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 problem is a classification problem, where a model will be trained to classify a given entry in the SRAG data between COVID-19 positive and negative sample, given the information contained in the individual file.

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.

The data has 72 obligatory columns, but has 134 columns in total, some of which have it's data's existence determined by the value of other columns (like the age of the subject of the given row, etc).

Each row represents a single registered case of SRAG, with information regarding from the symptoms of the case to the district and location on which the case was reported and registered.

There's a column representing the final diagnostic of the case, where there are 5 categories, two of them being not identified and COVID-19 positive cases. Those are the two main categories on which I'll be focusing, considering they represent most of the data (almost 90% considering 2019 and 2020). The aim will be to be able to identify if the cases are positive for COVID-19 or negative, considering the symptoms of patients registered before September 2019 which are classified as non identified and all patients which are identified as COVID-19 positive.

Using the data in this manner makes the lack of reliability in the 'not identified' data from 2020 make less of an impact in the precision for negative cases, considering that the disease was not around in 2019.

Solution Statement and Evaluation Metrics

The solution is to have a working classification model for COVID-19 cases reported midst SRAG not identified cases in Brazil's SRAG database. The classifier will be evaluated considering it's training, that is, it'll be evaluated using the 2019's data as negative samples and 2020's data as positive ones.

The model will be evaluated using the `F1 score`, and this evaluation will be compared with the XGBoost's model accuracy.

Benchmark models

The data is mostly tabular data, and I couldn't find any similar research on this specific dataset, but even then, tabular classification is a classic problem that has been tackled in many different forms, from SVMs with kernels to XGBoost and Random Forest and similar ensemble models.

In specific, I'll be comparing the model with XGBoost for classification.

Project Design

The project will consist in three main steps:

- Data exploration and processing
    - where I'll look into the data and normalize and change categorical data into numeric, and also remove unused columns (or rarely used ones)
- Feature engineering and dimensionality reduction
    - where I'll search for different features and correlation between columns, to make the data more usable and representative
- Model selection
    - where I'll look into different model architectures and work out the best one (and the best hyperparameters)
        - This part will contain some different model archtectures, including SVMs with kernels, pytorch fully connected models, SageMaker's linear learner and other classic classification models.