The dataset was provided by the Mexican government (link). This dataset contains an enormous number of anonymized patient-related information including pre-conditions. The raw dataset consists of 21 unique features and 1,048,576 unique patients. In the Boolean features, 1 means "yes" and 2 means "no". values as 97 and 99 are missing data

# Data Dictionary

* sex: 1 for female and 2 for male.
* age: of the patient.
* classification: covid test findings. Values 1-3 mean that the patient was diagnosed with covid in different degrees. 4 or higher means that the patient is not a carrier of covid or that the test is inconclusive.
* patient type: type of care the patient received in the unit. 1 for returned home and 2 for hospitalization.
* pneumonia: whether the patient already have air sacs inflammation or not.
* pregnancy: whether the patient is pregnant or not.
* diabetes: whether the patient has diabetes or not.
* copd: Indicates whether the patient has Chronic obstructive pulmonary disease or not.
* asthma: whether the patient has asthma or not.
* inmsupr: whether the patient is immunosuppressed or not.
* hypertension: whether the patient has hypertension or not.
* cardiovascular: whether the patient has heart or blood vessels related disease.
* renal chronic: whether the patient has chronic renal disease or not.
* other disease: whether the patient has other disease or not.
* obesity: whether the patient is obese or not.
* tobacco: whether the patient is a tobacco user.
* usmr: Indicates whether the patient treated medical units of the first, second or third level.
* medical unit: type of institution of the National Health System that provided the care.
* intubed: whether the patient was connected to the ventilator.
* icu: Indicates whether the patient had been admitted to an Intensive Care Unit.
* date died: If the patient died indicate the date of death, and 9999-99-99 otherwise.

# INSIGHTS

* All variables are categorical variables except for age.
* The death variable after converting date died to a categorical variable is highly skewed towards ‘no’ which is expected as the mortality rate of covid 19 is low.
* Sex variable is balanced with male and female being the two categories.
* Co – morbidities like pneumonia, asthma, copd, cardiovascular and renal chronic are skewed towards ‘no’.
* Tobacco consumption, diabetes, hypertension and obesity are skewed towards ‘no’ but have a substantial presence in ‘yes ’ as well.
* The distribution of ‘age’ is normal.
* DecisionTree, Random Forest, Extra trees and GaussianNB algorithms were used for prediction. All 4 models yielded an accuracy of more than 90 %