📊 Data Science Workflow Canvas*

Start here. The sections below are ordered intentionally to make you state your goals first, followed by steps to achieve those goals. You're allowed to switch orders of these steps!

Title: COVID EM MINAS GERAIS

1 Problem Statement

What problem are you trying to solve?
What larger issues do the problem address?

O objetivo deste trabalho é classificar os pacientes que tem chance de se recuperar, entrar em acompanhamento ou vir a óbito após se contaminar com a COVID-19.

O maior problema deste dataset é o desbalanceamento de classes. Esse desbalanceamento foi causado pela natureza do problema, onde por se tratar de um vírus que reage diferente em cada pessoa, houveram mais casos de pessoas recuperadas do que óbitos.

2 Outcomes/Predictions

What prediction(s) are you trying to make?
Identify applicable predictor (X) and/or target (y) variables.

A variável alvo do dataset é a EVOLUCAO, onde a classe 1 é designada para RECUPERADO, a classe 2 é designada para EM ACOMPANHAMENTO e a classe 3é designada para OBITO.

As variáveis de "input" são: SEXO, COMORBIDADE, INTERNACAO, UTI, RACA, MACRO, URS, FAIXA_IDADE.

3 Data Acquisition

Where are you sourcing your data from?
Is there enough data? Can you work with it?

Os dados foram retirados após uma busca ativa no site do Governo Federal e no site do Governo Estadual de Minas Gerais.

Ao todo, foram mais de 1 milhão de dados, quantidade esta suficiente para se realizar o Machine Learning.

4 Modeling

What models are appropriate to use given your outcomes?

Os modelos escolhidos foram: Decision Tree Classifier, Random Forest Classifier e Light Gradient Boost Machine.

Esses modelos foram escolhidos por terem suporte para problemas de classificasão categórica multiclasse.

5 Model Evaluation

How can you evaluate your model's performance?

Para avaliar a performace dos modelos, foi optado pelo classification report da biblioteca Sklearn, pois ele retorna a Acurácia, Precisão, Recall e F1 Score para cada classe.

Para este problema, o recall será avaliado com mais ênfase, pois é de extrema importância classificar corretamente as pessoas que tem a probabilidade maior de vir a óbito, para que os médicos possam se atentar a elas.

6 Data Preparation

What do you need to do to your data in order to run your model and achieve your outcomes?

Para treinar os modelos, os seguintes tratamentos de dados foram realizados:
- Análise de outliers, optando por eliminá-los.
- Eliminação de variáveis que não agregam informação relevante para os modelos.
- Eliminação de algmas variáveis que tinham como predominante, valores nulos.
- Preenchimento de valores nulos nas variáveis MACRO e URS.
- Balanceamento dos dados.
- Separação da variável alvo do resto dos dados.

✅ Activation

When you finish filling out the canvas above, now you can begin implementing your data science workflow in roughly this order.

1 Problem Statement → 2 Data Acquisition → 3 Data Prep → 4 Modeling → 5 Outcomes/Preds → 6 Model Eval

*** Note:** This canvas is intended to be used as a starting point for your data science projects. Data science workflows are typically nonlinear.