categories
* Set threshold to 85%
* Removed categories with above 85% patient data missing
* (5644 patients, 23 categories)

Adding back important categories:

* Went back to excluded columns and chose ones below 90% missing based on function of tests and how it could play a role in COVID-19
* Removed nan values from dataset
* (5644 patients, 35 categories)

1)Remove all patients with missing information

(242 patients, 81 categories)

2)Separate categories by types into object(words) and numerical(float + int) (working with missing data)

* Object categories: change existing data from object into int (0 / 1) + separate missing data into a new column (using one-hot encoding)
* Numerical categories: fill in missing values with the mean of each columns (using SimpleImputer)

Joined object-categories dataframe with filled-in numerical-categories dataframe

(5644 patients, 72 categories)

[tried this dataset with machine learning models. Because the ratio of tested positive and tested negative COVID-19 patients in the dataset is very uneven, predictions was very inaccurate (a lot of false negative and false positives)]

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Separate patients based on COVID-19 status (positive/negative)

Keep all rows of tested positive COVID-19 patients (all information / categories)

(558 patients, 72 categories)

Keep tested negative COVID-19 patients with minimal data (other categories) missing

Randomly select the same number of negative COVID-19 patients as positive patients

(558 patients, 72 categories)

Combine the 558 positive patients with the 558 negative patients (final dataset)

(1116 patients, 72 categories)

Export data

[tested this dataset. In comparison with the previous dataset, machine learning models performed much better (more accurate)]