Covid-19 classifier for under reported SARS cases in Brazil

This project will attempt to make a classifier for under reported cases of Covid-19 caused SARS in Brazil, through K-Means clustering and a recall focused linear classification model.

Domain Background and Problem Statement

Covid-19 is the main news in 2020, but in Brazil it's been neglected by the government and has not been notified nearly enough, due to lack of tests and public investiment in research and medical support.

The aim of this project is to estimate (in a sort of statistical and non formal manner) what is the rate of undernotification of Covid-19 deaths and hospitalizations midst the notified SARS (SRAG in portuguese) cases.

Datasets and Inputs

The project will use the openDataSUS SRAG data from 2020 and 2019. Both of the datasets are CSV files, for which the column data is described in the dictionary files.

The data is pretty organized and is described in a tabular manner, which makes it easier to input and decode. The data is also stored in AWS, perfectly fitting for downloading from SageMaker.

Solution Statement and Evaluation Metrics

The aim of the project is to create a model that can identify non-reported Covid-19 data from SRAG data. This data is not labelled (since it's not been identified), so the evaluation will be split into 2 parts.

- recall on identifying Covid-19 confirmed data.
- precision on 2019's data

The idea behind these metrics is that if the model has high recall on 2020's data, it identifies Covid-19 cases easily, and if it has high precision on 2019's data, it means that the number of actual false positives is low.

As for the benchmark model, I didn't really find any other projects that attempt on doing what this is doing.

Project Design

The design will be based on two main approaches. The first approach is an analysis driven approach, on which the data will be clustered and these clusters will be compared between cases that were and weren't diagnosed with Covid-19.

The second approach will be a binary classifier driven approach, where a model will be trained with 2019's data as negative and 2020's data as positive and test data. The data will have two different sets of test files, one for 2019, on which I'll test the model for precision and recall, and one for 2020, on which I'll test for how much the precision drops, and, with that, estimate the ammount of unreported cases in the data.